
# Statistical Methods of Detecting Change Points for the Trend of Count data

Taghreed Mohammed Jawa

Department of Mathematics and Statistics

University of Strathclyde

Glasgow, UK

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science

July 2017

# Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.



Signed:

Date:

# Acknowledgements

First and foremost, praises and thanks to the Almighty Allah (God), the almighty Merciful, who showered me with health, strength and ability to complete my thesis. Above all, I would like to thank my father Mohammed and my mother Latifa for their endless prayers, love and great patience, I am beyond grateful. I would like to express my deepest thanks to my husband Omar for all his love, support and encouragement and for his patience of being away from his mother and family during the period of this research.

The success of this thesis is attributed to the comprehensive support from my supervisors; Doctor David Young and Professor Chris Robertson. I would like to express my grateful gratitude to them for their suggestions, valuable advice, supervision, encouragement, providing feedback for my thesis and for their kindness during the course of my study. I also thank the examiners Dr. Andrea Sherriff and Dr. Kimberley Kavanagh for spending time on reading my thesis and giving me valuable advice, comments and suggestions to get my PhD in the best way.

I take this opportunity to thank the Ministry of Education in Saudi Arabia and Taif University for their financial support during my study. I would like to acknowledge the academic and technical support I received at the university of Strathclyde, particularly from Dr. Stephen Corson, Dr. Ronnie Wallace, Mr.

Ian Thurlbeck, Mrs. Irene Spencer, my colleagues; Dr. Abdullah Almarashi, Dr. Wafa Al-fwzan, Dr. Fatemah Al-mukahal and Maha Alsharari as well as from all staff and colleagues in the Mathematics and Statistics Department. I also thank Health Protection Scotland for providing me with the data for this study.

Last but not least, many thanks for those who gave me suggestions, constructive comments and supported me during my study period. Thank you to all members of my family in particular my siblings; Alaa, Baraa, Rawan and Duha, my nephew Mohammed and my sister-in-law Sara for giving me the strength to achieve such success. I wish to extend my thanks to my friends; Shaima, Rania, Noha, Manal, Danya, Afnan and Alya for their continues encouragement and support. I also thank Alaa, Heba, Saja and Heila for supporting me while we were studying in the United Kingdom.

Finally, I express my deepest thanks to my daughter Rateel for her love. This work is dedicated to her, my husband, my parents, my sisters and my brother.

# Abstract

In epidemiology, controlling infection is a crucial element. Since healthcare associated infections (HAIs) are correlated with increasing costs and mortality rates, effective healthcare interventions are required. Several healthcare interventions have been implemented in Scotland and subsequently Health Protection Scotland (HPS) reported a reduction in HAIs [HPS (2015b, 2016a)]. The aim of this thesis is to use statistical methods and change points analysis to detect the time when the rate of HAIs changed and determine which associated interventions may have impacted such rates.

Change points are estimated from polynomial generalized linear models (GLM) and confidence intervals are constructed using bootstrap and delta methods and the two techniques are compared. Segmented regression is also used to look for change points at times when specific interventions took place. A generalization of segmented regression is known as joinpoint analysis which looks for potential change points at each time point in the data, which allows the change to have occurred at any point over time. The joinpoint model is adjusted by adding a seasonal effect to account for additional variability in the rates. Confidence intervals for joinpoints are constructed using bootstrap and profile likelihood methods and the two approaches are compared. Change points from the smoother trend of the generalized additive model (GAM) are also estimated and bootstrapping is used to construct confidence intervals.

All methods were found to have similar change points. Segmented regression detects the actual point when an intervention took place. Polynomial GLM, spline GAM and joinpoint analysis models are useful when the impact of an intervention occurs after a period of time. Simulation studies are used to compare polynomial GLM, segmented regression and joinpoint analysis models for detecting change points along with their confidence intervals.

# Publications and presentations

- Some of the work in Chapters 3 and 4 has been drafted for journal publication in the Journal of Hospital Infection and is titled as "The Impact of NHS Infection Control Interventions on Rates of Healthcare Associated Infections".

- A full account of the work in Chapter 5 has been drafted for publication in Communications in Statistics journal and is titled as "Change Points Analysis for the Trend of Count Data".

- Parts of the work in Chapters 4 and 7 has been drafted for journal publication in Communications in Statistics journal and is titled as "Estimation of Change Points from Regression models of Count Data".

- A full account of the work in Chapter 6 is currently in preparation for publication.

- Aspects of the work described in Chapters 3, 4 and 5 have been presented at the 38th Research Students' Conference in Probability and Statistics in Leeds in 2015 and at the 9th Saudi Students' Conference in Birmingham in 2016.

- The main results described in Chapters 4, 5 and 7 will be presented at the Royal Statistical Society International Conference in Glasgow in 2017.

# Contents

# Chapter 1

# Introduction to Healthcare Associated Infections

Everything in the world tends to change over time (e.g. economic stability, education strategies and epidemiology). It is of interest to know when changes occur in order to identify interventions associated with these changes. The development of medical interventions in epidemiology and public health is an essential aspect worldwide.

## 1.1 Overview of the thesis

This thesis investigates statistical methods for the detection of changes in rates of infection. This change may be associated with changes to practice. The main methods used to detect these changes are polynomial generalized linear models (GLM), segmented regression analysis, joinpoint analysis and generalized additive models (GAM). This research also explores methods of constructing confidence intervals for the change points. All these methods are considered within the scope of healthcare associated infections (HAIs). Simulation studies are then carried out to investigate the best method of detecting change points

and the best method to construct confidence intervals for these change points.

In Chapter 2, a literature review of statistical modelling and detecting change points is presented. Statistical methods of modelling count data with polynomial generalized linear regression models and generalized additive models are reviewed. Also, detection of change points in several types of data is presented. Finally, a review on some selection methods of constructing confidence intervals is discussed.

In Chapter 3, methods for modelling the rates of HAIs over time are implemented to describe the change in trend. Statistical methods to model count data using Poisson and quasi-Poisson polynomial models are presented. Additionally, descriptive analysis of rates of methicillin-resistant staphylococcus aureus (MRSA) bacteraemia, methicillin-sensitive staphylococcus aureus (MSSA) bacteraemia and clostridium difficile infection (CDI) in Scotland are described. This is done on Scottish data and also by each of the 15 health boards in Scotland. Power of the models and sample size issues are investigated and funnel plots are then used to compare health boards rates of HAIs. A brief summary and discussion of findings conclude this chapter.

In Chapter 4, the best fitted polynomial generalized linear models (GLM) from Chapter 3 are used to estimate the time when the rates of HAIs change (turning points) and to determine which interventions had an impact on healthcare associated infection. Confidence intervals for estimated turning points are constructed using bootstrapping. A simulation study is carried out to compare bootstrap and delta methods for constructing confidence intervals for a single turning point.

In Chapter 5, a general approach which detects change points where rates change significantly and determines which interventions are associated with these changes is discussed. Segmented regression analysis is used to detect change points where the rates of HAIs change significantly after a specific intervention and then identify if some or none of these interventions have an impact on these rates. Joinpoint analysis is a generalization of segmented regression and it is used to estimate the existence of change points at unspecified times and estimate their location. The joinpoint method tests all data points to identify change points. The confidence intervals of joinpoints are constructed using bootstrap and profile likelihood methods. A simulation study compares these methods in order to find the best method of constructing confidence interval for joinpoint.

Chapter 6 aims to investigate and discuss the change point methods which were used in Chapters 4 and 5 through a simulation study. This identifies and compares particular changes in trends to determine which methods detect changes more easily and more accurately than others. Polynomial models including quadratic and cubic models are used to estimate the change in trend. Segmented regression identifies changes at particular times where the smallest deviance occurs and joinpoint analysis estimates the number of changes and their location from all possible times during the period of study. This simulation is carried out with different sample sizes and different scenarios of the original assumptions and values.

In Chapter 7, a new method is introduced to detect change points which uses generalized additive models (GAM). GAM is considered as a more flexible model than generalized linear models (GLM) and can fit a smoother trend to the data. A spline function is used within GAM models and the best model

is fitted. The method of estimating change points from the GAM model is demonstrated and confidence intervals of estimated change points are constructed using bootstrapping. The method is then used to detect change points for the HAI data.

Finally in Chapter 8, the conclusion and future work recommendations are listed. Some of the future work recommendations suggest further developing specific techniques using simulation studies and modifying the change point methods to better suit HAI data.

In the rest of this chapter (Chapter 1), background and some previous studies on HAIs including MRSA, MSSA and CDI are detailed. The data which were provided by the Health Protection Scotland (HPS) [HPS (2013)] on staphylococcus aureus bloodstream infections (SAB) including MRSA bacteraemia and MSSA bacteraemia and clostridium difficile infection (CDI) will be analysed. We will look at the trend of infections over time and detect time points when the pattern of data changes to assess the effect of preventative health care interventions.

In public health care and as a result of health care interventions, infections could take place either outside or inside a hospital. If an infection occurs at home or in health care centers and diagnosed by GP testing or diagnosed in a hospital within 48 hours of admission and are relevant attributable to community exposure, they are termed as community-acquired infections. If the infections developed during a hospital stay and diagnosed after 48 hours of admission, they are known as hospital-acquired infections [Wertheim (2005)]. Also, the definitions can vary dependent on the organism. Therefore, the community and hospital acquired distinction is based on an organism dependent

epidemiological definition based on previous healthcare exposures. Such terms are collectively referred to as healthcare associated infections (HAIs).

Healthcare associated infections are a major cause of patient morbidity and mortality. Staphylococcus aureus (SA) and clostridium difficile (C. diff) are among the most common infection-causing bacteria. SA bloodstream infection (SAB) is a type of HAIs [Tong et al. (2009)]. Bloodstream infection (bacteraemia) is clinically defined as the isolation of bacteria from one or more peripheral venous blood-culture samples collected from patients with associated relevant symptoms and signs of systemic infection [Thwaites et al. (2011)]. About 13% of all hospital-acquired bloodstream infections are caused by staphylococcus aureus [Wertheim (2005)]. Clostridium difficile infection is also a type of HAIs and causes serious bowel problems [NHS (2016a)].

This chapter covers four main topics:

1. The historical and biomedical background of healthcare associated infections (HAIs) including MRSA, MSSA bacteraemias (see Section 1.2.1) and CDI (see Section 1.3.1).

2. An impact of risk factors on MRSA and MSSA bacteraemias (see Section 1.2.2) and CDI (see Section 1.3.2).

3. Some studies showing the impact of healthcare interventions on the rate of HAIs (see Section 1.4).

4. A discussion of methods used for collecting data on MRSA and MSSA bacteraemias and CDI and the interventions conducted in Scotland (see Section 1.5). Such data will be used for analysis in this research.

## 1.2 Introduction to staphylococcus aureus bacteraemia

### 1.2.1 Historical background

Staphylococcus aureus (SA) was initially discovered by a Scottish surgeon (Alexander Ogston) in 1881 when he described the presence of grape-like clusters of globular micro-organisms in pus from abscesses [Özgen (2008) and Ekkelenkamp (2011)]. SA is a gram positive bacterium of about one micrometer in diameter [Plata et al. (2009)], which colonises the skin of about 30% of the healthy human population without causing infection [Wertheim (2005)]. Colonization with SA may occur any time after birth, and its carriage may be temporary or permanent [CFSPH (2011)]. Majority of individuals carry SA in their nose or on their skin without knowing that they are carrying it. They do not have skin infections or any other signs or symptoms of illness. Although this colonisation is usually harmless, SA is an important cause of serious infections [Wertheim (2005)]. These infections are commonly associated with healthcare interventions due to failures of implementing infection control methods [ECDPC (2012)]. One of the reasons that SA is causing infections is that it can survive for about one month on any type of surface [Wertheim (2005)]. Therefore, a simple wound infection can become contaminated by SA which can enter the bloodstream where it is transported to internal organs, skin and bone. This can cause severe infections with high mortality rates [Wertheim (2005)].

SA has two types of strains. Some SA bacteria are more resistant to the antibiotic methicillin which is the first member of this class of antibiotics [NHS (2016a)]. These are called methicillin-resistant staphylococcus aureus (MRSA) and often require different types of antibiotics to treat them. However,

methicillin-sensitive staphylococcus aureus (MSSA) can effectively be treated by antibiotics. Both MRSA and MSSA are endemic in many UK hospitals, causing a range of infections [NHS (2016a)] such as various skin and soft tissue infections (surgical wound infections), pneumonia, endocarditis, septic arthritis, osteomyelitis, meningitis and bacteraemia (bloodstream infection or blood poisoning which is commonly referred to as staphylococcus aureus bacteraemia (SAB)) [Wertheim (2005)].

MRSA is the most commonly identified antimicrobial-resistant pathogen in hospitals in many countries in the world including Europe, the United States, North Africa, South-East Asia and the Middle- and Far East [Thomas (2014)]. Health Protection Scotland reported that SAB is the serious type of infection leads to increased morbidity and mortality which requires treatment by antimicrobial therapy courses [HPS (2017)]. The first MRSA case was recognised in October 1960, and the first MRSA isolate was detected a few months later in February 1961 at a hospital in the United Kingdom. After a few years, MRSA was found in other European countries as well as Japan and Australia [ISMR (2006)]. Some MRSA strains, called epidemic strains, are more prevalent and tend to spread within or between hospitals and countries [CFSPH (2011)].

Staphylococcus aureus is usually transmitted by direct contact, often via hands, with colonized or infected people. It may also be spread by sharing personal hygiene items that have been touched by people with staphylococcus aureus like towels, soaps or clothes [NHS (2016a)].

The symptoms of an SA infection and the symptoms of an infection arising from other staphylococcus are similar. For example, pimples, rashes and pus-filled boils are indicative of SA skin infection which may be considered as a

minor infection especially when the rashes, pimples or boils are warm, painful, red or swollen [NHS (2016a)]. However, the symptoms for very serious infections can also occur which include severe skin infection, surgical wound infections, bloodstream infections and pneumonia include high fever, swelling, heat and pain around a wound, headache, fatigue and other symptoms [CFSPH (2011)].

If an antibiotic is required for treatment, MSSA infection can be treated by a penicillin-based antibiotic such as flucloxacillin. This antibiotic is prescribed for patients who are not allergic to penicillin, otherwise alternative antibiotics may be prescribed [NHS (2016a)]. However, the treatment of MRSA can be challenging because it can only be treated with antibiotics based on susceptibility testing [CFSPH (2011)]. MRSA is resistant to a family of penicillin-related antibiotics such as methicillin, oxacillin and flucloxacillin [NHS (2016a)]. MRSA cannot be identified without specific lab tests and it is not always recognised and treated correctly when antibiotic treatment is needed. SA infections are diagnosed by culture and identification of the organism [CFSPH (2011)]. Sometimes doctors diagnose an MRSA infection as a common staph infection and they prescribe antibiotics that do not kill MRSA. Such potential delay with appropriate antibiotics to treat MRSA infections can effectively result in prolonged illness and rare life-threatening diseases in the blood, heart and bones [CFSPH (2011)].

### 1.2.2 Impact of risk factors

In this section, some selected studies investigating risk factors of MRSA and MSSA infections and especially of MRSA and MSSA bacteraemias are reviewed.

Increasing SA infections causes mortality, morbidity and high expenditures. MRSA bacteraemia is associated with significant increase in mortality rates compared with MSSA bacteraemia (odds ratio = 1.93) [Cosgrove et al. (2003)]. Patients with MRSA infection in the USA during 2004-2006 had a higher mortality rate (23.6%) than patients with MSSA infection (11.5%) [Filice et al. (2010)]. In addition, patients with MRSA infection have more co-morbidities than patients affected with MSSA infection [Filice et al. (2010)]. In the UK, about 12,500 cases of SAB each year are reported, with associated mortality rates of about 30% with 95% confidence interval (15% - 60%) compared with bacteraemia caused by other pathogens [Thwaites et al. (2011)]. In 2016 in Scotland, 1,599 incidences of SAB were reported by Health Protection Scotland where 5.5% were MRSA bacteraemia and 94.5% were MSSA bacteraemia and these reported 1.9% increase in the overall incidence of SAB since 2012. In addition, between 2011 and 2015, Health Protection Scotland reported that 26.2% of mortality per month are caused by MRSA bacteraemia and 19.2% by MSSA bacteraemia [HPS (2017)].

MRSA infections are independently associated with increased costs and therefore efficient infection prevention programs are needed to reduce the incidence of these costly infections. Bacteraemia treatment is expensive and difficult [Tong et al. (2009)]. For example, in New York state, bacteraemia and pneumonia caused by SA is responsible for about 60% of the total direct medical costs and 97% of the mortality compared to other types of infection [Rubin et al. (1999)]. Furthermore, staphylococcus aureus bacteraemia (SAB) is the second most common serious bacterial infection worldwide and is a major cause of increased length of hospital stay, antibiotic use and associated costs [Wertheim et al. (2004) and Wertheim (2005)]. In North Carolina, the mean initial hospitalization cost was significantly greater for patients with complic-

ated SAB versus uncomplicated SAB ($32,462 vs $17,011, p=0.002) [Engemann et al. (2005)]. Between December 1993 and March 1995, hospital-acquired bacteraemia due to SA significantly extended the duration of hospitalization for about four days for MSSA bacteraemia patients and 12 days for MRSA bacteraemia patients. Therefore, direct costs increased three fold due to MRSA ($27,083) compared with those due to MSSA ($9,661) [Abramson and Sexton (1999)]. In the UK between April 1991 and December 1992, the hospital costs was a total of £403,600 for patients with an MRSA infection [Cox et al. (1995)]. Recent studies in the UK illustrated the costs associated with MRSA screening for all admissions and recommended to improve the policy of screening patients [Robotham et al. (2016)].

There are several risk factors associated with increasing SA infections where these include age, some diseases, teaching hospital and unavailable treatment. Elderly patients over 64 years are more likely to have MRSA infections in Scottish hospitals [Van Velzen et al. (2011)]. Age is the most common consistent predictor of mortality in patients with SAB [Van Hal et al. (2012)]. Recurrences of SAB occurred in 9.4% following anti-staphylococcal therapy where elderly patients with severe disease who have MRSA bacteraemia are more likely to experience delay in appropriate antimicrobial therapy (DAAT). This is associated with increased in-hospital mortality [Marchaim et al. (2010)].

Several diseases are associated with contracting SA infections inside or outside hospitals. Healthcare associated MRSA strains are the main causes of nosocomial infections associated with indwelling medical devices and surgical sites [CFSPH (2011)]. From 1994 to 1998 in the United States, the incidence of MRSA increased by 37% in patients hospitalized in the intensive care unit [Cosgrove et al. (2003)]. From 1994 to 2000, patients with surgical site infec-

tions (SSIs) in Durham in the USA infected with MRSA had a greater mortality rate than patients infected with MSSA (odds ratio 3.4). Those infected with MRSA also had a greater duration of hospitalization after infection (five additional days) which increased the median hospital cost to about $52,791 for patients with MSSA SSI, and $92,363 for patients with MRSA SSI [Engemann et al. (2003)]. Human immunodeficiency virus (HIV) is also associated with a high risk of acquiring an SA infection. Deep soft tissue infections have been observed in HIV positive patients which is associated with increased rates of SAB [Wertheim (2005)]. Patients with an orthopedic device infection (ODI) also had a higher relapse of SA infection, compared to bacteremic patients without ODI [Lalani et al. (2008)]. Hemodialysis dependent patients hospitalized with MRSA bacteraemia have a higher mortality risk (odds ratio = 5.4 with 95% confidence interval (1.5, 18.7)) at 12 weeks. Longer hospital stay leads to higher inpatient costs which are about $21,518 for initial hospital stay and $25,518 after 12 weeks [Reed et al. (2005)]. There are several factors associated with an increased risk of developing SAB, including colonization or previous MRSA infection, skin ulcers at hospital admission, existent central venous catheters, urinary catheter insertion, surgical site infection, injecting drug use, presence of immunosuppressive conditions, use of corticosteroids as well as liver disease and lung disease [Naber (2009)]. Furthermore, it was reported that 33% of patients in Scotland developed an MRSA infection after pancreatoduodenectomy (PD) [Sanjay et al. (2010)]. Also, renal failure, and open wounds were significantly associated with acquiring of MRSA in Scotland [Van Velzen et al. (2011)].

Teaching hospitals are associated with increasing MRSA and MSSA infection. In a general teaching hospital in Brazil with a high prevalence of MRSA strains, MRSA bacteraemia had a high mortality rate (39% within 14 days)

[Conterno et al. (1998)]. In contrast, a tertiary-care teaching hospital in Boston demonstrated no significant difference between the mortality in patients with MRSA and MSSA bacteraemias (p=0.53) however, MRSA bacteraemia was associated with significant increases in length of hospitalization and hospital charges [Cosgrove et al. (2005)].

MRSA bacteraemia infections are widespread and difficult to treat. The incidence of MRSA bacteraemia and associated complications have increased in the United States and in some European countries because of the increased resistance of SA strains to available antibiotics [Naber (2009)].

## 1.3 Introduction to clostridium difficile infections (CDI)

### 1.3.1 Historical background

Clostridium difficile (C. diff) is a bacteria that infects the digestive system of about 1/30 healthy people and can cause diarrhoea. C. diff bacteria live in the digestive system without causing any problems in healthy individuals because it is controlled by the presence of other bacteria. However, occasionally treatment with antibiotics can affect these bacteria which subsequently results in uncontrolled C. diff bacteria, resulting in infection [NHS (2016a)]. Clostridium difficile infection (CDI) can especially occur in patients recently treated by broad-spectrum antibiotics or those using different antibiotics in the same period of treatment [NHS (2016a)].

C. diff bacteria are passed out of a patient through diarrhoea and can be transmitted to other individuals. C. diff bacteria can live outside the body (on

different surfaces) for a long time and can infect other individuals if they enter the body.

Symptoms of CDI sometimes occur during or after finishing the course of antibiotics. These symptoms include bloody diarrhoea, stomachache, dehydration, headaches and fever. Serious complications can be developed such as drowsiness and damage to the bowel [NHS (2016a)]. In serious cases of infection, patients may need a surgery to remove a damaged part of the bowel [NHS (2016a)].

### 1.3.2 Some previous studies on risk factors

Several risk factors are associated with contracting or a recurrence of CDI. Some strains of clostridium difficile produce toxins (types of bacteria release poisons -antigenic poison-) which cause an infection called toxin clostridium difficile infection (TCDI). Also, non-toxigenic clostridium difficile (NTCD) strains are found in hospitals and cause the infection NTCDI [Gerding et al. (2015)]. A study between 2012 and 2014 in Australian hospitals showed a reduction in the annual trend of toxigenic C. diff with associated significantly high rates in the summer. Non-toxigenic C. diff was associated with some chronic diseases such as kidney failure [Furuya-Kanamori et al. (2017)].

CDI is also more likely to occur in patients with cancer and kidney disease, those who have weak immune systems or who have had surgery on their digestive system in comparison to patients without these diseases [NHS (2016b)]. In patients with kidney disease there are significantly high risks of CDI [Phatharacharukul et al. (2015)]. In 2011 in the USA the incidence of CDI was higher in females than males, and in patients over 65 years old. Mortality was estimated

to be 9.6% per 100,000 patients [Lessa et al. (2015)]. A long hospitalization time is also a risk factor of CDI and patients over 65 years old are more likely to be infected with CDI compared to patients under 65 [NHS (2016b)].

Age is also associated with recurrence of CDI where elderly patients are more likely to experience a recurrence of CDI after taking antibiotics to treat it. Oral vancomycin (type of antibiotics) reduces the risk of recurrence of CDI in patients who are re-exposed to antibiotics [Carignan et al. (2016)]. Also, increased antibiotic usage and specific foods and drinks is linked with the recurrence of CDI [Carpenter et al. (2016) and Oman Evans II et al. (2016)]. Furthermore, a recent study in Scotland demonstrated that long time of using antibiotics is associated with CDI and high risk occurs after one month of antimicrobial exposure [HPS (2017) and Kavanagh et al. (2017)].

These risk factors are associated with increased the mortality and cost. In Czech Republic from January 2008 to December 2013, mortality by CDI was observed in patients over 85 years old which was associated with the number of antibiotics used (48%), presence of pressure ulcers (42%) and fever (37%) [Bielakova et al. (2016)]. In Scotland between 2012- 2016, the incidence of CDI in patients above 65 years have been decreased by 7.9% and HPS reported that between 2011- 2015 there was a year on year reduction in mortality rate of CDI by 3.5% [HPS (2017)]. Between 2005 and 2015, the total annual CDI cost in the USA was $6.3 billion where the total annual hospitalization is estimated to be 2.4 million days [Zhang et al. (2016)].

## 1.4 Infection control

As part of quality of patient care and safety, infection prevention should be considered a priority and integrated at all strategy levels within healthcare

to avoid HAIs [Lindberg (2012)]. Medical staff's knowledge about infection control, patient behaviours and medical examinations and treatments should be integrated to prevent and control HAIs.

Controlling infection within healthcare requires that staff have knowledge about infection control measures. Hospital staff also need to improve their knowledge regarding the best strategies to ensure effective infection control practices [Solberg (2000)]. They should realize the importance of prevention of infection and work together to limit the spread of infection by improving patient safety in healthcare facilities. For instance, commitment to hand disinfection before and after contact with patients is an important action for ensuring patient safety [Lindberg (2012)]. Hand washing is also important to prevent HAIs transmission among patients and hospital staff [NHS (2016a)]. A study in England and Wales suggests that the national Cleanyourhands campaign, including a hand hygiene campaign can reduce healthcare associated infections by controlling the spread of HAIs through the contamination of healthcare staffs' hands [Stone et al. (2012)]. This study showed that increasing usage of alcohol hand rub and soap is associated with reduction in CDI rates and falling of MRSA bacteraemia in the last year of the study. Detergent and hydrogen peroxide decontamination are used to clean rooms occupied by patients with MRSA infection following discharge. A study by Mitchell et al. (2014) in Australia demonstrated a reduction in the incidence of MRSA bacteraemia when using detergent (0.16/10,000 patient care days) and when using hydrogen peroxide decontamination (0.11/10,000 patient care days). It is worth mentioning that the reduction associated with the detergent is not significantly different from the reduction by hydrogen peroxide decontamination ($p=0.58$). However, hydrogen peroxide decontamination has a significant reduction for the incidence of MRSA colonisation and infection (5.3/10,000 patient care days)

compared to the reduction by detergent (9.0/10,000 patient care days) with $p<0.001$.

To control and prevent HAIs by patients themselves, people should keep their hands clean and dry and wounds and cuts should be cleaned and covered where patients should use their own hygiene equipment such as towels, toothbrushes, etc. [NHS (2016a)]. Outpatients with MRSA skin wounds should keep them covered with clean and dry dressings [CFSPH (2011)].

There is some evidence within European hospitals that MRSA bacteraemia can be reduced with developed medical interventions and infection control [Borg et al. (2014)]. Solberg (2000) suggested infection control strategies including screening and isolation of newly admitted patients and implementation of an infection control program to prevent transmission of resistant strains between patients and hospital personnel. Several studies by Lawes et al. (2012), Currie et al. (2014) and Coia et al. (2014) recommend MRSA screening for Scottish patients before hospital admission. The infection control practices in Scotland including universal admission screening and antibiotic stewardship were associated with decreasing MRSA bacteraemia [Lawes et al. (2012)]. However, most healthcare associated staphylococcus aureus infections are caused by the patient's own staphylococcus aureus cells, where patients carry the organism in their noses. To prevent these infections, staphylococcus aureus is eradicated from the nose by treatment with mupirocin nasal ointment [Wertheim (2005)]. Also, carriers of MRSA should be treated with intranasal antibiotics such as mupirocin to eliminate carriage [Solberg (2000)]. A Scottish study found that some changes should be implemented on infection control interventions to reduce and control MRSA infection in Scotland [Lawes et al. (2015)].

Several infection control polices to reduce the clostridium difficile infection are recommended by Vonberg et al. (2008) such as; early diagnosis of CDI, staff education, hand hygiene, environmental cleaning and cleaning of medical equipment and good antibiotic stewardship. Also, the usage of hydrogen peroxide vapour to disinfect the patient's room is associated with reduced clostridium difficile infection [McCord et al. (2016)]. National infection control procedures were adopted in Scotland over a period of time from 2004 to 2011 and are discussed in Section 1.5.3.

## 1.5 Data for analysis

Healthcare associated infections (HAIs) including data on MRSA, MSSA bacteraemias and CDI are described in this section. Health Protection Scotland (HPS) provide the data on HAIs from 2003 for MRSA and other infections from later years. Also, the data about the interventions is provided by HPS (2015b) from 2004 to 2011.

### 1.5.1 MRSA and MSSA bacteraemias data

The occurrence of SA bacteraemia is monitored by the HPS SA bacteraemia surveillance programme in Scotland as a notifiable diseases [HPS (2013)]. This includes SA bacteraemia occurring in patients who are under the healthcare system (in both acute and non-acute hospitals and in primary care settings) and those who have acquired SA bacteraemia in the community, without any healthcare contact. Many countries restrict surveillance of SA bacteraemia to those caused only by MRSA. However, the surveillance programme in National Health Service (NHS) Scotland and within individual NHS health boards includes data on both MRSA and MSSA bacteraemias. The quarterly SA bacteraemia data produced by HPS are based on tentative data for both bed

occupancy and incident SA bacteraemia. These data are subject to revision as finalised data become available [HPS (2012)].

Data was collected by HPS from January 2003 to June 2016 and includes 15 NHS boards, (see Table 1.1) [HPS (2016a)]. Data is collected regularly every three months (i.e. quarterly (Qu)) and records the number of patients with MRSA and MSSA bacteraemias and the number of acute occupied bed days (AOBDs) in 15 health boards in Scotland. AOBDs are based on the daily counts of occupied beds that are undertaken in every hospital at midnight. These counts exclude day patients who, by definition, do not occupy a bed at midnight [HPS (2012)]. Each case is reported to an NHS health board according to the location of the diagnostic laboratory where a patient was screened. In addition, if a patient is diagnosed twice within 14 days with two positive tests, duplicate cases have been removed [HPS (2015c)]. Rates of MRSA and MSSA bacteraemias are presented per 100,000 AOBDs and this gives an indication of the number of cases relative to the size of the population at risk. In 2009, NHS in National Waiting Times Centre was joined to NHS Scotland so MRSA and MSSA bacteraemias data were collected from April 2009. For example, Figure 1.1 explains the sort of MRSA and MSSA bacteraemias data which were used in this research. It also clearly shows that the trend of infections has changed over time where the trend of MRSA bacteraemia showed reduction over time but MSSA bacteraemia showed a little increase from 2013 up to 2016.

Several factors associated with risk are recorded to adjust the rate of MRSA and MSSA bacteraemias. Information Services Division (ISD) provides data which specify the percentage of acute surgical procedure (ASP) in each individual health board from April 2009 to December 2013 [ISD (2014)]. ASP is the percentage of patients who had surgical operations in each health board. Fur-

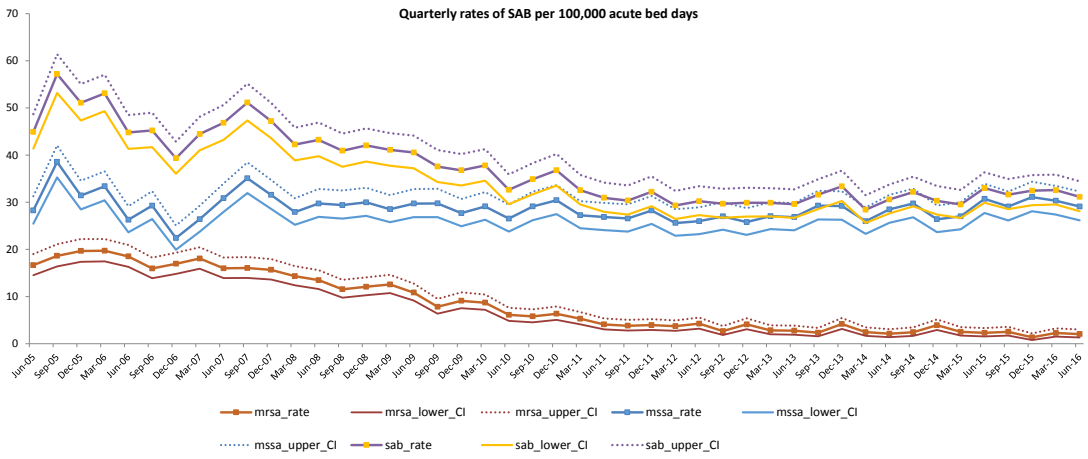



Figure 1.1: Overall quarterly SA, MRSA and MSSA bacteraemia incidence rates for Scotland (per 100,000 AOBDs) from April 2005. This figure has been taken from HPS quarterly report up to June 2016 and is available on HPS (2016a).

thermore, a teaching hospital (TH) is used by medical schools where training and clinical education are provided in teaching hospitals to medical students and current health professionals. TH is considered as a risk factor due to training staff increasing the risk of infection because they may be unaware of infection control policies. The health boards which have teaching hospitals are Great Glasgow and Clyde, Tayside, Grampian and Lothian.

### 1.5.2 CDI data

The mandatory surveillance programme of CDI in Scotland started in October 2006 and focused on the incidence of CDI in patients aged over 65 years. From April 2009, patients aged 15-64 years were added to the mandatory surveillance programme of CDI [HPS (2016a)]. As in SA bacteraemia (SAB) data, the CDI data is collected every three months and records the number of patients with CDI and the number of acute occupied bed days (AOBDs) in 15 health boards

Table 1.1: Health boards in Scotland.

| Health board | Abbreviation |
|---|---|
| Ayrshire and Arran | A.A |
| Borders | BOR |
| Dumfries and Galloway | DG |
| Fife | Fife |
| Forth Valley | FV |
| Grampian | GR |
| Greater Glasgow and Clyde | GGC |
| Highland | HI |
| Lanarkshire | LA |
| Lothian | LO |
| National Waiting Times | NWTC |
| Orkney | ORK |
| Shetland | SH |
| Tayside | TAY |
| Western Isles | WI |

in Scotland, (see Figure 1.2). Figure 1.2 showed a slightly reduction in CDI and the trend of CDI in patients aged over 65 years and CDI in patients aged 15-64 years are similar from July 2011 to June 2016. AOBDs are different for CDI in patients over 65 than for CDI in patients aged 15-64 years and is recorded by age groups. Rates of CDI are presented per 100,000 AOBDs. The percentage of acute surgical procedure relating to CDI is not recorded for CDI data. The classification of TH is the same as in the SAB data.

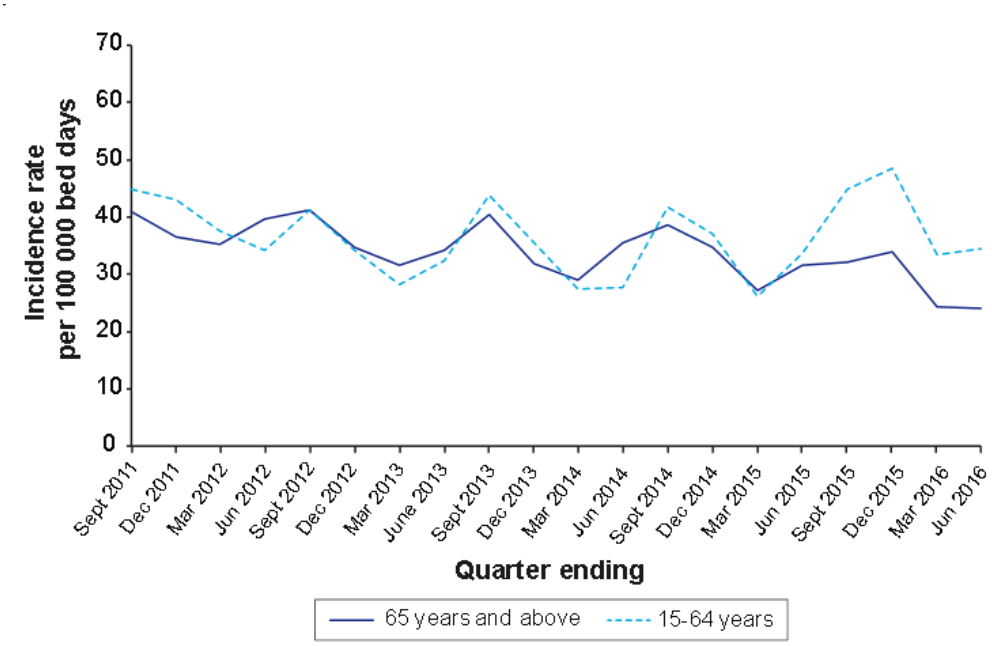



Figure 1.2: Overall quarterly CDI incidence rates for Scotland in patients over 65 years and in patients aged 15-64 years (per 100,000 AOBDs) from July 2011 to June 2016 and is available on HPS (2016a).

### 1.5.3 Healthcare interventions

Some healthcare interventions took place in NHS Scotland from 2004 to 2011. Table 1.2 explains the year, month and impacted infection (either staphylococcus aureus bacteraemia (SAB) or clostridium difficile infection (CDI) or both) of those interventions. The quarter in the Table 1.2 was coded starting from 2003 where Qu1, 2003 takes 1 and Qu1 means the months from January to March, (see Table 3.1 in Chapter 3 for explaining what months are included in each quarter).

Table 1.2: Healthcare interventions that took place in Scotland from 2004 to 2011 provided by HPS [HPS (2015b)] and ISD [ISD (2014)].

| Month | Quarter | Intervention | Targeted infection |
|---|---|---|---|
| May 2004 | 6 | NHS Scotland Code of Practice on HAI management published. | SAB and CDI |
| February 2005 | 9 | CNO letter on alcohol based hand rubs and infection control. | SAB |
| March 2005 | 9 | CNO requested that all G Grade Sisters/ Charge Nurses (Senior Charge Nurses) undertake the Cleanliness Champions Course commenced. | SAB and CDI |
| July 2005 | 11 | New IC structure in Boards, including ICM funding. | SAB and CDI |
| August 2005 | 11 | Antimicrobial Prescribing Policy and Practice in Scotland- Recommendations for good antimicrobial practice in acute hospitals. | SAB and CDI |
| January 2006 | 13 | Hand hygiene national campaign announced/launched. | SAB and CDI |
| February 2006 | 13 | Standard Infection Control Precaution model policies first launched. | SAB and CDI |
| April 2006 | 14 | HEAT targets introduced- target of 30% reduction in SAB by 2010 asked of all boards. | SAB |
| April 2006 | 14 | MRSA guidelines issued in Journal of Hospital Infection (although not Scottish initiative, widely recognised by infection control world)- screening practices changed. | SAB |

*Continued on next page*

Table 1.2 – *Continued from previous page*

| Month | Quarter | Intervention | Targeted infection |
|---|---|---|---|
| March 2007 | 17 | Scottish Patient Safety Programme (SPSP) announced. | SAB and CDI |
| December 2007 | 20 | First national hand hygiene compliance report issued. | SAB and CDI |
| January 2008 | 21 | HPS care bundles related to interventions first launched (SPSP). | SAB |
| March 2008 | 21 | Launch of QIS standards (followed by visits to Boards related to these- from 2008). | SAB and CDI |
| March 2008 | 21 | HPS CDI bundle launched | CDI |
| May 2008 | 22 | Transmission Based Precaution model policies first launched. | SAB and CDI |
| July 2008 | 23 | SAPG guidance control of 4Cs Antibiotic policy. | CDI |
| July 2008 | 23 | Cleanliness champions uptake at 2000 members of staff. | SAB and CDI |
| August 2008 | 23 | Letter to CE in Scotland outlining roles and responsibilities for HAI (performance management push). | SAB and CDI |
| August 2008 | 23 | National action plane for CDI | CDI |
| October 2008 | 24 | Credit card flyer issued (ABHR message). | SAB and CDI |
| December 2008 | 24 | National CDI guidance issued | CDI |
| January 2009 | 25 | Cabinet Secretary for Health announcement on zero tolerance to non compliance with hand hygiene. | SAB and CDI |
| January 2009 | 25 | HAIRT template introduce for hospital reporting at boards bi monthly. | SAB and CDI |
| March 2009 | 25 | Second Wave of NHS staff materials (with mandate from SGHD for compulsory placement). | SAB and CDI |
| April 2009 | 26 | Revised HEAT targets announced- CDI one introduced. | CDI |
| April 2009 | 26 | Public health act (inclusive of reporting SAB and CDI) implemented. | SAB and CDI |
| September 2009 | 27 | First HEI inspection carried out. | SAB and CDI |
| September 2009 | 27 | Second Wave of NHS staff materials reissued. | |
| December 2009 | 28 | HIIAT issued for managing outbreaks. | SAB and CDI |
| January 2010 | 29 | SAB 90 day programme launched. | SAB |
| March 2010 | 29 | MRSA Screening changes in all Boards- targeted universal in | SAB |

*Continued on next page*

Table 1.2 – *Continued from previous page*

| Month | Quarter | Intervention | Targeted infection |
|---|---|---|---|
| | | specialties interim policy. | |
| March 2011 | 33 | MRSA screening changes to CRA. | SAB |

**SAB:** Staphylococcus aureus bacteraemia, **CDI:** Clostridium difficile infection, **NHS:** National Health Service, **HAI:** Healthcare associated infection, **CNO:** Chief nursing officer, **IC:** Infection control, **ICM:** Intensive care medicine, **HEAT:** Health improvement, efficiency, access and treatment, **MRSA:** Methicillin-resistant staphylococcus aureus, **SPSP:** Scottish patients safety program, **HPS:** Health Protection Scotland, **QIS:** Quality improvement Scotland, **SAPG:** Scottish antimicrobial prescribing group, **4Cs Antibiotic:** Broad-spectrum antibacterials including clindamycin, co-amoxiclav, cephalosporins and fluoroquinolone, **CE:** Chief Executives, **ABHR:** Alcohol based hand rub, **HAIRT:** Healthcare Associated Infection Reporting Template, **SGHD:** Scottish Government Health Directorate, **HEI:** Healthcare environment inspectorate, **HIIAT:** Hospital infection incident assessment tools, **CRA:** Clinical risk assessment.

Table 1.2 presented the possible interventions which may impacted the rate of HAIs. However, the intervention may takes time to impact the rate of infection in Scotland overall or may do not impact the rate of infection at all. High awareness from people are required to implement these interventions in order to impact HAIs. It is required quick response to inform all health boards about the intervention and people should have high adherence towards implementing these intervention in a perfect manner. British medical association suggested effective work planning management is required as well as training for all temporary and permanent staff in health care systems [Raza (2011)]. Also, world health organisation recommended some essential components of effective infection prevention and control (IPC) including programmes, guidelines, education and training which identify evidence and evaluate healthcare interventions [Storr et al. (2017)]. Moreover, the real impact of some interventions is not much clear because of some interventions took place in the same month (e.g. interventions at April 2006) and some others took place in following months (e.g. interventions at February 2005 and March 2005) therefore, it is difficult to know which actual intervention had impacted the rate of infection.

# Chapter 2

# Modelling and Detecting Change in Count Data

The main aim of this research is to explore the trend of healthcare associated infections (HAI) over time and to describe the change in the rates with interventions taking place at various time points. The HAI reviewed in Chapter 1 is a count data collected regularly over time and the number of cases per risk population is reported (see Figure 1.1). Thus, Poisson regression models can be carried out for such data where the time is an explanatory variable that can be linear or as multiple covariates. There were some interventions that took place in Scotland during various periods of time thus, it is of interest to know how these interventions affect the rate of HAIs. Therefore, this chapter provides a review of the methods used in this research to investigate the change in count data. Section 2.1 includes statistical methods of modelling count data with generalized linear regression models and generalized additive models. In Section 2.2, a review of change point problem is presented. In the last Section 2.3, methods of constructing confidence intervals are described.

## 2.1 Modelling count data

### 2.1.1 Generalized linear regression models

The main use of statistical modelling in medical studies is to provide tools for description and interpretation of explanatory variables which explain the change in the response variable. It can also detect whether the relationship between an explanatory variable and a response variable is significant. Regression analyses are a common approach to illustrate the relationship among variables. In epidemiology, the response is usually a discrete variable which represents count data. Count data is modelled by generalized linear regression models (GLM).

The GLM is a generalization of a linear regression model that is used when the response variable follows an exponential family distribution. In GLM, the linear model relates to the response variable through a link function with mean $\mu$ where $g(\mu)$ =linear model. GLM can fit polynomial terms which are non-linear transformations of the original predictor. This is obtained by increasing the power of the original predictor, for example, quadratic regression has two variables; $x$ and $x^2$. This leads to a nonlinear model (polynomial regression) in the independent variable but linear in the parameters. The best order of polynomial can be identified by residual plots or by using the criteria of choosing the best fitted model such as Akaika information criterion (AIC) and the likelihood ratio test. For example, a cubic polynomial regression model was fitted to describe the trend of congenital malformations from 1999 to 2006 among two groups of patients [Agay-Shay et al. (2012)].

The term count data refers to the number of cases occurring during a period of time. Healthcare associated infection data considers the analysis of count

data where the incidence of infection is observed. The most common analysis for this type of data is Poisson regression models [Cameron and Trivedi (2013)]. A Poisson distribution is specified by one parameter which defines the mean and the variance of the distribution where they are equal [Coxe et al. (2009)]. The Poisson regression is a type of generalized linear model where the Poisson distribution is a member of the exponential family. The Poisson regression is a linear regression in the natural logarithm of the predicted count. Maximum likelihood estimates the parameters of Poisson regression and the deviance measures the accuracy for the model. A chi-square statistic tests the significance of model fit [Coxe et al. (2009)]. To assess model adequacy, the deviance residuals are plotted against the predicted outcome values [Pierce and Schafer (1986)]. Another way to assess model adequacy is to compare the observed values to predicted values of the outcome [Hilbe (2011)].

In real life applications, count data often exhibits over-dispersion which occurs when the variability of the Poisson response is significantly larger than the mean (expected value). This is due to the omission of some important independent variables that should be in the model. Another reason is the occurrence of autocorrelation between the response observations which are assumed to be independent events [Coxe et al. (2009)]. To tackle this issue, the Poisson regression model is modified and different regression models are assumed to deal with over-dispersion in count data.

The first model accounts for over-dispersion is a quasi-Poisson regression model. It includes an over-dispersion parameter which corrects the error distribution. The amount of dispersion in the model is determined by the Pearson chi-square goodness of fit test [McCullagh (1984)]. The deviance for the quasi-Poisson model is equal to the deviance of the Poisson model divided by

the over-dispersion parameter. A quasi likelihood estimator deals with over-dispersion [Davison and Tsai (1992)]. This approach of regression modelling is used to deal with over-dispersion in analysing our data of healthcare associated infections ( see Chapter 3).

Sometimes the over-dispersion occurs when there is additional heterogeneity between observations. Poisson regression does not account for that. A negative binomial regression model is then used to account for over-dispersion in the Poisson regression model [Gardner et al. (1995) and Land et al. (1996)]. The negative binomial regression model assumes a variability among observations that have the same predicted value. This variability leads to a large variance in the overall response without affecting the mean. The variance of the negative binomial model is given by $\mu + \alpha\mu^2$, where $\alpha$ is the over-dispersion parameter. If $\alpha = 0$ this gives the variance equal to the mean as in Poisson regression [Hilbe (2011)].

To investigate the presence of over-dispersion, the difference in deviances of Poisson regression model and over-dispersed models; quasi-Poisson regression model and negative binomial regression model are compared using chi-square test with one degree of freedom. If the test is significant, the over-dispersed models fit better than the Poisson regression model [Scott Long (1997)]. However, over-dispersion models cannot be compared because the quasi-Poisson model is not nested within the negative binomial model. Therefore, AIC and Bayesian information criterion (BIC) can be used alternatively to select the best fitted model [Coxe et al. (2009)].

Another common problem with count data models, including both Poisson and negative binomial models, is that empirical data often show more zer-

oes than would be expected under model assumptions and these are called zero-inflated models. Zero-inflated Poisson overcomes the issue of when the observed number of zero counts exceeds the predicted number of zero counts which leads to variability [Lambert (1992) and Cheung (2002)]. If the variability still occurs after accounting for excess zeros, then zero-inflated negative binomial is used to eliminate over-dispersion.

Hurdle models have similar ideas to deal with the high occurrence of zeros in the observed data but have one important difference in how they interpret and analyse zero counts. Zero-inflated models consider two sources of zero observations; sampling zeros (not always zero) that are part of the Poisson or negative binomial (sampling distribution) and structural zeros (always zero) that only take zero counts. In contrast, the Hurdle model considers all zeros to be structural zeros (i.e. they are not from sampling distribution) [Hu et al. (2011)]. The Hurdle model was first introduced by Mullahy (1986). Generalized Hurdle regression models have been studied by Gurmu (1998). In count data with many structural zeroes, the Hurdle Poisson models fit the data while in the case of over-dispersion the Hurdle negative binomial models are appropriate for large variability [Zuur et al. (2009)].

Another way to deal with count data is zero truncated models. These models deal with only positive integer observations (the response cannot take value of zero). If the observations are positive integer values and the mean of the response variable is small, the GLM estimated parameters and standard errors may be biased. This means that zero truncated Poisson and zero truncated negative binomial (in case of over-dispersion) are used to solve this problem [Zuur et al. (2009)].

Many studies used over-dispersion models to analyse different types of count data. In an example, a likelihood method for analysing over-dispersion in correlated count data among cluster varying covariates was investigated [Solis-Trapala and Farewell (2005)]. Saffaria et al. (2013) used over-dispersed regression models to analyse censored count data. In addition, Mian (2016) fitted a zero-inflated regression model to analyse count data with some missing values. In contrast, some authors studied the case of under-dispersion where when the variance was smaller than mean distribution, zero truncated regression model for count data with under-dispersion was used by Chou et al. (2015).

All regression models for count data can easily work with the rate of the response variable accounted for the population as dominators. The logarithm of the population can be included in the regression model as an offset variable. The offset gives fitted values which are always positive and it also allows for heterogeneity within the data [Zuur et al. (2009)].

Several studies compare regression models for count data. Differences between Poisson, mixed Poisson regressions and negative binomial regression were compared [Lawless (1987), Hutchinson and Holtman (2005) and Bugna (2015)]. Regression models accounting for over-dispersion have been compared. Quasi-Poisson versus negative binomial regression models were compared and for a given dataset, the quasi-Poisson regression model was shown to be better [Seyoum and Zewotir (2016)]. Also, the Poisson, negative binomial, zero-inflated Poisson and negative binomial, and Poisson and negative binomial hurdle models were compared by Hu et al. (2011). Zuur et al. (2009) used cod parasite data to detect the best GLM model among Poisson, quasi-Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, Poisson hurdle, and negative binomial hurdle models. The negative

binomial GLM was better than Poisson GLM according to a hypothesis testing approach. However, a frequency plot indicated zero-inflation model was preferred.

In conclusion, Poisson regression is generally used to model count data but if over-dispersion is observed, the quasi-Poisson or negative binomial are used. However, if a lot of zero counts are observed in the data, zero-inflated or hurdle models are used. Finally, if the observed count are always positive (i.e. observed counts >0), zero truncated models are used.

### 2.1.2 Modelling temporal data

A temporal data is the data which changes over time (i.e. dependent variable is a function of time). The temporal data can be yearly, quarterly and etc. Generalized linear regression models usually fit count data which change over time (i.e. count data is a function of time). This function can be polynomial regression which describe the change in the trend of data over time. The most important variable in GLM to describe change over time is the seasonal effect which should be built as a factor in the polynomial regression models. A polynomial Poisson regression was fitted to the data of birth with epilepsy and showed the effect of seasonality where the epileptic births increase in the winter [Procopio and Marriott (1998)]. The study of analysing the number of tourism arrival to Singapore from January 1989 to July 1990 used the backward selection method to choose the best order of polynomial regression and concluded that cubic polynomial multiply by seasonal indexes fit the data well during the period of study [Chu (2004)].

Moreover, the count data can be correlated over time where this can be modelled by time series analysis which deals with serial correlation to see how

sequential observations in a time series affect each other. Time series regression models explain the current data and identify the correlation structure which improve forecasting for the future. They easily modelling the seasonal effect [Box et al. (2015)]. For example, in epidemiology, if healthcare associated infection is high in one data point (quarter), the next quarter may be associated with this high rate because it is an infection which can be related to what happen before. These observations tend to be connected to each other therefore there could be correlation between quarterly data. However, the data were used in this research was found not correlated therefore polynomial regression models were used to describe the change in the rate of infection over time.

### 2.1.3 Generalized additive regression models

As means to get the best GLM model explaining the data, a high order of polynomial is usually needed. However, it is better to not use greater than three or four degrees of polynomial because more polynomial terms result in less degrees of freedom which negatively affects the precision of the parameter estimates and reduces the power of analysis. Also, the polynomial curve can over fit the model so large variance is observed [Winter and Wieling (2016)].

Rather than fitting a high order of polynomial over the full range of the predictor variable, an alternative approach is to divide the overall range of the predictor variable into $k$ regions and fit polynomial regression splines to different regions of the predictor variable. These polynomials connect smoothly through knot points. When the overall range of the predictor variable is divided into enough segments and the cubic model is fitted in each region requiring continuity at each knot point, the smooth and flexible curve will fit the data well and this is called a cubic smoothing spline model. This nonlinear method of fitting a smooth model to the response $Y$ with a single predictor $X$ is

called additive models. The independent variable can be extended to multiple predictors $X_i,\ i = 1, 2, ..., m$. The response $Y$ can be count observations and when the response $Y$ is predicted by spline functions of multiple predictors $X_i,\ i = 1, 2, ..., m$, this type of modelling is called generalized additive models (GAM) [Hastie and Tibshirani (1990) and Wood (2006)].

To fit count data using GAM, a Poisson distribution is used. The assumption of Poisson GAM is that the Poisson variable has the mean equal to the variance. The predictors are the exploratory variables or functions of them. The relation between the mean of the response and the predictors is a logarithm link which indicates that the fitted values are always $\geq 0$. In the case of over-dispersion, quasi-Poisson or negative binomial are used. The likelihood ratio test is used to compare the negative binomial GAM with its Poisson equivalent [Zuur et al. (2009)].

There are two main algorithms used in GAM to fit a smooth curve to the data. The first one uses a back fitting algorithm to estimate the smoothing parameter where it depends on a local regression smoother (LOESS) or a local polynomial regression [Hastie and Tibshirani (1990)]. The second algorithm uses spline functions (basis functions) with predictor variables where it needs more complicated mathematical methods but its more efficient and has more features [Wood (2006)]. Splines are piecewise polynomials joined together to make a smooth curve by applying continuity conditions between the pieces.

There are several types of spline functions which can be used with GAM. Since the explanatory variable is divided into many subintervals (say $k$), the polynomial function (with order $r$) is fitted in each interval. The intervals are then connected by knot points to build a piecewise polynomial regression

spline [Wood (2006)]. Specifically, when the polynomial has order $r = 3$, the explanatory variable is divided into a certain number of intervals and a cubic polynomial is fitted in each interval. The intervals are connected by knot points and the first and second derivatives are required to have a smooth connection at knot points. This basis function is called a cubic regression spline [Wood (2006)]. In order to get a cubic smoothing spline, the interval of explanatory variable is divided into a large number of segments and a cubic regression model is fitted on each segment with a smooth connection at the knot points. This is followed by minimising the penalized least squares functional in case of linear regression or the penalized likelihood Poisson regression functional to estimate the smoothing parameter to give a cubic smoothing spline [Wood (2006)]. Cyclic cubic regression spline is used with time series data where it meets the following: cubic spline function has the same value at the first and last points of the interval, the first derivative of the cubic spline function has the same value at the first and last points of the interval and the second derivative of the cubic spline function has the same value at the first and last points of the interval [Wood (2006)]. Another type of basis function is thin plate splines. This deals with estimating smooth spline functions for multiple explanatory variables and involves high order derivatives [Wood (2006)]. B-splines, p-spline, natural splines, shrinkage smoother spline and many other of regression splines are introduced in Wood (2006).

Comparisons between GLM and GAM is investigated by many authors. Guisan et al. (2002) provided an overview of GLM and GAM models in ecology. They discussed the pros and limitations of regression models and reviewed several statistical tools of model selection, diagnosis of collinearity and evaluation of interactions between explanatory variables. Also, Liu (2008) illustrated a back fitting algorithm to compare GAM and GLM where he used

GAM to explore the relationship between the dependant binary variable and independent variables. Winter and Wieling (2016) compared polynomial regression and GAM when analysing nonlinear change over time where there is correlation between nested models.

Generalized additive models are developed to cover various types of data and include more than one predictor to analyse the data. Simpsona and Anderson (2009) described serial correlation structures for time series data using GAM models. Marra and Wood (2012) used a Bayesian approach to study the properties of confidence intervals for the smooth functions of generalized additive models. Moreover, Clements et al. (2005) used generalized additive models (GAM) to predict cancer rates where they fitted GAM with two smoothing spline functions for age and time per year. However, Dukić et al. (2012) modelled meningitis disease using Poisson GAM with a smooth spline function of time and adjusted the model by adding different covariates related with climate which were reported monthly. This concluded that the presence of meningitis disease is associated with some of these covariates.

## 2.2 Change points analysis

Change points occur in many different areas where this is exemplified by change of stock price in finance and economic studies, change in the quality of the products in industry as well as climate change in environment and geological research. In cancer research, a change in mortality and morbidity rate is a crucial aspect to examine.

Change point detection problems are either off-line or on-line where the off-line approach detects the change point after the data has been completely gathered and no more data can be added to the analysis. Off-line change point

detection is usually associated with two issues in statistical inference. The first is the hypothesise tests to detect if there is any change in the observed data. The second is to estimate the number of changes and their locations. On-line change point detection is performed sequentially; while every new observation is added to the data, the change point detection method is implemented to detect the location of possible change points in the previous data.

Statisticians implement many different change point detection methods to estimate the number and location of significant change points. These approaches are chosen according to the type of data and the physical problem being modelled. Parametric approach is used when the change in parameter of distribution is being investigated while the non-parametric is used usually with experimental data where no prior information about the distribution is known. The assumptions of Bayesian approach is that a prior information about the occurrence of change point is known. These approaches can be implemented with different modelling such as change in temporal data with serial correlation, change in survival date, change in regression trend and sequential change point methods which are reviewed.

### 2.2.1 Classification of change points problem

The change points problem was first introduced by Page (1954) where the change was detected in the mean of quality control data. The methods used to detect change points differ depending on the type of data.

#### Parametric change point

Parametric change point analyses are traditional approaches assuming prior knowledge of the distribution before and after the change point with fixed unknown parameters. The likelihood function plays a major role in the detection

of change points in different scenarios. Chen and Gupta (2001) give a survey of detection change points in normal models using different methods such as, likelihood ratio technique, Bayes method, cumulative sum (CUSUM) method, Akaike information criterion (AIC), Schwarz information criterion (SIC) and the wavelet conversion method.

The exponential family distributions are considered continuous distributions which include Gaussian, Gamma, exponential and others. Siegmund (1988) used a likelihood ratio test to construct confidence intervals for the change point of independent variables from an exponential family. Loader (1992) used likelihood ratio tests to detect change points in log linear models and derived the power of the likelihood ratio test and the confidence intervals for the change point. An optimum categorization method was used by O'Brien (2004) to detect the number and location of change points. This divided the continuous explanatory variable into different partitions in order to explain the response variable from exponential family distributions. Habibi et al. (2005) introduced test statistics for detecting change points when the parameter of exponential family distributions (on one parameter) changes.

The change in parameters of discrete distributions (such as binomial and Poisson) can be detected using similar methods. Freeman (2010) detected a systematic change in parameter for a sequence of binomial variables. In addition, Ghorbanzadeh et al. (2016) considered the change occurring in the parameter of a Poisson distribution. Kihoro et al. (2017) investigated the power of the likelihood ratio tests for a change point in a binomial distribution and used neural network techniques to estimate its conditional means.

### Non-parametric change point

Non-parametric change point analyses consider no prior knowledge about the distribution (i.e. no prior information about the data or the data after the change). It is the most useful for many applications especially experimental data. Pettitt (1979) tested for a change in a sequence of observations when the initial distribution is unknown. Orváth and Kokoszka (2002) detected change points in non-parametric regression models by testing the right and left continuity of nonlinear function. Using binary data, a non-parametric empirical likelihood ratio (ELR) test was developed by Zou et al. (2007) to estimate change points. Matteson and James (2014) considered non-parametric approaches to estimate the number of change points and their positions in clustered multivariate data.

### Bayesian change point

A Bayesian approach considers prior information about where the change point occurs and integrates this prior knowledge into the model. Many authors used Bayesian methods to analyse change points in different situations. Lyazrhi (1997) introduced Bayesian criteria to estimate one change point. Diniz et al. (2003) used Bayesian inference to fit segmented linear regression with non-homogeneous error variance. Tiwari et al. (2005) developed Bayesian model selection methods to estimate the number and location of change points from joinpoint models. Ghosh et al. (2009a) developed parametric and semi-parametric Bayesian joinpoint models using continuous time method by the Dirichlet distribution. Martinez-Beneito et al. (2011) re-parameterized the Bayesian joinpoint models to be more flexible to consider count data as Poisson variables. Steward et al. (2016) fitted segmented regression using a Bayesian approach to detect change points in Bernoulli data using logistic regression models.

### Time series change point

Time series change point deals with correlated data where parametric, non parametric and Bayesian methods can be used to detect the change point. Kligienė (1977) detected change points in parameters of an auto-regression model. Zhang et al. (2002) used the multivariate delta method to estimate confidence intervals around the relative effect of response variables estimated from a segmented linear autoregressive error model and extended to generalized linear models. Western and Kleykamp (2004) used a Bayasian method to detect change points in time series. Verbesselt et al. (2010) studied breaks for additive seasonal and trend (BFAST) to detect change points in time series data. An abrupt change from time series using a non-parametric method based on Haar Wavelet and Kolmogorov-Smirnov (HWKS) statistic was detected [Qi et al. (2014)]. Interrupted time series method is dealing with correlated data to detect the change points.

### Survival analysis change point

Reliability analysis represents the issue of estimating the change point in a failure rate or hazard function. Many studies look at the change point in a hazard function. The kernel method was used to analyse the change in hazard rates in survival data [Müller and Wang (1990)]. Ghosh et al. (2009b) developed semi-parametric Bayesian joinpoint models based on hazard function. A segmented discrete time model was used to analyse time to event in fertility data [Muggeo et al. (2009)].

### Regression analysis change point

Linear regression models fit a linear trend to describe the relationship between a dependent variable and one or more independent variables. The pattern of data can change after some points which can be unknown. These change

points divide the data into more than one segment where the models of these segments are different. The first author to introduce the change point problem in regression analysis was Quandt (1958) who used likelihood ratio tests to estimate one change point. If the response variable follows a normal distribution, the change points will be detected from linear regression models. If the response variable is one of the exponential family (such as Poisson, exponential or Gamma distributions), the change point will be detected from generalized linear models with link functions.

A single linear regression model has one known or unknown change point with two segments. Many authors have studied the change point problem associated with single linear regression models. Blischke (1961) used least squares estimators to estimate joinpoint within two segments in linear regression models with assumption of continuous joinpoint. Some inferences about single linear regression models were investigated [Hinkley (1969) and Hinkley (1971)]. F-statistic was derived to estimate the change in the slope of a segmented linear regression model [Worsley (1983)]. The SIC was used with simple linear regression models to estimate the change points [Chen (1998)].

Multiple linear regression models have two or more known or unknown change points. Many studies addressed the multiple linear regression model where Ferreira (1975) used a Bayesian approach to study the change in a regression model with a known number of regimes. Kim (1994) used the likelihood ratio test for estimating change points in linear regression and studied the asymptotic behavior of the likelihood ratio test. Segmented linear regression models were examined to investigate the strategy shifts in data patterns and if there is a linear relationship between input and output variables [Luwel et al. (2001)]. Natarajan and Pednault (2002) introduced linear regression tree

methods for continuous response variables of segmented linear regression and described naive bayes tree for categorical response variables which were used in data mining applications.

Several researchers addressed applications of generalized linear regression models with a change point [Zhou et al. (2008), Huang and Lyu (2011) and Huh (2012)]. Vexler and Gurevich (2009) examined likelihood ratio tests for detecting a change in logistic regression parameters which splits the model into two segments. Park et al. (2012) used Poisson change point regression models to investigate the daily murder rates in Colombia. Further details on regression analysis change points are presented in Section 2.2.2.

#### Sequential change point

This approach refers to on-line change point analyses which are implicated in the field of quality control in industrial processes. The cumulative sum statistic (CUSUM) was introduced by Page (1954) to identify a change in the mean of a normal sequence with constant variance. However, Wu* and Tian (2005) used CUSUM to identify changes in both mean and variance of normal sequences. Many different methods; parametric or non-parametric are introduced to deal with sequential change point analysis.

There is a huge amount of literature covering the change point problem, but only a few relevant papers of segmented regression and joinpoint analysis will be reviewed in detail.

### 2.2.2 Change point in regression models

Regression analysis is an important statistical method implemented in many fields where simple linear regression models are more commonly used in sev-

eral studies. However, using linear regression models when the data pattern has changed after a specific time point makes the data fit poor. Switching linear regression model into more than one linear regression (two segments or more) introduces the statistical inference (hypothesis and estimation) of change point analysis. This analysis gives better fitted regression models to the data after the change points have been located in the regression models. The location of the change is usually called a segmented point, joinpoint or change point. The regression lines which are joined by change points are often referred to as phases, segments, stages or regimes.

Change point regression analysis includes pre specified time of change points and unknown change points. The known change points refer to segmented linear regression analysis and the unknown change points indicate joinpoint analysis.

### Segmented linear regression analysis

Segmented linear regression is a statistical method for modeling changes which is usually used when the change occurs at specific times (points) from when the intervention took place in the independent variable. The good fitted segmented linear regression model indicates the importance of the interventions in changing the level, the slope or both where those describe the pattern of data.

Wagner et al. (2002) used segmented regression analysis to evaluate policy interventions that improve the quality of medication use. Ansari et al. (2003) concluded that segmented regression analysis is a practical and robust method for measuring the impact of interventions to change. Zhang et al. (2009) calculated confidence intervals of relative changes in the response variable of segmented regression on time series data using delta and bootstrap methods

where bootstrap is better as it does not require a large sample size.

Segmented linear regression analysis was applied in several applications and gave good results. This analysis was used to investigate the association between national infection control interventions and HAI rates. Stone et al. (2012) showed that hand hygiene reduces HAIs in hospitals in England and Wales. The use of hydrogen peroxide decontamination in hospitals in Australia was shown to reduce the incidence of MRSA infections [Mitchell et al. (2014)]. The national and regional HAI campaigns which involves hand hygiene, day to day cleaning and MRSA screening impacted MRSA bacteraemia rates within acute hospitals in East Midlands in the UK [Newitt et al. (2015)]. Moreover, Pinlac et al. (2016) used segmented regression analysis to compare the trend of mortality caused by noncommunicable disease (NCD) in Philippine two years before the year of intervention and two years after the year of intervention. Additionally, segmented regression analysis was used to investigate the trend of smoking amongst children (aged 13 and 15 years) before and after the introduction of Smoke-free legislation in the UK [Katikireddi et al. (2016)]. Carter et al. (2016) used segmented regression analysis and suggested that the introduction of the family medicine groups (FMG) produced reductions in the weekly rate of avoidable visits to the emergency department (ED). The impact of the grading responsible hospitals for acute care (GRHAC) program on patients in an emergency department with myocardial infarction in Taiwan was evaluated by using a Poisson distribution on a segmented regression model with multi explanatory variables and one change point over time [Tzeng et al. (2016)]. Segmented regression models with a negative binomial distribution was used to investigate the impact of a healthcare provider educational intervention on the frequency of clostridium difficile in children [Kociolek et al. (2017)].

### Joinpoint analysis

Segmented regression models with unknown change points occasionally account for one response variable and one independent variable (segmented linear regression models) or multiple independent variables (segmented multiple regression models) which can be continues or discrete variables. The unknown joinpoints are the values of independent variables where the slope of the segmented regression model changes. Joinpoint analysis estimates the number and location of change points through partitioning variables (independent variable accounts for joinpoints) which describe the change in the response variable.

Joinpoint analysis simply detects the number and location of joinpoints which divides one exploratory variable in order to describe the response variable. There are several research studies on joinpoint analysis in different applications. The issues associated with joinpoint analysis are detecting the existence of joinpoints and counting their numbers as well as estimating the location of the joinpoints. Kim and Siegmund (1989) used likelihood ratio tests to detect one change point in segmented linear regression with change in the intercepts only or in the intercepts and slopes assuming the response variable is normally distributed. Then, Kim and Cai (1993) examined the power of the likelihood ratio test for one change point in segmented linear regression. The properties of likelihood ratio tests was discussed [Kim (1994)]. However, Jones and Dey (1995) detected the number and locations of unknown joinpoints by using a modified version of Akaike's information criterion (AIC). Kim (1996) detected one joinpoint assuming auto-correlated observations using LRT to detect change in the mean. Additionally, Chen (1998) detected the change point for single and multiple linear regression models by minimizing the Schwarz information criterion (SIC) [Schwarz et al. (1978)]. A non-parametric method

was used to detect the change point where Liu and Qian (2009) used an empirical likelihood test to detect change points in segmented linear regression models. If there is no change point assumed, the empirical likelihood test has an asymptotically Gumbel extreme value distribution.

To estimate the location of joinpoints, different methods were reviewed. Hudson (1966) assumed a technique to fit a segmented linear regression model with one estimated joinpoint occurring anywhere within the range of data. The final estimate of the location of joinpoint is obtained by looking for the minimum residual sum of squares. Bellman and Roth (1969) detected one joinpoint in a segmented linear regression model using a dynamic programming method introduced by Bellman and Dreyfus (1962). However, Lerman (1980) studied a grid search method to fit a segmented linear regression model where the estimated joinpoints occur at discrete time points, and studied statistical inference using asymptotic normality of the least squares estimators on the independent variable in segmented linear regression proved by Feder (1975). Furthermore, Stasinopoulos and Rigby (1992) detected one change point in generalized linear models using piecewise polynomial in one exploratory variable and allowing for additional terms. They also fitted the piecewise polynomial (quadratic) model and used Golden Section search to detect the change point with considering Poisson data of acquired immune deficiency syndrome (AIDS) cases in the UK. One partitioning variable (year) is considered to detect one change point in the trend of data and allow for seasonal effect as additional term. Jones and Dey (1995) estimated the location of joinpoints by nonlinear optimization using F-tests and minimizing the residual sum of squares. Julious (2001) used non-parametric bootstrap methods to estimate the location of joinpoints.

Statistical methods of detecting and estimating joinpoints in one independ-

ent variable were developed. Walkowiak and Kala (2000) used a grid search method to determine change point of two segments of nonlinear regression with smooth change. Kim et al. (2000) estimated the number of joinpoints using permutation tests from segmented linear regression models with one exploratory variable and detected the location of the joinpoints using a grid search method for continuous or discrete observations. Kim and colleagues applied joinpoint regression analysis to describe the rate of cancer where they used a grid search method to fit the segmented regression models at all possible discrete points assuming constant variance and uncorrelated errors. Permutation tests with the Bonferroni correction were used to adjust the overall significance level of $\alpha$ to determine the number of significant joinpoints. They extended the research to the situation of non-constant variance, Poisson variation and autocorrelated errors. Kim et al. (2004) used a permutation procedure to compare two segmented linear regression models with unknown joinpoints that are either identical or parallel with different intercepts. They also compared permutation tests with F-tests which detect the number of joinpoints. Fay et al. (2007) developed the technique of saving the computation time of permutation tests which they used to detect the number of joinpoints in their final joinpoint model. This technique depends on when the beginning replications indicate a large enough or small enough p-value, the truncated sequential probability ratio test boundary was used to end the re-sampling. Yu et al. (2007) developed the Hudson (1966) method which detects one joinpoint on the continuous scale to detect more than one joinpoint and compared the efficiencies of this method with a grid search method. They found Hudson's method to be better than a grid search method in detecting more than one joinpoint. Kim et al. (2008) used small samples to compare grid search and Hudson methods of fitting segmented linear regression models and detecting the location of joinpoints. They also investigated the behaviour and the robustness of asymptotic confidence inter-

vals of the joinpoint regression parameters. Czajkowski et al. (2008) compared backward and forward model selection to detect the location and number of change points in logistic joinpoint regression where with a large sample size they found that both methods are approximately similar but the forward procedure was computationally much more efficient. Kim et al. (2009) proved that the number of joinpoints selected by the permutation test is consistent under the assumption of large sample theory. The permutation procedure, Bayesian information criterion (BIC) and the generalized cross validation (GCV) were compared. Kim et al. (2014) clustered segmented linear regression models into sub joinpoint models based on variables of clustering such as age. Different methods were implemented to estimate the number of joinpoints and minimum difference worth detecting (MDWD) method for the number of clusters. Kim and Kim (2016) studied the consistency of some information based on selection criteria which was used to estimate the number of joinpoints in segmented linear regression models.

A segmented multiple regression model is a generalization of a segmented linear regression model which has been studied in different applications by many authors. This model includes one response variable which can be continuous or discrete, more than one explanatory variable and one or two partitioning variables with multiple change points. The explanatory and partitioning variables can be continuous or discrete. The change points are the values of partitioning variables where the slope of the segmented regression model changes. Liu et al. (1997) estimated the number of joinpoints using a modified Schwarz information criterion (SIC) from segmented multiple regression models. They used a model which has only one partitioning continuous variable with multiple joinpoints and independent variables which are not subject to change. Bai and Perron (1998) and Bai and Perron (2003) studied statist-

ical inference on segmented multiple regression models with one partitioning variable with multiple discrete changes and the model includes explanatory variables which have not changed over time. Kim and Kim (2008) developed Liu et al. (1997), Bai and Perron (1998) and Bai and Perron (2003) models and studied the asymptotic properties of the estimated joinpoints from segmented multiple regression (more than one explanatory variable) with more than one partitioning variable (each of them has more than one continuous or discontinuous joinpoint).

The aim of our developed method is detecting one or more change points in one partitioning variable (year) using a segmented Poisson regression model where the response variable (count/ rate) is adjusted by the categorical variable (quarterly seasonal effect), (see Chapter 5). The method is then applied to detect the interventions that had an impact on the rate of HAIs.

### Joinpoint in applications

A joinpoint software program was used to describe and detect the change in the trend of cancer mortality rates [NCISR (2017)]. This software has become a standard program in epidemiological research. Many research applications have been done in cancer research. Fernández et al. (2001) used joinpoint Poisson regression analysis and found a decline in cancer mortality in Spain between 1975 and 1998. Stracci et al. (2007) in Italy, Qiu et al. (2009) in Japan, Won and Park (2010) in Korea, Ilic and Ilic (2016) in Serbia and Mohammadi et al. (2016) in Iran described the change in cancer mortality trends using joinpoint regression analysis.

Some most recent applications in different medical research are presented. Liu et al. (2015) used joinpoint analysis to analyse primary care in China. Lee

(2016) fitted joinpoint regression at unknown times to examine the relationship between mortality and physical activity and sedentary behavior. Joinpoint analysis was used to describe the change in diabetes mortality trends in Serbia between 1991 and 2015 [Ilic and Ilic (2017)].

### 2.2.3 Detection of change point in polynomial regression with GLM models

In a change points problem, statistical inference for estimating the change point and other regression coefficients under the generalized linear model is of interest. Zhou et al. (2008) applied a similar procedure to the extended model with a linear quadratic model and adjusted the model with additional independent variables.

In GLM when the parameter is always between 0 and 1, the corresponding link function is the logit function, $g(\mu) = \log(\mu/(1-\mu))$ and the regression model under this assumption is a logistic regression model. Several studies proposed methods using logistic regression to detect change points. Pastor-Barriuso et al. (2003) introduced a method of modifying a least square algorithm to estimate change points in logistic regression where the distribution of the response variable is binary. Also, Gurevich and Vexler (2005) used a generalized maximum likelihood estimator to detect a change point in logistic regression. Zand et al. (2013) developed likelihood ratio tests and clustering methods to estimate the time of change in logistic regression. Fong et al. (2015) improved methods of estimating the coefficients of a logistic regression model with the change point variable. These methods were based on the maximum likelihood ratio test and maximum weighted scores test investigating whether the covariate (change point variable) appears in a main effect term and an interaction term.

A Poisson regression model with link function $g(\mu) = \log(\mu)$ is also used to estimate change points. Min and Park (2016) developed a Bayesian selection method of covariates in the Poisson change point regression model with both discrete and continuous variables.

Polynomial regression models have been used in the change point problem. MacNeill (1978) introduced tests based on raw regression residuals for detecting change in polynomial regression at unknown points. Jandhyala and Minogue (1993) derived a formula to test for multiple changes in the polynomial regression model. The Schwarz information criterion (SIC) was used to detect multiple change points in polynomial regression models [Tang and Fei (2004)]. Dianat and Kasaei (2010) improved polynomial regression to detect changes in sensing images. Our research uses polynomial model with Poisson regression to detect change points in the rate of infections, (see Chapter 4).

### 2.2.4 Change point detection in spline regression models

Early detection of change points in spline regression was introduced by Dathe and Müller (1980). Several authors used spline functions among GAM to detect the change points in the trend of data. Jiang (2012) used a spline function to detect degradation change points by estimating knot points. Curtis and Simpson (2014) modelled time series data in ecology with GAM by considering a cubic regression spline for the year and cyclic cubic regression spline for seasonal effect then, they used the method of finite differences to estimate period of change from the GAM model. Han et al. (2015) developed minimum operator to summation operator in terms of a smoothing parameter of partial spline model with change points to estimate change times which gives a smaller mean square error (MSE) for estimated change points. Jahren et al. (2016) used GAM with quasi-Poisson distribution to estimate the reduction in populations

of capercaillie and black grouse in 16 countries through 80 years. They detected the period of significant change in the trend of data by extracting first order derivatives. This PhD research uses GAM models to detect change point in HAI rates. It uses the idea of Curtis and Simpson (2014) for detecting change points but the GAM model is fitted with one smoothing spline function using cubic regression spline for the trend (time per year) and is adjusted with another predictor (quarterly seasonal effect), (see Chapter 7).

## 2.3 Constructing confidence intervals

In this section, some general approaches of constructing confidence intervals are reviewed. We then focus on methods of constructing confidence intervals for change points.

### 2.3.1 General approaches to confidence interval construction

A classical method of confidence interval construction is based on the asymptotic normality of the maximum likelihood estimate (MLE). For a small sample size, the robust construction of confidence intervals is derived from the asymptotic chi-square distribution of the generalized likelihood ratio test. Since classical methods for constructing confidence intervals are based on an asymptotic approximation which can be quite inaccurate in practice, alternative methods are needed. There are different approaches including bootstrap, profile likelihood and delta methods.

#### Bootstrap confidence intervals

The bootstrap method was first introduced by Efron (1979) where a random sample of size $n$ is observed from unknown distribution. Several authors have

developed bootstrap methods for constructing confidence intervals in parametric, semi-parametric and non-parametric forms. The bias corrected (BC) bootstrap confidence intervals introduced by Efron (1981) and Efron (1982) make some transformation on the parameter to be normally distributed with constant standard error. Later, he developed the BC method to bias corrected and accelerated (BCa) by considering a general assumption on the transformation of the parameter which adjusts for bias and skewness in the bootstrap distribution. This method provides reasonably narrow intervals [Efron (1987)]. Non-pivotal bootstrap methods are developed by reducing the error of the bootstrap distribution function estimate [Hall (1992)]. Several different bootstrap confidence interval methods are discussed by Efron and Tibshirani (1994) where these include student's $t$, the bootstrap-$t$, the percentile interval and the approximate bootstrap confidence interval (ABC) methods (see Efron and Tibshirani (1994) for more details). Bootstrap confidence interval methods are compared with each other and with other methods of constructing confidence intervals. DiCiccio and Efron (1996) described the theory of bootstrap confidence interval methods; BCa, bootstrap-$t$, ABC and calibration and compared them with likelihood based confidence intervals.

Regression analysis uses bootstrap confidence interval methods to construct confidence limits for the regression parameters. The bootstrap confidence interval was constructed using bias corrected and accelerated (BCa) for the slope parameter in simple linear regression with non-normal error and the result is compared with other confidence interval methods [Vos and Hudson (2003)]. The backwards selection method for a regression model was improved by bootstrapping and applied non-parametric percentile bootstrap to construct confidence intervals for each regression coefficient [Austin (2008)]. The bootstrap percentile confidence intervals were used to construct confidence levels

for the parameters of linear regression model with fuzzy response variable where the bootstrap algorithm is assessed by simulated and real data [Ferraro et al. (2013)].

Bootstrap confidence interval methods are modified to be used with GLM, GAM and nonlinear regression models and different studies have been done for these types of analysis [Wahrendorf et al. (1987), Härdle et al. (1995), Kim et al. (1999), Huet et al. (1999), Claeskens and Van Keilegom (2003), Härdle et al. (2004) and Karlsson (2009)].

### Profile likelihood confidence intervals

Profile likelihood methods reduce the log likelihood function to a function of a single parameter of interest by treating the others as inconvenience parameters and maximising the log likelihood over them. The profile likelihood confidence interval is derived from a chi-square with one degree of freedom [Venzon and Moolgavkar (1988)].

Several studies compared profile likelihood confidence intervals with other methods of confidence interval and used profile likelihood confidence intervals with various types of data. Stryhn and Christensen (2003) explained that a profile likelihood confidence interval is better than the Wald confidence interval to construct confidence intervals from correlated data. Profile likelihood confidence intervals were also used with nonlinear models [Royston et al. (2007)]. Fletcher (2008) used a simulation study that showed that the profile likelihood confidence interval was the best of three methods to construct confidence intervals for the mean of skewed data. In contrast, other methods can be better than profile likelihood confidence intervals. Wu and Hsieh (2014) developed a generalized pivotal quantities method to construct confidence intervals for the

mean of delta log normal data and stated that it is the best method compared with profile likelihood confidence intervals, bootstrap confidence intervals and other methods.

The profile likelihood confidence interval is used with generalized linear models; logistic and Poisson regressions [Pradhan and Banerjee (2008), Saha et al. (2012), Heinze et al. (2013) and Pradhan et al. (2013)].

#### Delta method of constructing confidence intervals

The delta method is a general approach for computing confidence intervals for functions of maximum likelihood estimates. It is used to estimate variance of nonlinear functions of random variables. The delta method is used with GLM to construct confidence intervals for the parameters. Confidence intervals for the ratio of proportions from logistic regression was constructed using the delta method [Oliveira et al. (1997)]. Roser and Nakano (2002) found that the delta method constructs confidence intervals for rare event data.

Other methods for constructing confidence intervals are better than the delta method. The delta method 95% confidence interval and Fieller 95% confidence interval [Fieller (1954)] are compared using time to event data [Rothmann and Tsou (2003)]. Type I error probability was approximately achieved using the Fieller approach [Read (2003) and Rothmann et al. (2003)]. Also, the delta method is compared with other methods (maximum likelihood, endpoint transformation for binary data and bootstrap methods) by using Stata program which provides two functions; **prvalue** and **prgen** in **SPost** package to construct confidence intervals for the predicted value in regression models [Xu and Long (2005) and Xu et al. (2005)].

### 2.3.2 Confidence intervals of change points

The change point problem in regression models analysis is of interest to analyse different types of data. The simplest way to construct a confidence interval for the change point when linear regression changes at an unknown point ($j$) to observe two fitted linear regressions $\hat{Y}_i = a_i + b_i X_i, \quad i = 1, 2$ where $Y_i$ are normally and independently distributed is introduced by Kastenbaum (1959). The abscissa of the point of intersection ($j$) of two lines was estimated by $\hat{j} = \frac{a_2 - a_1}{b_1 - b_2}$ and had a $t$-student distribution with $N + 4$ degrees of freedom where $N$ is the number of total observations. The statistical method produced by Mood et al. (1974) used the observed value of the parameter to construct confidence intervals for the estimated change point. The confidence interval was constructed by Piegorsch (1982) using the modified least square joinpoint estimator for segmented linear regression. Piegorsch et al. (1982) developed the maximum likelihood estimator (MLE) of the joinpoint and used the distribution of the joinpoint estimator to construct its confidence interval. The joinpoint in linear regression analysis was estimated by Kim et al. (2000) where they used Lerman (1980) approach to construct confidence intervals for the joinpoint. The Lerman (1980) approach of constructing confidence levels of joinpoints using the formula $S(X) \leq C^{\alpha}$. $S(X)$ is the function calculated from the adjusted residual sum of squares and $C^{\alpha}$ depends on likelihood ratio statistic (LRT) which follows an F distribution. This approach can be used with nonlinear segments.

The limited distribution of the estimate change point in a linear regression model is obtained by Bai (1997) and then confidence intervals are constructed. This result extended to multiple changes by Bai and Perron (1998) for the linear regression model. Elliott and Müller (2007) discussed the Bai (1997) approach where the asymptotic distribution was used to obtain the coverage rates of confidence intervals for the change points. This is not accurate when the size

of change is small and therefore confidence intervals for the change point in linear time series regressions was constructed using a UT test statistic.

A profile likelihood based confidence interval is considered for constructing a confidence interval for change points. A likelihood based method was investigated by Siegmund (1988) to construct a confidence interval for a change in mean of the linear regression model with independent normal observations. Generalization of this method has been done by Eo and Morley (2015) to a multiple regression model with time series data. They introduced inverted likelihood ratio (ILR) to calculate confidence intervals for the change points.

Bootstrapping is a common method to approximate the distribution of a change point and can be used to construct confidence intervals. In the case of independent and identically distributed errors bootstrap is used to generate random samples from adjusted error ($\hat{\varepsilon} - \text{mean}(\hat{\varepsilon})$) [Carpenter and Bithell (2000)]. A studentized confidence interval for the change point in dependent time series data was constructed by Hušková and Kirch (2010). They used a circular overlapping block bootstrap method of Hušková and Kirch (2008) and combined it with studentizing techniques to construct confidence intervals for the change point estimator. Furthermore, Bühlmann et al. (1997) suggested the sieve bootstrap to deal with serial correlation in the errors. This method was used by Chang and Perron (2016) to construct a confidence interval for the change point. In addition, Chang and Perron (2016) used the Wild bootstrap method of Liu et al. (1988) to construct confidence intervals in the case of heterogeneity.

A multivariate delta method was introduced by Casella and Berger (2002) and was used to estimate the variance of the relative effect estimate (the change

in the level and slope after intervention) in order to construct a confidence interval around relative effect estimate from a segmented linear model with autoregressive error [Zhang et al. (2002)].

Comparing different methods to construct confidence intervals of change points was investigated by Chang and Perron (2016) where they compared the performance of Bai (1997), Elliott and Müller (2007) and Eo and Morley (2015) and they found that Elliott and Müller (2007) method was the best.

## 2.4 Summary

This chapter reviewed the statistical methods of modelling count data and estimating change points and their confidence intervals. Generalized linear models (GLM) and generalized additive models (GAM) were explained to model count/rate data. Several methods of change points detection were presented and methods of estimating change points in regression models were discussed. Finally, constructing confidence intervals of change points was explained. In the next chapters we are going to use Poisson and quasi-Poisson regression with the exploratory variable of time (year) and fixed effect of seasonality to fit the best model to describe the change in the rate of HAIs (Chapter 3). Polynomial regression will then be used to estimate the change points (Chapter 4). Segmented and joinpoint regression are also used to detect the change points where the interventions impact the rate of HAIs (Chapter 5). The GAM is also modelled with fixed effect of seasonality to estimate the change points from a spline function (Chapter 7). Confidence intervals are constructed for estimated change points using bootstrapping and other methods.

# Chapter 3

# Modelling Changes in the Rate of HAIs

Since Health Protection Scotland reported a general decrease in the rates over time of healthcare associated infections (HAIs) [HPS (2014, 2015a)], more investigation to describe change in the trend of the rate is of interest. This chapter implements methods for modelling the rates of HAIs over time and describes the change in trend. Initially, statistical methods to model count data are reviewed and rates of MRSA and MSSA bacteraemias in Scotland are described. Modelling rates of clostridium difficile infection (CDI) is also analysed where these are carried out on the whole of Scotland and by each of the 15 NHS health boards. Power and sample size analysis are investigated and funnel plots are then used to compare health boards rates of HAIs. A brief summary and discussion are presented at the end of the chapter. The software **R** programming language [R Core Team (2014)] was used for all analyses.

The aims of this chapter are:

**(1)** Develop a trend model to describe changes in the rates of HAIs in Scotland using polynomial Poisson regression and quasi-Poisson regression

as appropriate for the data (Sections 3.2 and 3.3), and hence compare the trends in the target organisms, i.e. MRSA and MSSA bacteraemias and CDI.

**(2)** Consider the impact of small samples on the power of statistical models to detect changes in the rates, (Section 3.4).

**(3)** Compare the 15 NHS health boards (HB) using funnel plots for adjusted and unadjusted rates of HAIs to investigate differences in rates over HBs associated with sampling variation, (Section 3.5).

## 3.1 Statistical methods of analysing count data

The main statistical methods are reviewed to describe and model count data. First of all, Poisson regression is used for modeling count data by generalized linear models (Section 3.1.1). The case of over-dispersion (i.e. when the mean and variance of Poisson are different) is then considered (Section 3.1.2). Model fit is then assessed for the Poisson regression model using residual deviance, likelihood ratio tests, Akaika information criterion and F-tests. The Shapiro-Wilk test and Durbin-Watson test are used for checking the residuals of the model (Section 3.1.3). Model selection methods are then considered (Section 3.1.4). Finally, after fitting the best Poisson model to the rates, Byar's method is used to calculate their confidence interval (Section 3.1.5).

### 3.1.1 Poisson regression

In clinical trials and in epidemiological studies, there are situations where the outcome variable is in the form of counts. The outcome variable could be a count of rare events such as the number of cases of breast cancer occurring in a population over a certain period of time or a number of deaths due to some

disease, etc. The aim of regression analysis is to model the dependent variable $Y$ using explanatory variables $X_k,\ k = 1, 2, ..., m$.

Generalized linear models (GLMs) are widely used in data analysis. GLMs provide a flexible framework to describe how a set of explanatory variables can explain the variation in the dependent variable. The dependent variable can be continuous or discrete (integer values), and the explanatory variables can be either quantitative (covariates) or qualitative (factors). The model is assumed to have linear effects on some transformation of the dependent variable, defined by the link function, and the variable distribution can have various forms, such as Gaussian, binomial or Poisson [Cameron and Trivedi (2013)].

Poisson regression is one GLM where the response variable is a count that follows a Poisson distribution. The simplest distribution used for modeling count data, when the events being counted are somewhat rare, is the Poisson distribution. This distribution is completely characterized by one parameter called $\lambda$ which is the mean number of events and only takes positive values. If a discrete random variable $Y$ has a Poisson distribution with parameter $\lambda$ then the probability mass function can be written as:

$$Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!},\ y = 0, 1, 2, ...$$

The Poisson distribution has the property that its mean and variance are equal and it can be shown that:

$$E(Y) = Var(Y) = \lambda$$

Therefore, when the mean is estimated, the variance can also be estimated. As will be seen later, this can be quite limiting when data are over-dispersed, i.e. the variance is greater than what would be expected from a simple Poisson

distribution.

Poisson regression is a type of regression analysis used to predict counts of rare events given a set of explanatory variables [Cameron and Trivedi (2013)]. The response variable in Poisson regression is assumed to be generated from a Poisson distribution function. If the logarithm of the expected value of the response variable can be modeled by a linear combination of unknown parameters then, Poisson regression models are a type of GLM with a logarithmic link function [Dobson and Barnett (2011)]. Accordingly, assume a sample of $n$ independent observations $y_1, y_2, ..., y_n$, where each $y_i$, $i = 1, ..., n$ is an observation from Poisson random variable $Y_i$ and $Y_i \sim \text{Poisson}(\lambda_i)$, the Poisson regression model for the count data is obtained as:

$$\log(\lambda_i) = \beta_0 + \sum_{k=1}^{m} \beta_k X_{ik}, \tag{3.1}$$

where $X_{i1}, X_{i2}, ..., X_{ij}, ..., X_{im}$ are explanatory variables. In this model, increasing $X_{ij}$, $1 \leq j \leq m$ by one unit is associated with an increase of the regression coefficient $\beta_j$ in the log of the mean. This is often referred to as a Poisson log-linear model. The linear part of the GLM can consist of continuous $X$ or categorical $X$ or a mixture of both types of explanatory variables.

The Poisson distribution can be used for modeling rates (i.e. counts per unit) if the units of collection are different. If the data being investigated have different populations (or the data is collected over different amounts of time) then it would be appropriate to model a rate. The Poisson regression model for the rate is obtained as:

$$\log(\lambda_i) = \log(\text{population}_i) + \beta_0 + \sum_{k=1}^{m} \beta_k X_{ik},$$

where the log(population$_i$) is called the offset variable and has a known coefficient of one associated with it, which is needed to account for different population sizes in each period of time. The log(population$_i$) is an adjustment term and each individual response variable may have a different population value.

Maximum likelihood estimation is typically used to estimate the parameters of a Poisson regression model. To illustrate that, let $Y$ be a Poisson random variable that depends on a predictor $X$ then the likelihood function is:

$$L(\lambda; y) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!},$$

where $\lambda$ is given by Equation (3.1) and $n$ is the number of observations. Taking the logarithm of the likelihood function gives the log-likelihood function as:

$$l(\lambda; y) = \left( \sum_{i=1}^{n} y_i \log(\lambda) \right) - n\lambda,$$

and the goal is to find the values of $\lambda$ that maximize this function. Poisson regression is available in **R** programming language [R Core Team (2014)] by using the generalized linear model (**glm(., family=poisson)**) function.

### 3.1.2 Over-dispersion and quasi-Poisson regression

In generalized linear models, it is quite common for the variability of the response to exceed what is expected by the model. One of the weaknesses of Poisson regression is the assumption that the variance of a Poisson distribution is equal to the mean. When the variance is greater than predicted it is called over-dispersion. Over-dispersion exists if $Var(Y) = \phi E(Y), \quad \phi > 1$ and an over-dispersion parameter $\phi$ can be estimated by comparing actual counts ($y_i$)

to predicted counts ($\hat{y}_i$) as:

$$\phi = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i},$$

where $n$ is a sample size and $p$ is the number of predictors [Gardner et al. (1995)].

The problem of over-dispersion can be caused by characteristics of the data. For example, some important predictor variables may be needed in the model. One way to diagnose over-dispersion is to look at the deviance statistic (see Section 3.1.3). If the residual deviance greatly exceeds the residual degrees of freedom, then that is an indication of over-dispersion [Dean and Lawless (1989)].

One way of dealing with over-dispersion is to use the mean regression function and the variance function from the Poisson GLM but let the dispersion parameter $\phi$ to be estimated from the data. This strategy leads to the same coefficient estimates as the standard Poisson model but standard errors are too small as they assume a Poisson distribution is valid when it is not. This means that inferences based on them are too precise as confidence intervals are very narrow. Consequently, quasi-Poisson regression models can be used for over-dispersed count data. As the variance of a quasi-Poisson model is a linear function in the mean, large and small counts get weighted differently in quasi-Poisson regression which can adjust the standard errors.

### 3.1.3 Goodness of fit

Several measures of goodness of fit for the Poisson regression model have been proposed in the literature.

### Residual deviance

For GLMs, a statistic called the residual deviance (DV) is computed which measures how close the predicted values from the fitted model match the actual values from the raw data. The maximum likelihood function is generally used to estimate the parameters for GLMs. The likelihood function is simply the probability density computed from the observed data values with the parameters replaced by their estimates. One of the fundamental goals of statistics is to determine a simple model with as few parameters as possible. The saturated model has as many parameters as observations and hence it provides no simplification at all. However, we can compare any proposed model to the saturated model to determine how well the proposed model fits the data [Gail and Benichou (2000)]. The residual deviance is defined as:

$$dv = 2[\text{log-likelihood(saturated model)} - \text{log-likelihood(proposed model)}] \tag{3.2}$$

A common problem with Poisson regression is that the response is more variable than expected by the model. Letting $\hat{y}_i$ denote the predicted response from the Poisson model, the measure of difference between observed ($y_i$) and fitted values is the deviance. In Poisson responses the residual deviance takes the formula [Frome (1983)]:

$$dv = 2\sum y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i).$$

If the model is a good fit to the data, then the deviance should be roughly equal to the deviance degrees of freedom ($n - p - 1$) where, $p$ is the number of predictors and $n$ is the sample size. Asymptotically, the deviance follows a chi-square distribution on these degrees of freedom if the model is correctly specified. In **R**, the **glm** function calls this the residual deviance. Thus, the deviance can be used directly to test the goodness of fit of the model.

### Likelihood ratio test

The likelihood ratio test (LRT) is used to compare the fit of two models. One of them is the null hypothesis (null model) which is a special case of the other (the alternative hypothesis (alternative model)) [Hilbe (2014)]. The LRT depends on the difference between the maximum likelihood estimates of the parameters under the null hypothesis and the alternative hypothesis. Likelihood ratio tests for Poisson regression models can easily be constructed in terms of residual deviances. In general, the difference in residual deviances between two nested models has approximately a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the null and alternative models [Keeping (1962)]. Consider the full model:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + \beta_{q+1} x_{q+1} + ... + \beta_p x_p,$$

and the null hypothesis $H_0 : \beta_{q+1} = ... = \beta_p = 0$ is tested versus the alternative hypothesis that at least one of these coefficients differs from zero. If $H_0$ is true, then the variables $x_{q+1}, ..., x_p$ are redundant in the full model and can be dropped. In order to test $H_0$ in practice, the alternative model and the null model are fitted and their respective deviances compared. The test statistic is $X^2 = dv_{\text{null}} - dv_{\text{alternative}}$, which can be written as:

$$X^2 = 2[\text{log-likelihood(alternative model)} - \text{log-likelihood(null model)}]. \quad (3.3)$$

If $H_0$ is true, then the test statistic $X^2$ has an approximate chi-squared distribution (provided the sample size is sufficiently large) with degrees of freedom equal to the difference in the number of parameters between the alternative and null models $(p - q)$. If $H_0$ is false, then the test statistic tends to be too large to be considered as deriving from the chi-squared distribution on $(p - q)$ degrees of freedom and $H_0$ is rejected. If we are testing at a level of significance

$\alpha$, then we reject $H_0$ if $X^2 > \chi^2_{\alpha,p-q}$ ($\chi^2_{\alpha,p-q}$ is a critical value of the chi-squared distribution on $(p-q)$ degrees of freedom). The test statistic given by Equation (3.3) is based on the notion of a likelihood ratio test.

### Akaika Information Criterion

The Akaike information criterion (AIC) is the most important criteria used to measure the relative goodness of fit of a statistical model and is used as criteria for variable selection. In the case of Poisson regression, the AIC is defined as follows [Akaike (1973)]:

$$\text{AIC} = -2l(\hat{\beta}; y) + 2k,$$

where $k$ is the number of estimated parameters in the fitted model and $l(\hat{\beta}; y)$ is the log-likelihood function of the parameters given by data, which is obtained as:

$$l(\hat{\beta}; y) = \sum_{i=1}^{n} (y_i x_i \hat{\beta} - e^{\sum x_i \hat{\beta}}),$$

where $n$ is the number of observations. Although the log-likelihood function could be used to measure the goodness of fit, the AIC includes $k$ to adjust the number of independent variables. The best fitted model has the smallest value of AIC.

### F-test

F-test is a statistic that can be used to compare quasi-Poisson nested models and looks at the reduction in error between two models. The F statistic is obtained as:

$$F = \frac{(SSE_1 - SSE_2)/(n_2 - n_1)}{SSE_2/(N - n_2)},$$

where $SSE_1$ is the sum of squared errors of the model with fewer parameters (model 1), $SSE_2$ for the other model (model 2), $n_1$ and $n_2$ is the number of parameters in model 1 and 2 respectively, and $N$ is the sample size. If the

F statistic is larger than the critical F value, with $(n_2 - n_1, N - n_2)$ degrees of freedom, then the p-value will be smaller than the significance level ($\alpha$) and therefore the model with more parameters is better [Draper et al. (1966), Breslow (1984) and Ludden et al. (1994) ].

### Shapiro-Wilk test

The Shapiro-Wilk statistics is a test to check whether a sample came from a normally distributed population. Assuming a random sample $x_1, x_2, ..., x_n$, a small value of $W$ indicates the normality of the random sample. $W$ is the Shapiro-Wilk test obtained as:

$$W = \frac{(\sum_{i=1}^{n} \alpha_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $x_{(i)}$ are the ordered sample values, $\bar{x}$ is the sample mean and $\alpha_i$ are constants calculated from the means, variances and covariances of the order statistics of a sample of size $n$ from a normal distribution and is obtained as:

$$(\alpha_1, ..., \alpha_n) = \frac{E^T V^{-1}}{\sqrt{E^T V^{-1} V^{-1} E}},$$

where $E$ is a vector of expected values of the order statistics of random variables from standard normal distribution and $V$ is the covariance matrix of these order statistics [Shapiro and Wilk (1965)]. At a specific significance level, the null hypothesis of the Shapiro-Wilk statistics is that the population is normally distributed. The null hypothesis is rejected if the p-value is less than the significance level.

Residuals (errors) from a Poisson regression model can be positive or negative. These residuals are expected to follow a normal distribution if the counts are large. However, when the counts are small, the residuals are not likely

to follow a normal distribution because they become much more discrete values. A normal quantile-quantile (Q-Q) plot is a graphical method for assessing whether the residuals are approximately normally distributed [Thode (2002)]. If the points lie approximately around $y = x$ line, the residuals are normally distributed.

#### Durbin-Watson test

The Durbin-Watson (DW) test is a simple numerical method for checking serial correlation. The test is used to detect the existence of autocorrelation in the prediction errors from a regression analysis. The hypotheses usually considered in the Durbin-Watson test are the null hypothesis that the residuals from statistical regression analysis are uncorrelated ($H_0 : \rho = 0$) against the alternative that the residuals are autocorrelated ($H_1 : \rho > 0$) [Durbin and Watson (1950)]. The DW statistic test is:

$$dw = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2},$$

where $\varepsilon_i = y_i - \hat{y}_i$ and $y_i$ and $\hat{y}_i$ are the observed and predicted values of the response variable for individual $i$, respectively and $n$ is the sample size. The range of the Durbin-Watson statistic is always (0, 4). A value of $dw = 2$ means that there is no autocorrelation between the residuals. Values toward 4 indicate negative autocorrelation while values approaching 0 indicate positive autocorrelation.

### 3.1.4 Model selection methods

A probability model is a useful concept to understand the data, but the best fitting model is usually too complex to be described in every detail from the information available. In order to decide, among all possible exploratory variables that have been included in a multiple regression model, which are important to describe the dependent variable and should be retained, and which

one could be dropped, model selection methods are used. A full model and all subset models (nested within the full model) are defined. Model selection methods are techniques to determine which of these models should be retained. Selection of variables depends on the likelihood ratio test (see Section 3.1.3) and residual deviance (DV) (see Section 3.1.3) where the model with the smallest residual deviance is chosen.

A common issue is that there can be a large set of explanatory variables and the aim of statistical analysis is to choose the most effective predictors from the set of explanatory variables. This should then produce a parsimonious model with good predictive ability [Christensen (1996)]. There are three common methods to select the variables from the set of exploratory variables.

1. Forward regression is a simple strategy which starts from the simplest model and sequentially adds the most significant variables. The procedure of forward regression is to fit all simple regression models and consider the predictor with the lowest p-value. Variables are added to the regression equation one at a time. DV is computed then the predictor variable with the second lowest p-value is added to the regression and DV is recalculated. The process of adding more variables stops when all of the available explanatory variables have been included or when it is not possible to make a statistically significant improvement in DV using any of the variables not yet included. Therefore, all of the independent variables selected for inclusion will have a statistically significant relationship to the dependent variable [Halinski and Feldt (1970)].

2. Conversely, backward regression starts with all predictor variables in the model and non-significant variables are removed from the model one by one. The algorithm for this approach is to fit the full model with all possible predictors and remove the variable with the highest

non-significant p-value then, refit the model without this variable. The procedure can be repeated until all variables remained in the model are significant [Halinski and Feldt (1970)].

3. Another approach to select a model is stepwise regression. In stepwise regression, the independent variables are entered according to their statistical effect on describing differences in the dependent variable. Stepwise regression is designed to find the most parsimonious and accurate model which excludes variables that do not contribute to explaining variation in the response variable [Christensen (1996)]. The basic idea of the procedure combines both forward selection and backward deletion. It starts from a given model which is often the null model and proceeds by either deleting a variable already in the model or adding a variable from all possible explanatory variables. The algorithm of stepwise regression is to start like forward selection and add a new variable to the model which must have $p<0.05$ then refit the model with this variable and those already in the model and remove variables which have $p>0.05$ and keep variables which have $p<0.05$. This is then continued until there is no justifiable reason to enter or remove more variables from the model. Eventually, the final model is the parsimonious model that explains the variation in the dependent variable.

### 3.1.5 Byar's method for confidence interval of the rate

Statistically, when the rate $r$ is low and the denominator (population) at risk $n$ is large, the variability in the observed count $O$ is described by the Poisson distribution (see Section 3.1.1). A confidence interval for $O$ and $r$ is given by using Byar's method which gives very accurate approximations to the exact Poisson probabilities, even for small counts [Breslow et al. (1987)]. The 95%

confidence interval limits for the rate $r$, where $z$ is the $100(1-\alpha/2)\%$ value from the standard normal distribution, are given by [Eayres (2008)]:

$$r_{\text{lower}} = \frac{O_{\text{lower}}}{n}, \quad \text{where}, O_{\text{lower}} = O \times (1 - \frac{1}{9O} - \frac{z}{3\sqrt{O}})^3, \tag{3.4}$$

$$r_{\text{upper}} = \frac{O_{\text{upper}}}{n}, \quad \text{where}, O_{\text{upper}} = (O+1) \times (1 - \frac{1}{9(O+1)} + \frac{z}{3\sqrt{O+1}})^3. \tag{3.5}$$

## 3.2 Modelling HAIs in Scotland

This section describes the change in the rates of healthcare associated infections (HAIs) (MRSA, MSSA bacteraemias, CDI in patients aged over 65 years and CDI in patients aged 15-64 years). Models for the rate of HAIs are fitted. The data are available in Health Protection Scotland (HPS) for MRSA bacteraemia from January 2003 to December 2013 and for MSSA bacteraemia from April 2005 to December 2013 [HPS (2013)]. Additionally, HPS started the mandatory surveillance programme of clostridium difficile infection (CDI) in October 2006 and focused on the incidence of CDI in patients aged over 65 years in Scotland. From April 2009, patients aged 15-64 years were added to the mandatory surveillance programme of CDI [HPS (2015a)]. This type of data is considered as a short period of time data (i.e. short time series).

For all HAIs data, two different datasets are analysed. First, the dataset shows the total number of MRSA, MSSA bacteraemias, CDI in patients aged over 65 years and CDI in patients aged 15-64 years and related AOBDs in Scotland overall which include time ($t$) (per year) and seasonal effect ($Qu$). This dataset will be used in this section to describe the overall change in the rates of HAIs in Scotland. The second dataset shows the number of MRSA and MSSA bacteraemias, CDI in patients aged over 65 years and CDI in patients aged 15-

64 years and related AOBDs in Scotland by health board which include time ($t$) (per year), seasonal effect ($Qu$), health boards ($HB$), teaching hospital ($TH$) and acute surgical procedure (ASP) variables (see Section 3.3). Notice that the last two variables are only used in Section 3.5 for adjusting funnel plots because $TH$ describes the variable $HB$ and it is a subset of $HB$. The ASP is provided from Qu2, 2009 (i.e. April- June, 2009) for the data of MRSA and MSSA bacteraemias however, ASP is not provided for CDI data. For simplicity, time and ASP were recoded by centering time 0 at the beginning of the data and ASP in the middle of the range of ASP values (i.e. $t$ takes the values of 0, 0.25, 0.50, 0.75, 1, 1.25, ... and $asp$=ASP-25). See Table 3.1 for a description of the variables in HAIs data.

The general process to build polynomial regression model for HAIs using the first dataset (Scotland over all data) was adopted. The **glm** function in **R** was used to fit a Poisson regression and quasi-Poisson regression with a log-link function and the **log** of AOBDs as the **offset** (the denominator of the rate) which has the coefficient of 1. Model selection was done by stepwise selection and the significance level was set at 5%. The model started with the linear effect of time and seasonal effect ($Qu$) then the measurements of goodness of fit were calculated. Then the quadratic effect of time was added to the model and the LRT was calculated to compare the models. If the model with quadratic effect of time is not significant different from the model with linear effect of time, the quadratic term is omitted and then the model with linear effect of time describes the trend of HAI data. Otherwise, adding the cubic term to the quadratic model and then using LRT to compare quadratic and cubic models. If the cubic model is significant different from the quadratic, adding the quartic term to the cubic model and then comparing cubic and quartic models using LRT. The maximum power for the temporal trend is considered as 4 because of short period of time was analysed and more increase in the power of polynomial

Table 3.1: Description of variables in HAIs data.

| Variable | Variable description | baseline |
|---|---|---|
| no.MRSA | Number of cases of MRSA bacteraemia (numeric). | |
| no.MSSA | Number of cases of MSSA bacteraemia (numeric). | |
| no.CDI65 | Number of cases of CDI in patients over 65 years (numeric). | |
| no.CDI64 | Number of cases of CDI in patients aged 15-64 years (numeric). | |
| AOBDs | Number of acute occupied bed days (numeric). | |
| $t$<br>$t^2$<br>$t^3$<br>$t^4$ | Year, linear effect of time (numeric).<br>Quadratic effect of time (numeric).<br>Cubic effect of time (numeric).<br>Quartic effect of time (numeric). | |
| *Qu* | Seasonal effect (categorical).<br>Qu1 is from January to March,<br>Qu2 is from April to June,<br>Qu3 is from July to September,<br>Qu4 is from October to December. | Qu1 |
| *HB* | Health boards (categorical).<br>See Table 1.1. | GGC |
| *TH* | Teaching hospital (categorical).<br>TH1 is health boards have teaching hospital which are Greater Glasgow and Clyde, Tayside, Grampian and Lothian.<br>TH0 is health boards do not have teaching hospital which are the rest of health boards in Table 1.1. | TH0 |
| *asp*=ASP-25 | ASP is percentage of acute surgical procedure (numeric).<br>25 is approximately the average point. | |

will reduce the degree of freedom of the model fit. Moreover, the polynomial model with power 5 was found not significant different from the quartic model. The seasonal effect was kept in the model from the beginning even if it has not significant impact because previous studies explained its impact on the rate of HAIs.

### 3.2.1 Modelling for MRSA bacteraemia

Since 2003, the general trend of MRSA bacteraemia rates in Scotland has dramatically decreased over time until December 2013 and there has been a seasonal fluctuation as shown in Figure 3.1. However, the total number of MRSA bacteraemia in Scotland during Qu4, 2013 increased by 79.7% compared with the previous quarter (Qu3, 2013). There was also a 1.4% increase on the overall MRSA bacteraemia rates in Qu4, 2013 from the corresponding quarter in the previous year [HPS (2014)]. Therefore, this indicates that the rates of MRSA bacteraemia differ annually as well as seasonally.

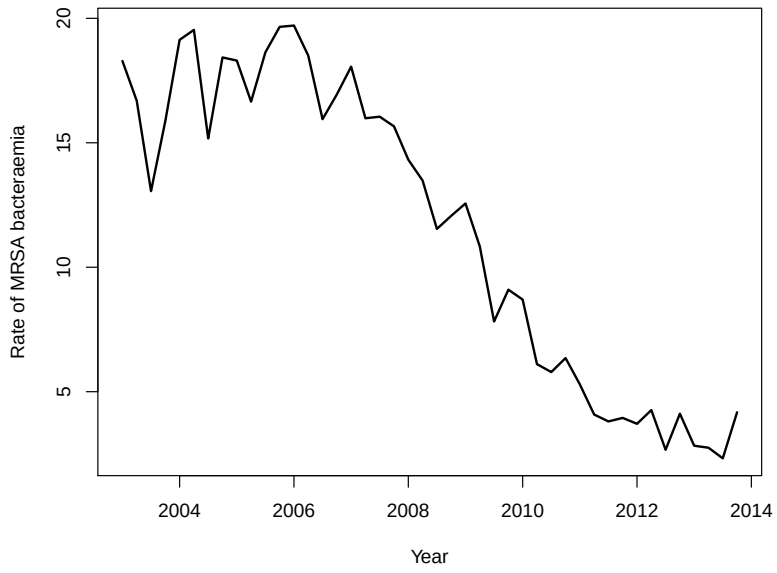



Figure 3.1: General trend of MRSA bacteraemia rates per 100,000 acute occupied bed days in Scotland from January 2003 to December 2013.

The procedure of modelling MRSA bacteraemia rates in Scotland is as follows. The observed rates in Figure 3.1 show that a quadratic model may fit the data well. Using Poisson regression with the terms of $t, t^2$ and seasonal effect, goodness of fit shows that all coefficients in the quadratic model were significant and there is no over-dispersion (p=0.067) however, the Shapiro-Wilk

normality test shows that the residuals of the model are not normal ($p < 0.001$) and Figure 3.2 shows there is a relation between residuals and predicted values. The Akaike information criterion (AIC) is 393.85 and the residual deviance is 88.271 on 38 degrees of freedom. These measures indicate lack of fit of the quadratic model.

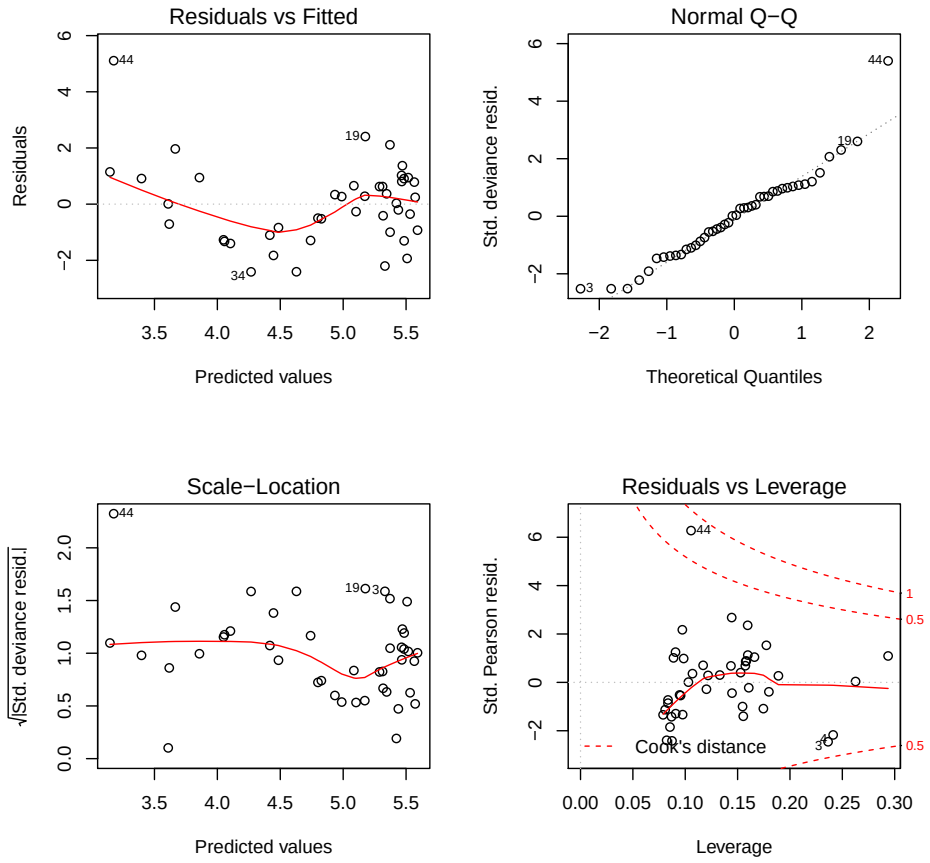



Figure 3.2: Residual plots of quadratic model of MRSA bacteraemia.

Adding a cubic term $t^3$ into the quadratic model results in the goodness of fit indicating no over-dispersion ($p=0.063$) but the residuals of the cubic model (3.6) were still not normal ($p < 0.001$). Figure 3.3 shows that there is an approximately random scatter plot of residuals against predicted values. The AIC is 374.37 and the residual deviance is 66.79 on 37 degrees of freedom. A

chi-square test indicated that the cubic model is significantly fitted better than the quadratic model ($p < 0.001$).

$$\log(\text{no.MRSA}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + Qu \tag{3.6}$$

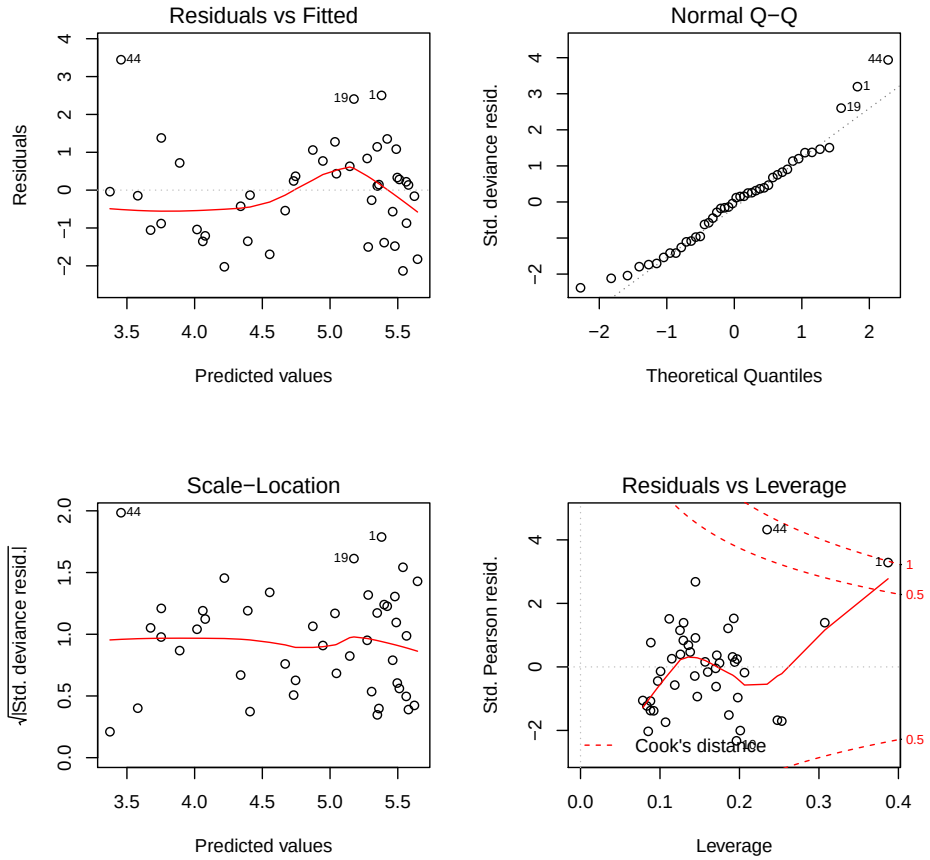



Figure 3.3: Residual plots of cubic model (3.6) of MRSA bacteraemia.

Adding a quartic term $t^4$ to the model (3.6) gives the AIC as 347.61 with residual deviance of 38.033 on 36 degrees of freedom. Figure 3.4 shows that the points in the plot of residuals against fitted values are randomly scattered with no particular pattern and the residuals and the fitted values of model (3.7) are uncorrelated. Also, the normal $Q - Q$ indicates that the residuals follow a normal distribution (Shapiro-Wilk normality test p=0.289). The chi-square test shows that the quartic model is a significantly better fit than the cubic model

(p<0.001). Over-dispersion was tested (p=0.807) and showed that the Poisson distribution fits the model well. Therefore, of those fitted models, model (3.7) is the best model to explain the change in rate of MRSA bacteraemia. Table 3.2 shows the parameter estimate and standard error of the coefficients of model (3.7).

$$\log(\text{no.MRSA}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + t^4 + Qu \qquad (3.7)$$

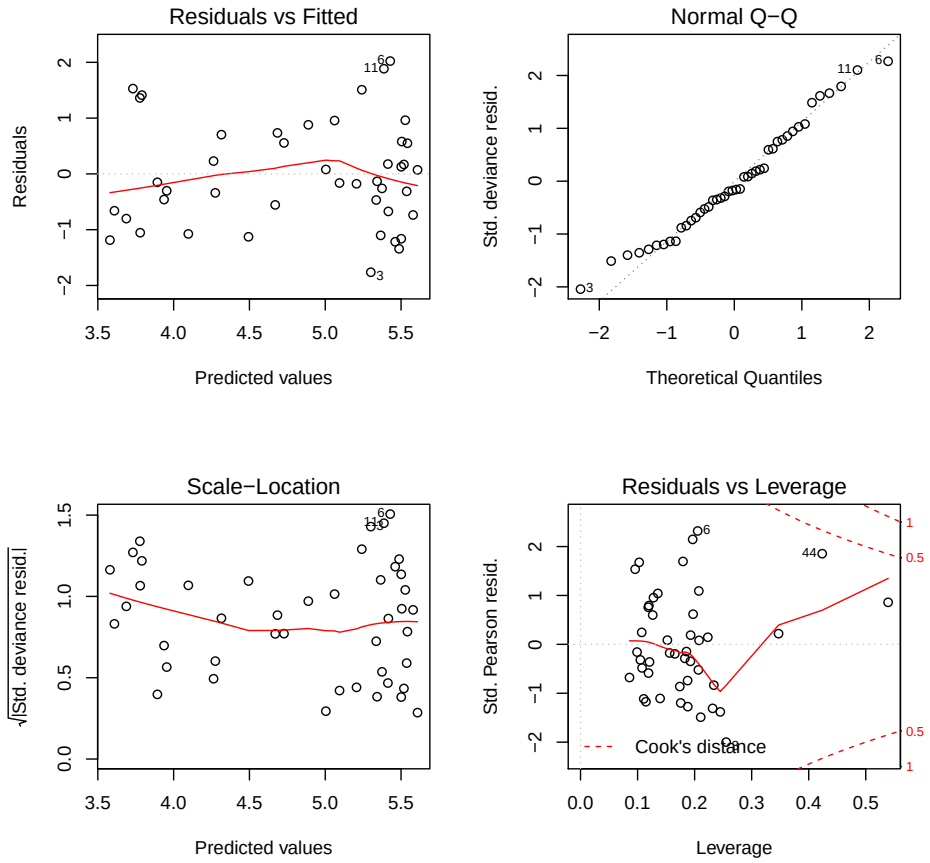



Figure 3.4: Residual plots of quartic model (3.7) of MRSA bacteraemia.

Figure 3.5 shows the best fitted model of the rates of MRSA bacteraemia with 95% confidence intervals computed using Byar's method (see Section 3.1.5). The fitted line is decreasing over time from January 2003 to December

Table 3.2: The coefficients of quartic model (3.7) of MRSA bacteraemia.

| Coefficient | Estimate | Standard error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -8.6433 | 0.0468 | -184.5711 | 0.0000* |
| $t$ | -0.0287 | 0.0643 | -0.4459 | 0.6557 |
| $t^2$ | 0.0693 | 0.0269 | 2.5710 | 0.0101* |
| $t^3$ | -0.0192 | 0.0041 | -4.6900 | 0.0000* |
| $t^4$ | 0.0011 | 0.0002 | 5.4042 | 0.0000* |
| $Qu2$ | -0.0634 | 0.0332 | -1.9076 | 0.0564 |
| $Qu3$ | -0.1709 | 0.0347 | -4.9206 | 0.0000* |
| $Qu4$ | -0.0333 | 0.0337 | -0.9863 | 0.3240 |

* : Significant coefficient at $\alpha = 0.05$ .

2013 and the rate of MRSA bacteraemia in Qu3 (July - September) is reduced over time and is significantly less than Qu1 (January - March).

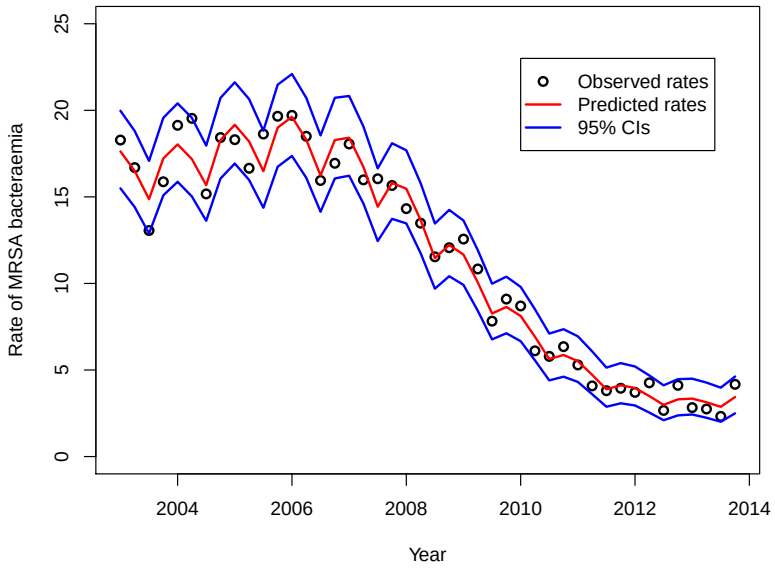



Figure 3.5: Fitted line of MRSA rate (red line) with 95% CIs (blue lines) in Scotland from January 2003 to December 2013 using quartic model (3.7).

### Prediction and residual autocorrelation

The goodness of fit of a statistical model describes how well the model fits the data. Measures of goodness of fit typically summarize the discrepancy between observed rates and predicted rates under the model. Prediction and

testing autocorrelation for residuals are two methods to detect the goodness of fit of the model.

In order to decide which model is the best for future predictions, model (3.7) was used to predict the rates of MRSA bacteraemia for five years after 2013, assuming AOBDs stay the same as in 2013. To detect the behaviour of the trend of the rates of MRSA bacteraemia before 2003, five previous years were added to the new data and AOBDs were assumed the same as the first year (i.e. 2003). Model (3.7) was then used to predict the rates of MRSA bacteraemia. Figure 3.6 shows that model (3.7) gave sudden and unexpected increasing rates of MRSA bacteraemia in the future although, MRSA bacteraemia rates were decreasing over the last ten years.

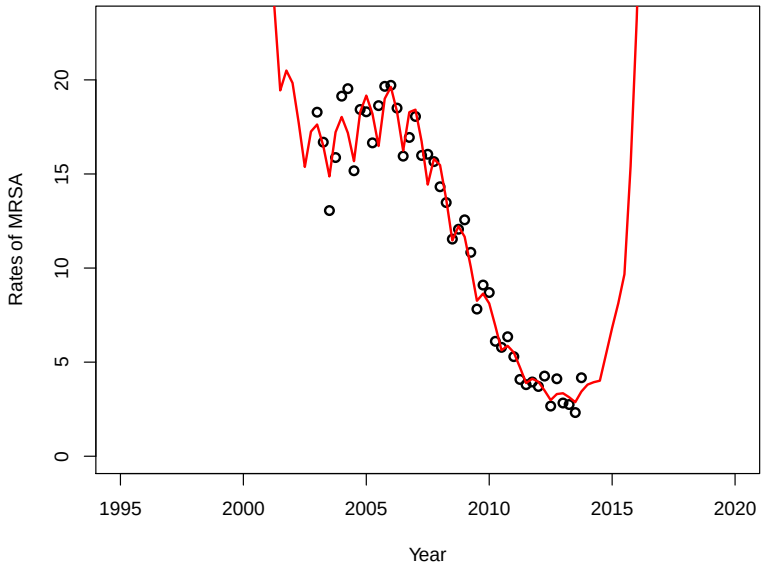



Figure 3.6: Prediction in the future using quartic model (3.7).

Using a cubic model (3.6) gives predicted rates of MRSA bacteraemia as shown in Figure 3.7. Consequently, the quartic model is the best model statistically to describe the change in the rates of MRSA bacteraemia in Scotland in

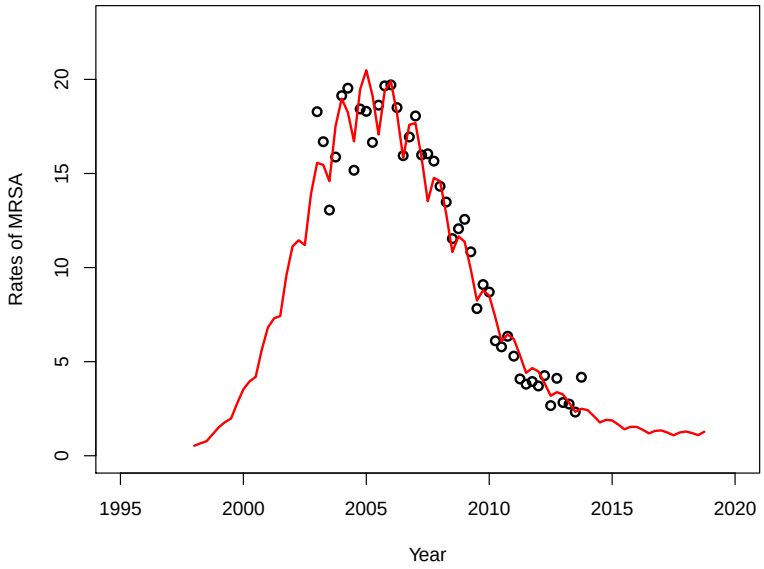



Figure 3.7: Prediction in the future using cubic model (3.6).

the period (2003- 2013) but is not good for future predictions. The cubic model gives a good fitted line in the same period and is likely to be more realistic for prediction in the future where the rates of MRSA bacteraemia will continue to decrease if nothing happens in the next five years to increase the rates of MRSA bacteraemia.

To illustrate this prediction, data from Qu1, 2003 to Qu4, 2013 was used to fit cubic and quartic models and predict ten periods of time forward (Qu1, 2014 to Qu2, 2016) where observed values for those periods are known. Comparing observed values with predicted values for those periods is shown in Table 3.3. This shows that the cubic model gives a good prediction for the rate of MRSA bacteraemia in all quarters. However, the quartic model overestimates the rate of MRSA bacteraemia.

Autocorrelation often occurs with time series data and it is the correlation between values of the data at different times. Another method to detect the

Table 3.3: Observed and predicted rates of MRSA and MSSA in different quarters from 2014 to 2016.

| Quarter | Obs Rate of MRSA | Pred Rate of MRSA | | Obs Rate of MSSA | Pred Rate of MSSA (linear) |
|---|---|---|---|---|---|
| | | Cubic | Quartic | | |
| Jan-March 2014 | 2.456 | 2.426 | 3.808 | 26.019 | 26.054 |
| April-June 2014 | 2.115 | 2.115 | 3.937 | 28.597 | 25.590 |
| July-Sep 2014 | 2.441 | 1.772 | 4.018 | 29.850 | 27.949 |
| Oct-Dec 2014 | 3.922 | 1.910 | 5.419 | 26.378 | 26.156 |
| Jan-March 2015 | 2.504 | 1.877 | 6.834 | 27.016 | 25.615 |
| April-June 2015 | 2.300 | 1.656 | 8.142 | 30.699 | 25.159 |
| July-Sep 2015 | 2.525 | 1.407 | 9.692 | 29.084 | 27.478 |
| Oct-Dec 2015 | 1.355 | 1.538 | 15.440 | 31.092 | 25.715 |
| Jan-March 2016 | 2.259 | 1.535 | 23.290 | 30.309 | 25.183 |
| April-June 2016 | 2.043 | 1.376 | 33.535 | 29.094 | 24.735 |

**Obs:** Observed, **Pred:** Predicted.

goodness of fit of the model is the Durbin-Watson (DW) test which was used to test the autocorrelation of residuals at lag1 where if the residuals are not correlated, the model fits the data well. Note that, autocorrelation at lag1 means the residual at point 1 (Qu1) is correlated with the residual at point 2 (Qu2) and Qu2 is correlated with Qu3 and so on (i.e. correlation between each quarter). While, autocorrelation at lag4 means the residual at Qu1, 2003 is correlated with residual at Qu1, 2004 and so on (i.e. correlation between the same quarters every year). By testing three different models (quadratic, cubic and quartic), $dw$=1.0037 (p<0.001) was found for the quadratic model and therefore the quadratic model does not fit the data well. There is autocorrelation at lag1 where the correlation coefficient is $\rho = 0.3135$ with 95% confidence interval (CI) (0.0183, 0.5584) which was calculated using Fisher's z-transformation [Fisher (1925)] (see Fisher's z-transformation code in Appendix A.1) as shown in Figure 3.8. On the other hand, cubic and quartic models show $dw$=1.5531 (p=0.156) and $dw$=2.1077 (p=0.719), respectively. This indicates that the cubic and quartic models had uncorrelated residuals, $\rho = 0.1920$ with 95% CI (-0.1112, 0.4625) for the cubic model and $\rho = -0.0842$ with 95% CI (-0.3717, 0.2181) for the quartic model, (see Figure 3.8).

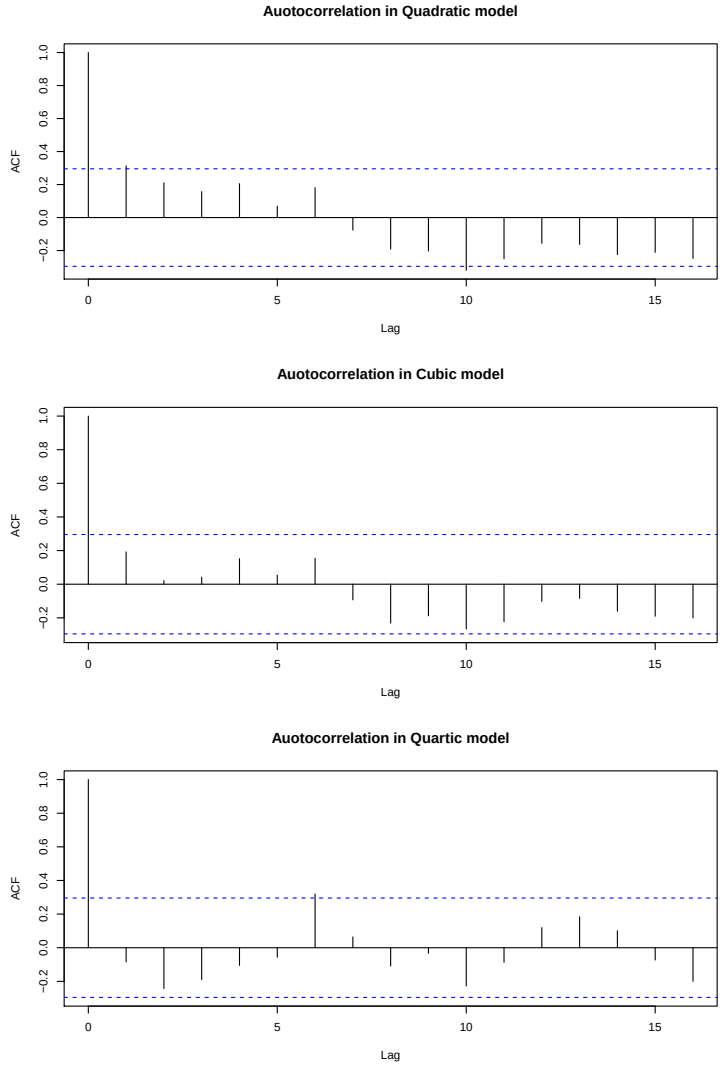



Figure 3.8: Autocorrelation functions of residuals in quadratic, cubic and quartic models.

In conclusion, the cubic model is a good model for prediction since it gave the best future predictions for MRSA bacteraemia and there was no autocorrelation for residuals at lag1. In contrast, the quartic model is significantly the best to describe and detect the change in the current data. In general, the predictions are not likely to be robust description for the future trend because if the model parameters are changed and the predicted values are increasing in the future rather than decreasing, the cubic model would be not good for pre-

diction. Therefore, the predictive ability of the models depends on the model parameters and an inherent mathematical structure is applied.

### 3.2.2 Modelling for MSSA bacteraemia

Figure 3.9 shows the general trend of MSSA bacteraemia rate in Scotland from April 2005 to December 2013. The total number of MSSA bacteraemia in Scotland during Qu4, 2013 increased by 1.6% compared with the previous quarter (Qu3, 2013) and overall MSSA bacteraemia rates for Scotland in Qu4, 2013 increased by 12.8% from the corresponding quarter in the previous year. The rate in Qu1, 2006 is higher than rate in Qu2, 2006 while, the rate in Qu1, 2007 was less than the rate in Qu2, 2007. This indicates an unclear seasonal pattern.

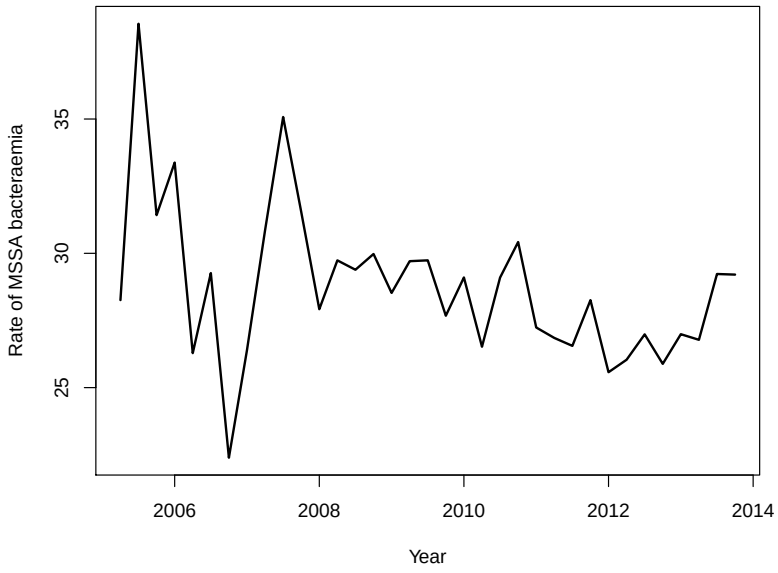



Figure 3.9: General trend of MSSA bacteraemia rates per 100,000 acute occupied bed days in Scotland from April 2005 to December 2013.

Using Poisson regression with a linear effect of time to model MSSA bacteraemia rate in Scotland observed over-dispersion (p=0.029). Adding the quadratic term to the model did not eliminate the over-dispersion (p=0.027) and

the quadratic term was not significant (p=0.259). The quadratic term was then removed and the seasonal effect was added to the model where this showed over-dispersion (p=0.033). Quasi-Poisson regression was used to fit the model with linear effect of time and seasonality. The F-test showed that the model with a seasonal effect was not significantly different from the model without seasonal effect (p=0.147) under the quasi-Poisson assumption, (see Table 3.4 for the seasonal effect and linear term coefficients of the model (3.8)). However, seasonality can explain the change in the rates of MSSA bacteraemia in individual health boards (see Section 3.3).

The best fitting model of MSSA bacteraemia is obtained in model (3.8) which has residual deviance 89.69 on 30 degrees of freedom. Figure 3.10 shows the normality of the residuals of model (3.8) where the Shapiro-Wilk normality test gave p=0.362. Figure 3.11 shows that the fitted line and 95% confidence interval of MSSA bacteraemia rates is slightly decreasing over time.

$$\log(\text{no.MSSA}) \sim \text{offset}(\log(\text{AOBDs})) + t + Qu \tag{3.8}$$

Table 3.4: The coefficients of linear model (3.8) of MSSA bacteraemia.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -8.1037 | 0.0396 | -204.7561 | 0.0000* |
| $t$ | -0.0170 | 0.0059 | -2.8597 | 0.0076* |
| $Qu2$ | -0.0140 | 0.0432 | -0.3245 | 0.7478 |
| $Qu3$ | 0.0786 | 0.0425 | 1.8488 | 0.0744 |
| $Qu4$ | 0.0159 | 0.0431 | 0.3698 | 0.7142 |

* : Significant coefficient at $\alpha = 0.05$ .

### Prediction and residual autocorrelation

Model (3.8) was used to predict the future rates of MSSA bacteraemia where it shows a decreasing trend in the future, (see Table 3.3). However, the observed rates of MSSA increased after December 2013. This indicates that linear trend

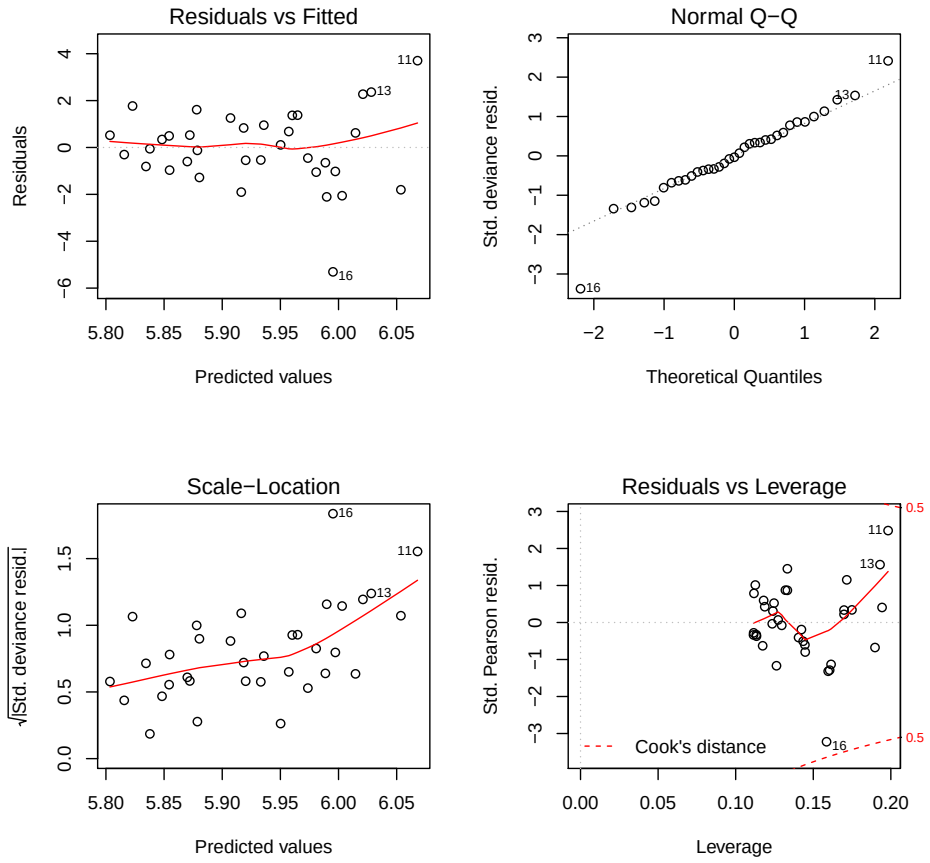



Figure 3.10: Residual plots of the linear model (3.8) of MSSA bacteraemia.

is not appropriate to predict rates in the future. In addition, observed and predicted rates show a rise in the third quarter over the first quarter which may indicate the effect of a seasonal pattern. As a result, the quadratic model is better than the linear model to explain the rates of MSSA until Qu2, 2016. There is no autocorrelation of residuals at lag1 where $dw$=1.5422 (p=0.166) and the correlation coefficient is $\rho$ = 0.2043 with 95% CI (-0.1382, 0.5033) .

### 3.2.3 Modelling CDI in patients over 65 years

Figure 3.12 shows the general trend (observed rates) of the rate of CDI in patients over 65 years in Scotland from October 2006 to September 2014 which slowly increases up to 2008 and then dramatically falls. The total number of

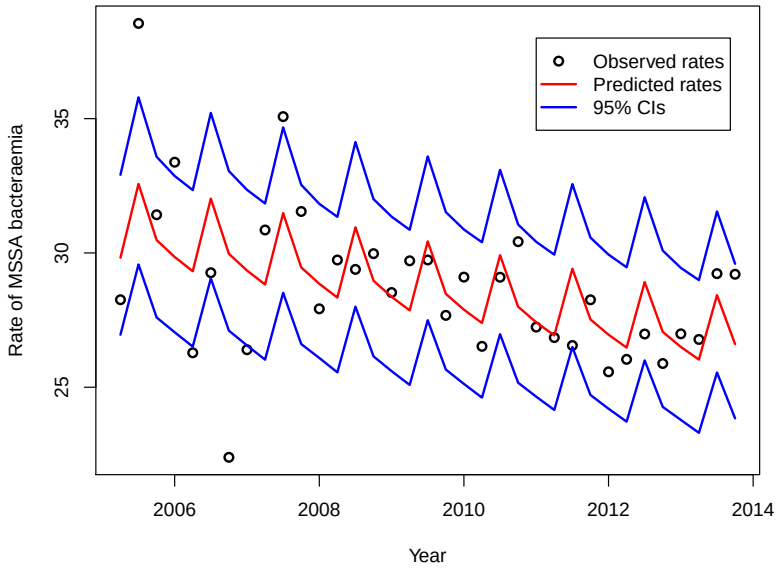



Figure 3.11: MSSA rates with 95% CIs in Scotland from April 2005 to December 2013 using model (3.8).

CDI in patients over 65 years in Scotland during Qu3, 2014 increased by 3.4% compared with the previous quarter (Qu2, 2014). There was a 4.5% reduction on the overall rate of CDI in patients over 65 years in Scotland in Qu3, 2014 in comparison to the corresponding quarter in the previous year.

Quasi-Poisson regression was used to fit the model of rates of CDI in patients over 65 years with a quartic effect of time since the over-dispersion test gave p<0.001 when Poisson regression was used. Although the seasonal effect was not significant (p=0.419), seasonality was a significant explanatory variable in previous reported studies [Rodriguez-Palacios et al. (2009) and Reil et al. (2012)]. The best model to describe the rate of CDI in patients over 65 years is obtained as:

$$\log(\text{no.CDI65}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + t^4 + Qu \qquad (3.9)$$

Model (3.9) includes a quartic effect of time and seasonality (see Table 3.5)

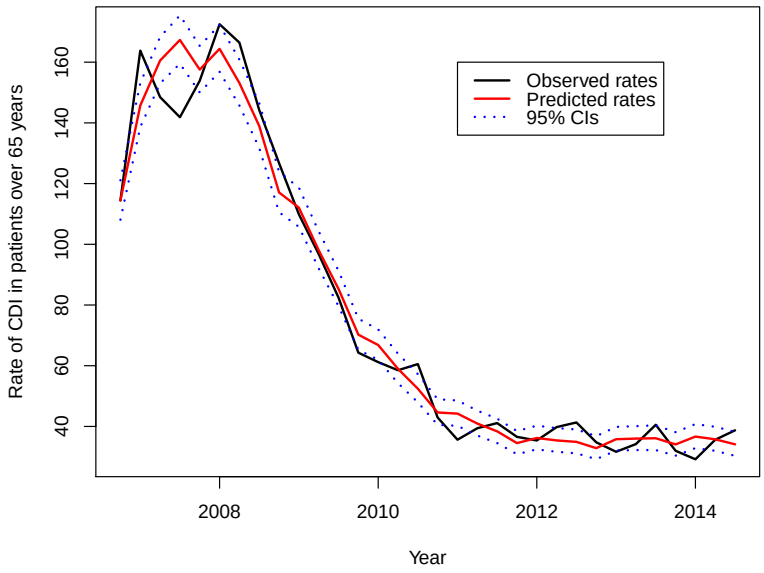



Figure 3.12: Observed and predicted rates of CDI in patients over 65 years with 95% CIs in Scotland from October 2006 to September 2014.

and has residual deviance of 194.82 on 24 degrees of freedom. Figure 3.12 shows that the fitted line of rates of CDI in patients over 65 years increased up to 2008 then decreased dramatically up to 2011 then slightly decreased until September 2014.

Table 3.5: The coefficients of model (3.9) of CDI in patients over 65 years.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -6.6958 | 0.0660 | -101.4097 | 0.0000* |
| $t$ | 0.7946 | 0.1234 | 6.4370 | 0.0000* |
| $t^2$ | -0.5693 | 0.0732 | -7.7777 | 0.0000* |
| $t^3$ | 0.0997 | 0.0154 | 6.4580 | 0.0000* |
| $t^4$ | -0.0055 | 0.0010 | -5.2030 | 0.0000* |
| $Qu2$ | -0.0057 | 0.0500 | -0.1144 | 0.9099 |
| $Qu3$ | -0.0133 | 0.0515 | -0.2573 | 0.7992 |
| $Qu4$ | -0.0769 | 0.0502 | -1.5315 | 0.1387 |

* : Significant coefficient at $\alpha = 0.05$ .

In order to evaluate model (3.9) for prediction, the rates were observed until Qu2, 2016. Table 3.6 shows a slight decrease in the observed rates of CDI in

patients aged over 65 from Qu4, 2014 to Qu2, 2016 but a dramatic reduction in the predicted rates during the same period. This indicates that model (3.9) expected the rates to fall rapidly. Model (3.9) showed uncorrelated residuals where $\rho$ = 0.2227 with 95% CI (-0.1365, 0.5302) and $dw$=1.5271 (p=0.171). This indicates that model (3.9) fit the data well.

### 3.2.4 Modelling CDI in patients aged 15-64 years

Figure 3.13 shows that the general trend of the rate of CDI in patients aged 15-64 years in Scotland from April 2009 to September 2014 is decreasing over time. The total number of CDI in patients aged 15-64 years in Scotland during Qu3, 2014 increased by 51% compared with the previous quarter (Qu2, 2014). There was a 4.1% reduction on the overall rate of CDI in patients aged 15-64 years in Scotland in Qu3, 2014 from the corresponding quarter in the previous year.

Table 3.6: Observed and predicted rates of CDI in different quarters from 2014 to 2016.

| Quarter | CDI in patients aged over 65 | | CDI in patients aged 15-64 | |
|---|---|---|---|---|
| | Obs | Pred | Obs | Pred |
| Oct-Dec 2014 | 34.7 | 29.9 | 37.1 | 28.6 |
| Jan-March 2015 | 27.1 | 29.2 | 26.3 | 24.0 |
| April-June 2015 | 31.5 | 25.2 | 33.6 | 23.0 |
| July-Sep 2015 | 32.1 | 20.7 | 44.7 | 26.5 |
| Oct-Dec 2015 | 33.9 | 15.2 | 48.4 | 18.5 |
| Jan-March 2016 | 24.3 | 12.0 | 33.4 | 14.3 |
| April-June 2016 | 24.2 | 8 | 34.4 | 12.5 |

**Obs:** Observed, **Pred:** Predicted.

Poisson regression was used to fit a model of the rates of CDI in patients aged 15-64 years. The best model to describe the rate of CDI in patients aged 15-64 years is obtained as:

$$\log(\text{no.CDI64}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + Qu \tag{3.10}$$

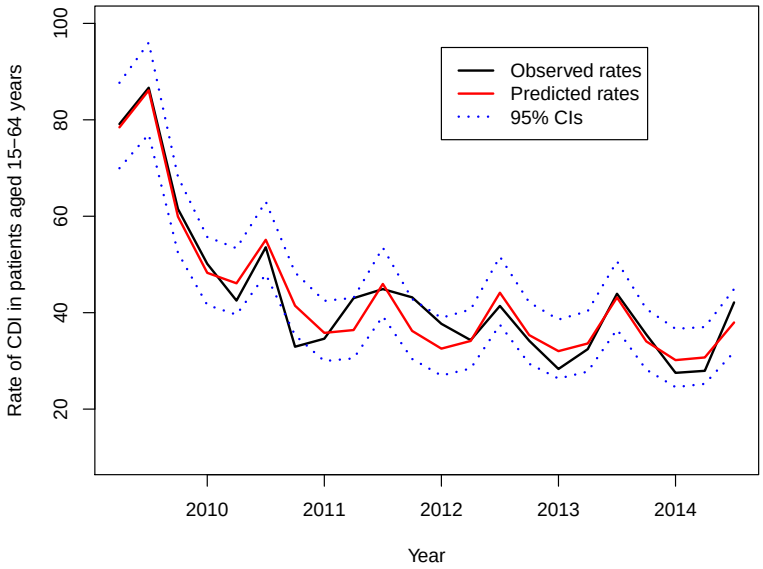



Figure 3.13: Observed and predicted rates of CDI in patients aged 15-64 years with 95% CIs in Scotland from April 2009 to September 2014.

Model (3.10) includes a cubic effect of time and seasonality and has residual deviance of 27.172 on 15 degrees of freedom and the AIC is 192.14. This model is significantly better than the model without a seasonal effect (p<0.001) and is significantly better than the model with quadratic effect of time (p<0.001). Figure 3.13 shows that the fitted line of rates of CDI in patients aged 15-64 years is decreasing over time with a clear pattern of seasonality with infection increasing in the summer, (see Table 3.7).

Table 3.7: The coefficients of CDI in patients aged 15-64 years model (3.10).

| Coefficient | Estimate | Standard error | z-value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -7.2049 | 0.0619 | -116.3860 | 0.0000* |
| $t$ | -0.7205 | 0.0879 | -8.1971 | 0.0000* |
| $t^2$ | 0.2094 | 0.0419 | 5.0037 | 0.0000* |
| $t^3$ | -0.0206 | 0.0054 | -3.8006 | 0.0001* |
| $Qu2$ | 0.0544 | 0.0525 | 1.0362 | 0.3001 |
| $Qu3$ | 0.3151 | 0.0498 | 6.3322 | 0.0000* |
| $Qu4$ | 0.0954 | 0.0534 | 1.7876 | 0.0738 |

* : Significant coefficient at $\alpha = 0.05$ .

Using model (3.10) to predict the rate of CDI in patients aged 15-64 years in the future expected a reduction in the rate. However, the observed rates from Qu4, 2014 to Qu2, 2016 increased, (see Table (3.6)). This indicates that although model (3.10) describes the data well, it is not good for prediction. Furthermore, model (3.10) fits the data with no evidence of serial correlation where $\rho$ = 0.1641 with 95% CI (-0.2765, 0.5478) and $dw$=1.6107 (p=0.349). This indicates that model (3.10) is good to describe the rates in the period of study.

### 3.2.5 Modelling CDI from April 2009

In order to understand practically the reason on changing in the trend of CDI, data from April 2009 was analysed and the pattern of the trend for CDI in patients over 65 years and CDI in patients aged 15-64 years were compared. Poisson regression was used to fit a model of the rate of CDI in patients over 65 years and the best model described the rate is obtained as:

$$\log(\text{no.CDI65}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + Qu \qquad (3.11)$$

Models (3.10) and (3.11) have same degree of time trend; up to cubic effect of time and seasonal effect is significant factor in both models. Figure 3.14 shows that the trend of CDI in patients over 65 years and CDI in patients aged 15-64 years are similar.

The pattern of CDI overall was adjusted by the age factor (over 65 and below 65) and no significant different was found in the rate of CDI in to groups of age. Also, the interaction between temporal trend and the age was found not significant. Therefore, the trend of CDI in both groups of age has same pattern and whatever intervention impacted the rate of CDI in patients over 65 years is also affects the rate of CDI in patients aged 15-64 years. The differences

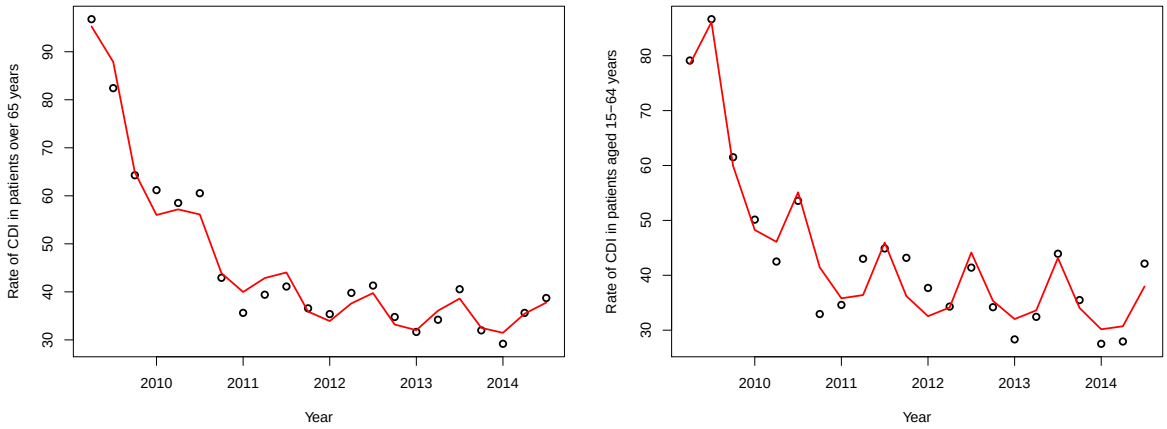



Figure 3.14: [Observed (black circles) and predicted (red line) rates of CDI in patients over 65 years (left figure) and CDI in patients aged 15-64 years (right figure) in Scotland from April 2009 to September 2014.

between models (3.9) and (3.10) are associated with different data points which was used to analyse the rate of CDI.

## 3.3 Modelling HAIs by health boards

Since the high order polynomial models are not parameterised to detect a change in the rate of HAIs in health boards, the polynomial model is adjusted by a health board factor (*HB*) to examine potential differences between the HB. Using the second dataset (health board data) and add *HB* factor to the best fitted model from Section 3.2. Then the interaction between *HB* and linear effect of time was added to the model and LRT was used to compare the models. If the model with interaction is significant, the interaction between *HB* and quadratic effect of time was added. The process of adding the interaction between *HB* and higher effect of time (up to quartic) was adding until there is no significant different from the previous model. Then the interaction between the *HB* and *Qu* was added to the last accepted model and LRT was used to assess the impact of this interaction.

### 3.3.1 MRSA bacteraemia

The previous section concluded that the quartic model (3.7) is statistically the best to describe the change in the trend of MRSA and it was used to fit individual health board rates. Figure 3.15 shows that the predicted values of Great Glasgow and Clyde (GGC) are close to the observed MRSA rates. The line for GGC is fitted well because the number of MRSA cases are large in that HB. However, the fitted lines in other health boards lie above most of the observed MRSA rates such as in AA, BOR, DG, HI and GR which indicates overprediction. In contrast, the fitted lines in Fife, FV, LO and TAY are underprediction. Therefore, the model (3.7) is not a good fit to individual health boards. The model (3.7) was developed by taking into account the health board as a factor (see model (3.12)).

$$\log(\text{no.MRSA}) \sim \text{offset}(\log(\text{AOBD})) + t + t^2 + t^3 + t^4 + Qu + HB + t \times HB + Qu \times HB \tag{3.12}$$

Fitting model (3.7) to the dataset with health boards showed over-dispersion with Poisson regression ($p<0.001$) so, quasi-Poisson regression was used to fit the model to the rates of MRSA. Adding health board (*HB*) as a factor in the model (3.7) was significant. The F-test shows that the model with health board (*HB*) is significantly different from the model without *HB* ($p<0.001$). An interaction between *HB* and time was included. The interaction arises when the effect of an explanatory variable depends on the particular level or value of another explanatory variable. The interaction between time and health board was added to the model and indicates that the rates are different in some health boards over time (i.e. at a specific time point the rate is high in some health boards but it is low in others). The model with the interaction compared to the model without the interaction was significantly better ($p<0.001$). Adding the interaction term ($t^2 \times HB$) did not make any difference from the previous

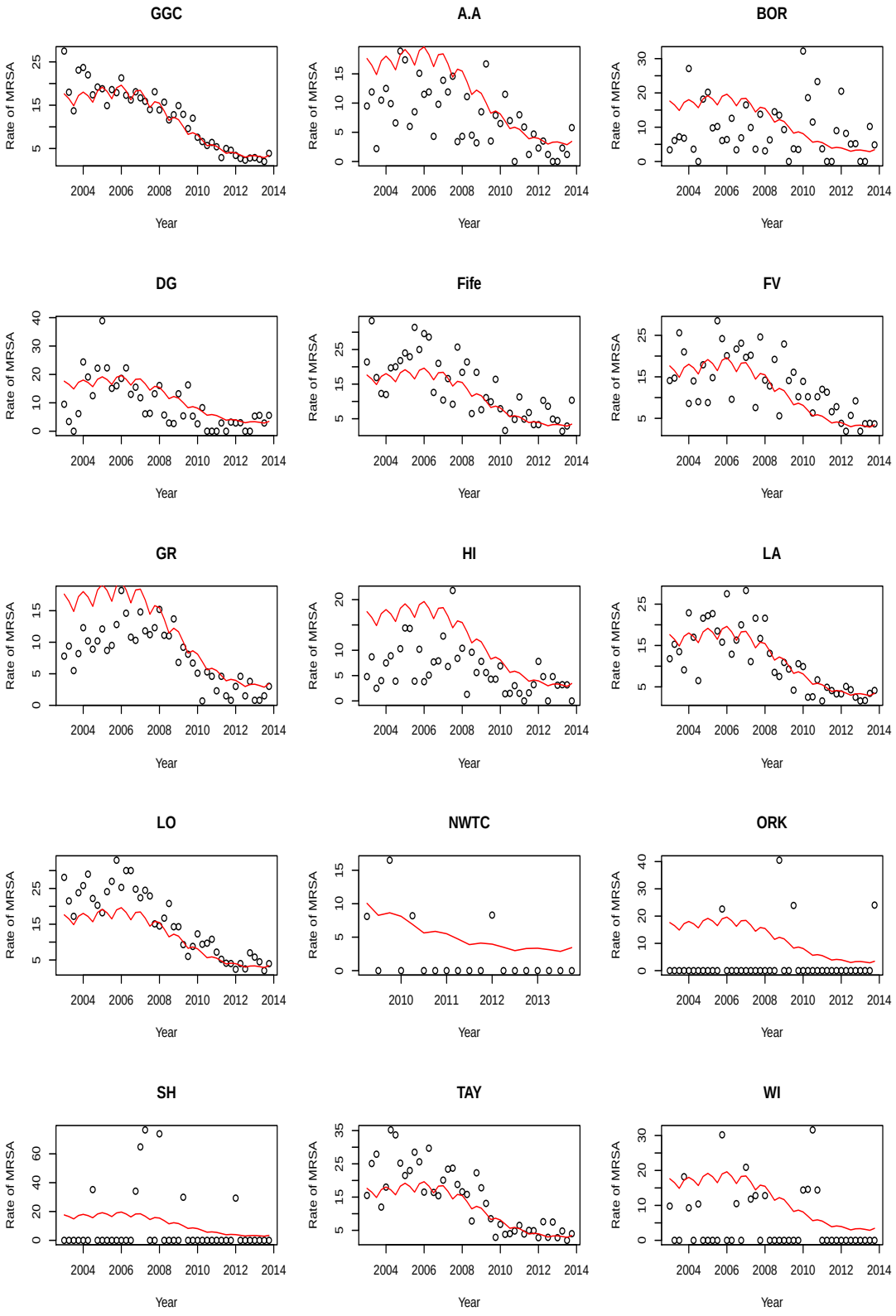



Figure 3.15: Fitted lines vs observed rates for MRSA in each health board (January 2003- December 2013) using model (3.7).

model (p=0.255). The interaction between (*HB*) and seasonal effect was investigated and the result showed that the rates of MRSA bacteraemia in some health boards differ in different quarters (p=0.038). This is simply another way of saying that the change in the rate over time is different at different health boards yearly and seasonally. Thus, model (3.12) best describes the trend of MRSA bacteraemia rate in different health boards in Scotland where the rate of MRSA decreases significantly in Qu2 and Qu3, (see Table A.1).

Figure 3.16 illustrates model (3.12) which was fitted to individual health boards. Predicted rates in GGC, GR, LO and TAY are close to the observed MRSA rates. Model (3.12) predicts the rates of MRSA bacteraemia better than model (3.7) in most individual health boards where the trends are decreasing over time. Trends of MRSA bacteraemia rates in BOR and HI were increasing slowly until about 2007- 2008 where they have decreased. In some health boards such as A.A, BOR, Fife, FV, and HI, the observed rates have more variation around fitted lines . Also, NWTC, ORK, SH and WI have a lot of zero rates which makes the model fit poorly using quasi-Poisson. Zero-inflated regression is therefore better in this case to be investigated in future work.

Table A.1 explains that although the health boards A.A, BOR, FV, GR and HI have decreasing trends over time, the interaction of health board with time shows that they have significantly different trends from GGC (see Figure 3.16). In addition, the interaction between health board and seasonal effect showed that the seasonal effect impacts the change in some individual health boards, especially in LO and TAY.

As a result, model (3.12) gave the best prediction of MRSA bacteraemia rates in most health boards. Small health boards need simple models to describe the

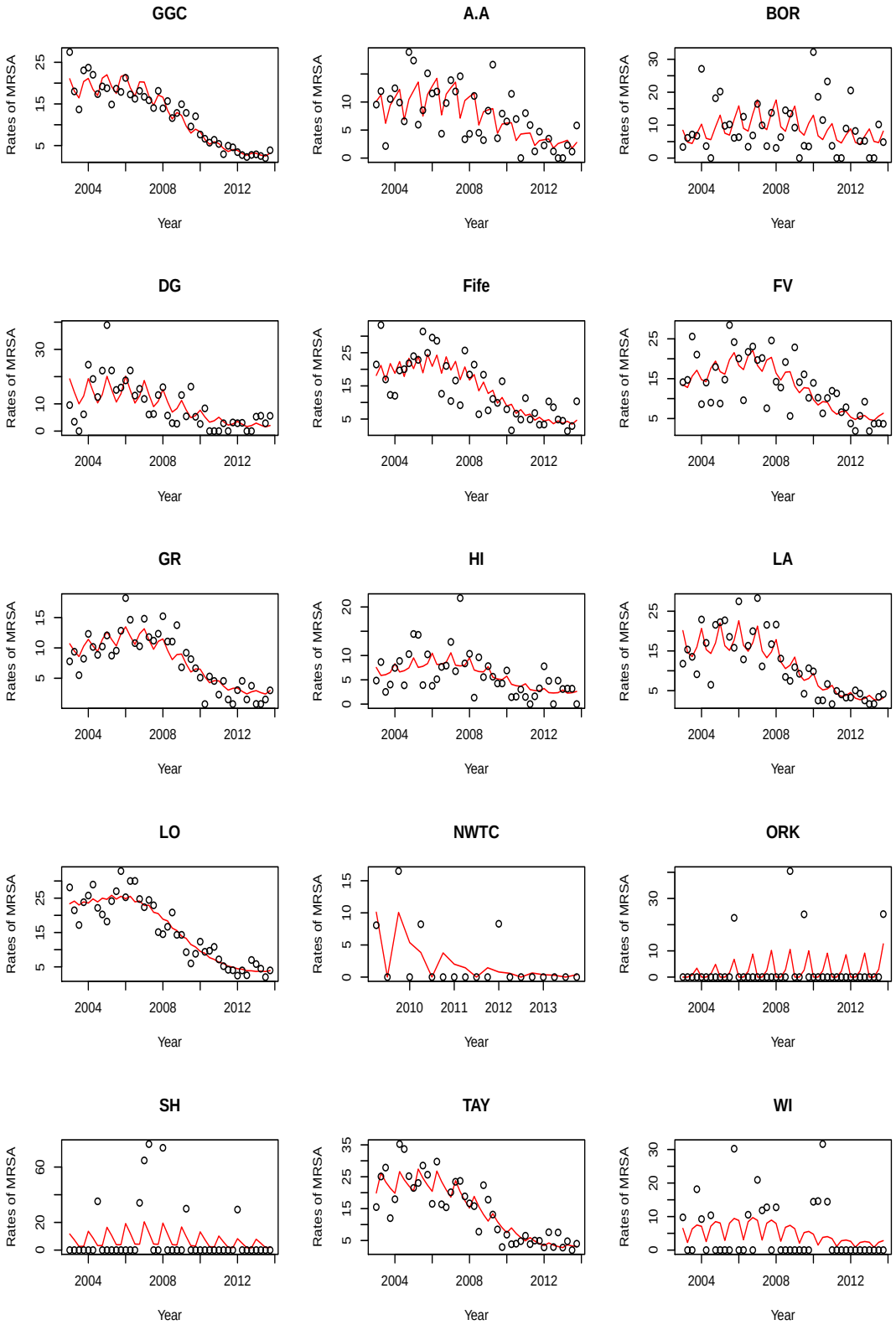



Figure 3.16: Fitted lines vs observed rates for MRSA in each health board (January 2003- December 2013) using model (3.12).

changes in the rates. Power and sample size implications are investigated in Section 3.4.

### 3.3.2 MSSA bacteraemia

As in MRSA analysis, the fitted model of MSSA (3.8) is not appropriate to fit some individual health boards. Health board was added as a factor to further develop the model. Quasi-Poisson regression was used to fit the model for the rate of MSSA bacteraemia using the health board data from April 2005 to December 2013 since over-dispersion was found ($p<0.001$). Health board (*HB*) was added to model (3.8) as a factor and the F-test shows that the model with *HB* is significantly different from the model without *HB* ($p<0.001$). The interaction between time and health board was significant when added to the model ($p<0.001$). However, the interaction of health boards and seasonal effect does not make a difference from the previous model ($p=0.913$). Model (3.13) illustrates the rates of MSSA bacteraemia in different health boards in Scotland over time.

$$\log(\text{no.MSSA}) \sim \text{offset}(\log(\text{AOBDs})) + t + Qu + HB + t \times HB \tag{3.13}$$

Figure 3.17 shows fitted lines for the rates of MSSA bacteraemia against the observed rates in individual health boards using model (3.13). In GGC, the observed MSSA bacteraemia rates are close to the predicted rates and the trend is decreasing over time. However, in all other health boards the observed rates have more variation as seen in A.A, DG and LO. Some health boards are significantly different from GGC where the trend increases over time such as the rates in BOR, HI and WI. In addition, some other health boards have a stable trend and are significantly different from GGC as witnessed in GR, LA and TAY. The interaction between time and health board shows that Fife and FV are significantly different from GGC over time. Table A.2 shows the overall

impact of the seasonal effect where Qu3 is significantly different from Qu1 in all health boards. However, there is no special effect of seasonality on some health boards over others.

In conclusion, the health board factor showed the impact of the seasonal effect on the trend in all health boards. It also showed the difference between health boards where the rates in some health boards decreased or did not change over time. This explains that the general trend from model (3.8) is linear where the rates in some health boards decreased, increased or remained stable. It also explains why the seasonal effect was kept in the model (3.8) although it was not significant.

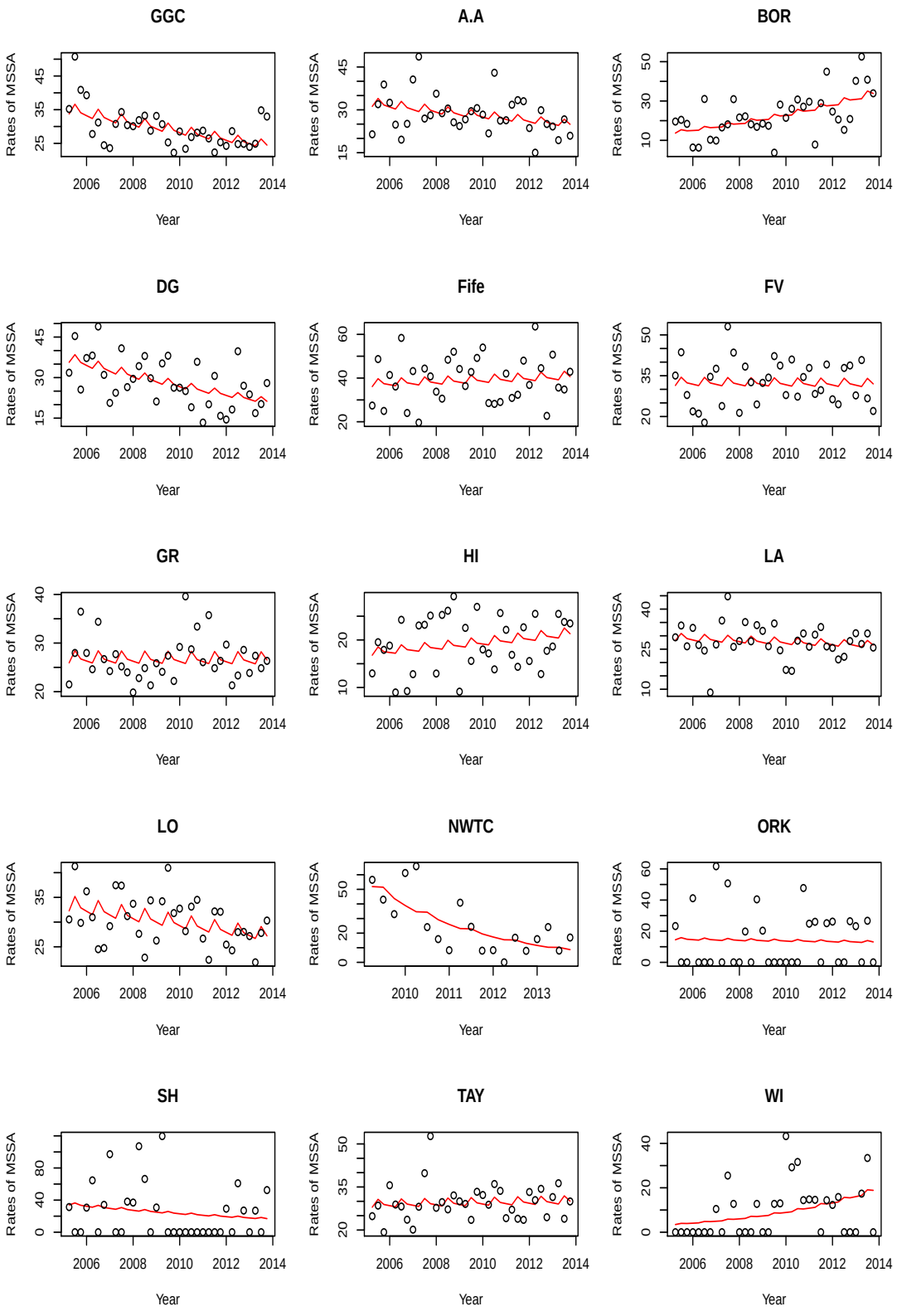



Figure 3.17: Fitted lines vs observed rates for MSSA in each health board (April 2005- December 2013) using model (3.13).

### 3.3.3 CDI in patients over 65 years

Model (3.9) is developed by adding health board to investigate the change in the rate of CDI in patients over 65 years in individual health boards. Quasi-Poisson regression is used due to over-dispersion ($p<0.001$). Model (3.14) is the best fit to the data where it describes the change in the rate of CDI in patients over 65 years in each health board. This model is significantly better than the model with an interaction up to $t^3 \times HB$ ($p<0.001$). The interaction ($Qu \times HB$) did not affect the rate of CDI in patients over 65 years ($p=0.412$) and the seasonal effect did not impact the rate between individual health boards. The general impact of seasonality shows that Qu4 is significantly different from Qu1, (see Table A.3).

$$\begin{aligned} \log(\text{no.CDI65}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + t^4 + Qu \\ + HB + t \times HB + t^2 \times HB + t^3 \times HB + t^4 \times HB. \end{aligned} \tag{3.14}$$

Figure 3.18 shows that the quartic model (3.14) is not appropriate to fit some health boards such as HI, NWTC, ORK, SH and WI. In contrast, some other health boards are fitted well with the quartic model but their models are significantly different from GGC quartic model. These health boards are A.A, DG, GR and LO. FV and TAY have a similar quartic model to GGC, (see Table A.3).

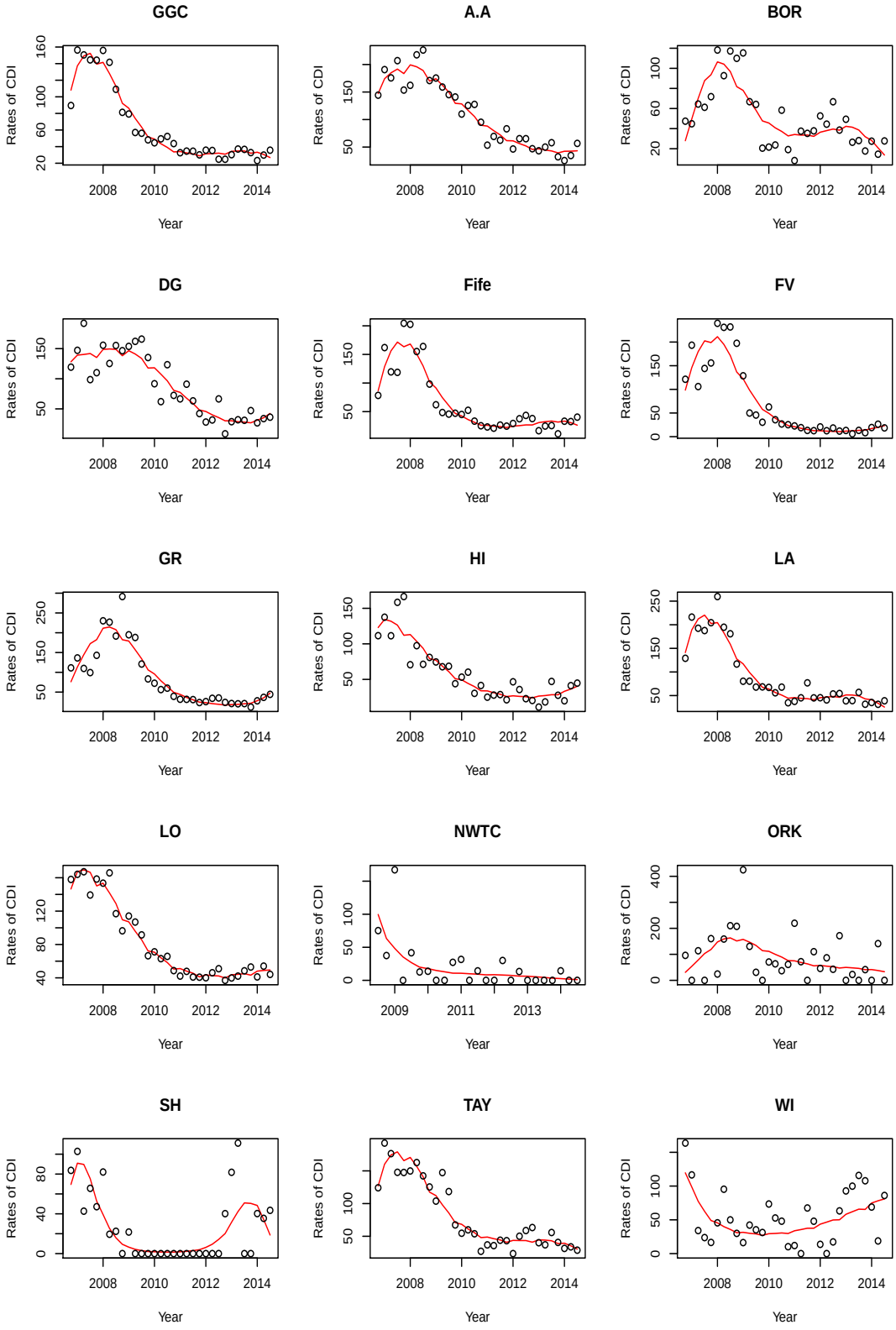



Figure 3.18: Fitted lines vs observed rates for CDI in patients over 65 years in each health board (October 2006- September 2014) using model (3.14).

### 3.3.4 CDI in patients aged 15-64 years

Although model (3.10) of CDI in patients aged 15-64 years was fitted using Poisson regression, over-dispersion was observed ($p<0.001$) when adjusting the model by health board and therefore quasi-Poisson regression was used. Model (3.15) is the best fit to the health board data and this model is significantly better than the model without the interaction between time and health board ($p<0.001$). The general impact of seasonality shows that Qu3 is significantly different from Qu1 where the infection is higher in summer than it is in winter, (see Table A.4). In addition, variables $t^2 \times HB$ and $Qu \times HB$ did not affect the change in the rate of CDI in patients aged 15-64 years in individual health boards.

$$\log(\text{no.CDI64}) \sim \text{offset}(\log(\text{AOBDs})) + t + t^2 + t^3 + Qu + HB + t \times HB \quad (3.15)$$

Figure 3.19 illustrates that model (3.15) is not appropriate for some health boards such as NWTC, ORK, SH and WI which have a lot of zero rates since quasi-Poisson was used to fit the data. Figure 3.19 shows that some health boards have a significantly different pattern to GGC where A.A, GR and TAY decrease over time but FV has an increasing trend over time. It also shows that LO has the same pattern to GGC but the data fit model (3.15) significantly better than GGC. The health board DG has a significantly different model from GGC. Other health boards have similar patterns to GGC such as Fife and HI, (see Table A.4).

In conclusion, although the fitted model of CDI rates from Scotland overall describes the data well, the pattern of CDI rates is different for individual health boards. The trend in some individual health boards is different from the general trend of Scotland, especially for small health boards.

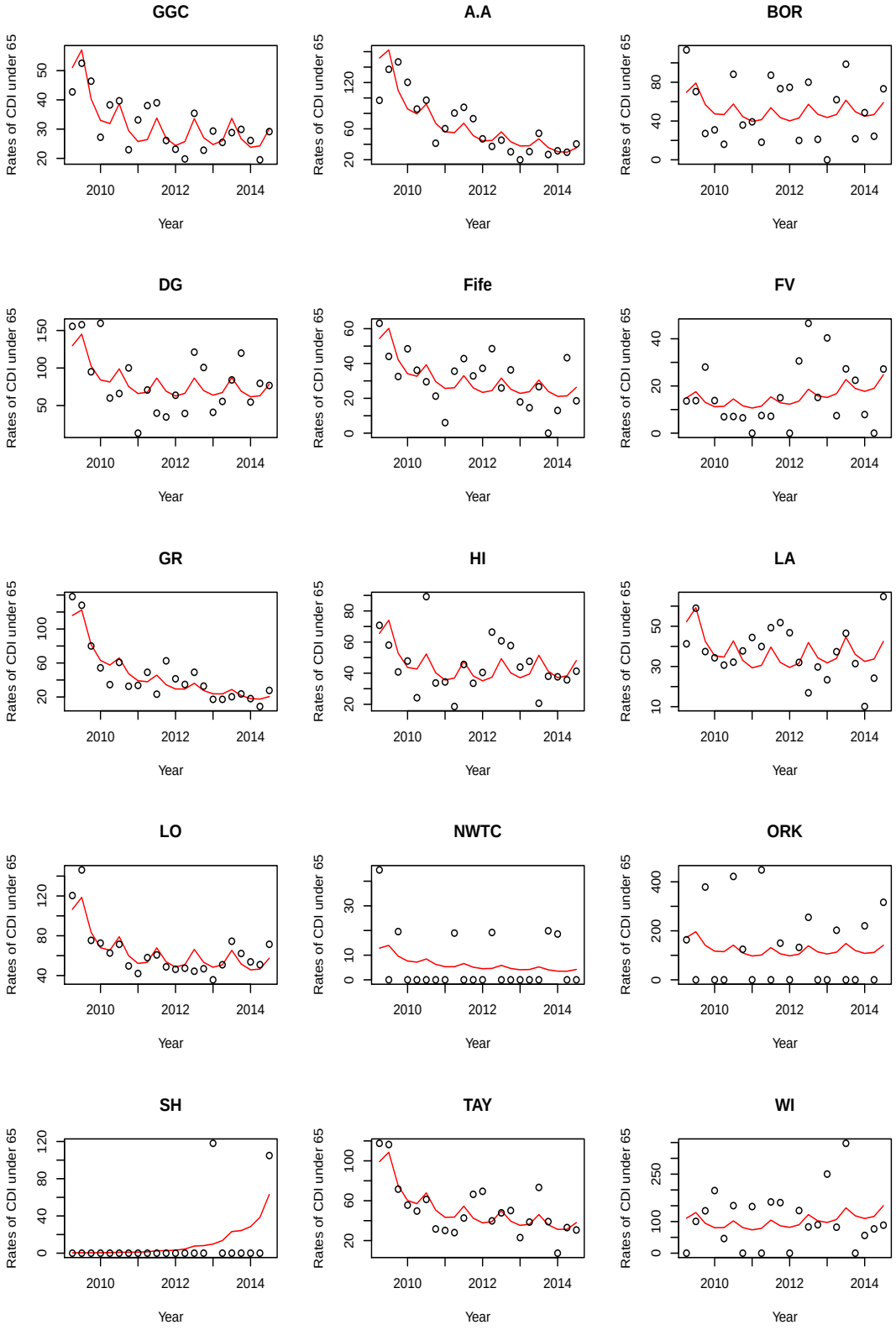



Figure 3.19: Fitted lines vs observed rates for CDI in patients aged 15-64 years in each health board (April 2009- September 2014) using model (3.15).

## 3.4 Power and sample size analysis

Since the polynomial models for MRSA (3.7), CDI in patients over 65 years (3.9) and CDI in patients aged 15-64 years (3.10) are often not suitable to fit data of some individual health boards, power and sample size are investigated.

### 3.4.1 Power and sample size test

Power analysis is an important consideration in research. The size of a sample, effect size, significance level and power are the components of statistical power and sample size analysis where if three are known, the fourth can be determined. Sample size is the number of units taken from the population and is the most important component affecting the statistical power. Effect size is a quantitative value to measure the strength of an effect such as the coefficients of regression models or the correlation between two variables. The significance level ($\alpha$) is the probability of rejecting the null hypothesis when it is true; $\alpha$ = 0.05 or 0.01 are commonly used. Given a true effect, the power ($1-\beta$) is the probability of detecting that effect and is the probability to reject null hypothesis when it is false. Ideally power should be in the range of 80% - 95% [Park (2008)].

#### Poisson ratio Test

Power analysis is used to compare two Poisson rates in order to detect the power and sample size under a significance level of 5%. Consider two independent Poisson rates obtained from different sample sizes, $X_i \sim \text{Poisson}(\lambda_i)$, $\lambda_i = t_i\gamma_i$ where $t_i$ is a total number of population, $\gamma_i$ is the rate and $i = 1, 2$. To make inference on the ratio of two Poisson rates ($R' = \gamma_1/\gamma_2$) suppose the null hypothesis of the ratio is unity (i.e. $H_0 : \gamma_1/\gamma_2 = R = 1$) and one side alternative hypothesis is $H_1 : \gamma_1/\gamma_2 = R' > R$ where $R'$ is a pre-specified positive number

which is the ratio of two Poisson rates, [Ng and Tang (2005)]. Then the Wald statistic is used to test the difference between two rates [Gu et al. (2008)].

$$PRT = \frac{X_1 - X_2\rho}{\sqrt{X_1 + X_2\rho^2}}, \quad \rho = \frac{R}{d}, \quad d = \frac{t_2}{t_1}, \tag{3.16}$$

where $PRT$ is the Wald statistic to test two Poisson rates. If $PRT \geq z_{1-\alpha}$, reject null hypothesis and the approximate power can be obtained as [Gu et al. (2008)]:

$$\text{Power} = \Phi\left(\frac{z_{1-\alpha}\sqrt{(\frac{\rho}{c} + \rho^2) * t_2 * \gamma_1 * R'} - (\frac{\rho}{c} - \rho) * t_2 * \gamma_1 * R'}{\sqrt{(\frac{\rho}{c} + \rho^2) * t_2 * \gamma_1 * R'}}\right), \tag{3.17}$$

where $\Phi(.)$ is the cumulative standard normal distribution function, $c = \frac{R}{R'}$ and $z_{1-\alpha}$ is a critical value of the significance level.

To detect the effect size ($R'$), power, sample size and significance level (0.05) are used. The approximate sample size ($ss$) can be obtained as [Gu et al. (2008)]:

$$ss = \frac{((c/\rho) + c^2)(z_{1-\alpha} + z_{1-\beta})^2}{(1-c)^2}, \tag{3.18}$$

where $z_{1-\beta}$ is a critical value of significance power.

### 3.4.2 Power analysis

Published reports often have statements such as there has or has not been a significant change in the HAI rates from one year to the next year [HPS (2008b)]. This research is going to consider the magnitudes of the effect sizes that can be detected in Scotland and in a selection of typical health boards such as Glasgow, Lothian, Grampian, Tayside and Fife. This covers a variety of population sizes.

The difference between the rates of MRSA bacteraemia in two following years 2005 and 2006 and between the rates of MSSA in two following years 2006 and 2007 (these years are chosen to be in the beginning of period of study when the rates were high) in Scotland and in some health boards such as Glasgow, Lothian, Grampian, Tayside and Fife was tested. Statistical power and sample size are then obtained. Applying Equation (3.16) to compare Poisson rates (in 2005 and 2006) in Scotland shows that the ratio between two Poisson rates of MRSA bacteraemia is not significantly greater than one where the p-value =0.264 and the ratio estimate is 1.03 (see the calculation in Appendix A.3.1). Because there is not enough power to detect this difference, the sample size and power statistic analysis have been investigated.

Statistical power of a test is the probability of detecting an effect that actually exists. MRSA bacteraemia data in 2005 and 2006 was used to test the ratio and detect the power. The approximate power of the ratio (effect size) is calculated from Equation (3.17), (see the calculation in Appendix A.3.2).

Table 3.8 shows the power for Scotland and some health boards when the ratio between two MRSA rates is $R' = \frac{\text{average rates in 2005}}{\text{average rates in 2006}} = 1.03$. If sample size (cases of MRSA bacteraemia) decreases, the power of the test will decrease and the test will have insufficient power to detect the observed effect size ($R' = 1.03$). On the other hand, given the rate of MRSA bacteraemia in Scotland in 2005 with a significance level of 5%, the population number in 2006 and an increase of the effect size (ratio change) by 25% (i.e. 1.25), the 99% power can detect 25% change in the rate of MRSA bacteraemia in Scotland. If the effect size in Lothian is 25%, change in the rate of MRSA bacteraemia can be detected with 73% power. 99% power is achieved to detect a 50% change in the rate of MRSA bacteraemia in Lothian, (see Figures 3.20).

Table 3.8: Power calculation for MRSA rates in Scotland and some health boards.

| Health boards | $R' = 1.03$ | $R' = 1.25$ | $R' = 1.5$ | $R' = 2$ | $R' = 2.5$ |
|---|---|---|---|---|---|
| Scotland | 0.1662 | 0.9999 | 1 | 1 | 1 |
| Glasgow | 0.1002 | 0.9354 | 0.9999 | 1 | 1 |
| Lothian | 0.0832 | 0.7329 | 0.9989 | 1 | 1 |
| Grampian | 0.0777 | 0.6182 | 0.9918 | 1 | 1 |
| Tayside | 0.0735 | 0.5172 | 0.9688 | 1 | 1 |
| Fife | 0.0673 | 0.3615 | 0.8499 | 0.9999 | 1 |

$R' = 1.03$: Effect size, i.e. the ratio of the rate of MRSA in Scotland in 2005 and 2006 where $\gamma_{2005} = 18.312$ is the rate of MRSA in Scotland in 2005 and $\gamma'_{2005} = \gamma_{2006} R'$. Also, $R'$ can be 1.25 (25%), 1.5 (50%), 2 (100%) and 2.5 (150%).

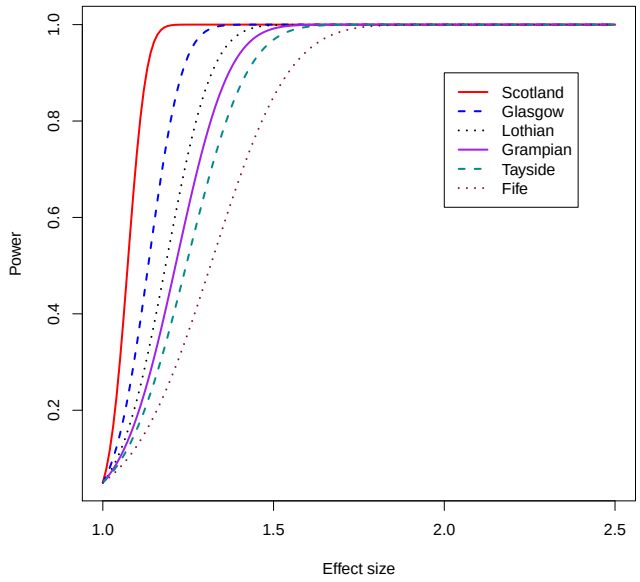



Figure 3.20: Relationship between the effect size and the power using MRSA data.

The Poisson ratio test shows that the ratio between two Poisson MSSA rates is significantly different from 1 (p=0.014) where the estimated ratio is 1.11. Table 3.9 shows that the power for Scotland is 93% when the ratio between two years of MSSA rates is $R' = \frac{\text{average rates in 2007}}{\text{average rates in 2006}} = 1.11$. If the sample size (cases of MSSA bacteraemia) decreases, the power of the test will decrease under the same size effect and the test will have insufficient power to detect the observed effect size ($R' = 1.11$) in Glasgow, Lothian, Grampian, Tayside and Fife. On the other hand, if the effect size is increased by 25%, there will be 98% power to detect

change in the rates of MSSA bacteraemia in Glasgow. When the effect size in Scotland, Glasgow, Lothian, Grampian, Tayside and Fife is 50%, the change in the rate of MSSA bacteraemia can be detected with 95% power or more, (see Table 3.9). There is a positive relationship between effect size, the power and sample size, (see Figures 3.21).

Table 3.9: Power calculation for MSSA rates in Scotland and some health boards.

| Health boards | $R' = 1.11$ | $R' = 1.25$ | $R' = 1.5$ |
|---|---|---|---|
| Scotland | 0.9314 | 1 | 1 |
| Glasgow | 0.5218 | 0.9883 | 1 |
| Lothian | 0.3336 | 0.8756 | 0.9999 |
| Grampian | 0.2714 | 0.7713 | 0.9995 |
| Tayside | 0.2270 | 0.6629 | 0.9959 |
| Fife | 0.1707 | 0.4811 | 0.9530 |

$R' = 1.11$: Effect size, i.e. the ratio of the rate of MSSA in Scotland in 2006 and 2007 where $\gamma_{2006} = 27.8$ is the rate of MSSA in Scotland in 2006 and $\gamma_{2007} = \gamma_{2006}R'$. Also, $R'$ can be 1.25 (25%) and 1.5 (50%).

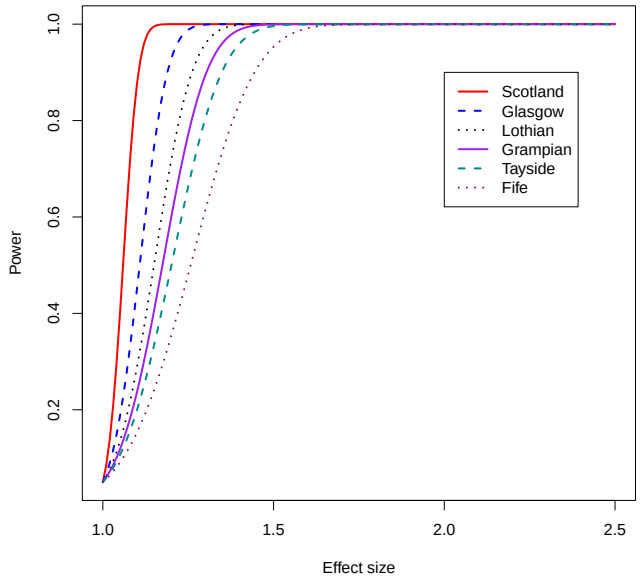



Figure 3.21: Relationship between the effect size and the power using MSSA data.

The Poisson ratio test shows that the ratio between two Poisson rates of CDI in patients over 65 years is not significantly different from 1 (p=0.983), where the estimate ratio is 1.003. The Poisson ratio test of two Poisson rates of CDI in patients aged 15-64 years is significantly different from 1 (p<0.001) where the estimate ratio is 1.081. Figures 3.22 and 3.23 and Tables 3.10 and 3.11 show the power detected at different effect sizes for CDI data. CDI in patients over 65 years detects at least 89% power for just a 20% effect size in small health boards. However, CDI in patients aged 15-64 years detects at least 70% power for a large effect size of 50% in small health boards.

Table 3.10: Power calculation for CDI in patients over 65 years in Scotland and some health boards.

| Health boards | $R' = 1.003$ | $R' = 1.10$ | $R' = 1.20$ |
|---|---|---|---|
| Scotland | 0.0703 | 0.9999 | 1 |
| Glasgow | 0.0600 | 0.9186 | 0.9999 |
| Lothian | 0.0572 | 0.7243 | 0.9982 |
| Grampian | 0.0558 | 0.5694 | 0.9808 |
| Tayside | 0.0554 | 0.5234 | 0.9665 |
| Fife | 0.0544 | 0.4087 | 0.8926 |

$R' = 1.003$: Effect size, i.e. the ratio of the rate of CDI in patients over 65 years in Scotland in 2007 and 2008. Also, $R'$ can be 1.10 (10%) and 1.20 (20%).

Table 3.11: Power calculation for CDI in patients aged 15-64 years in Scotland and some health boards.

| Health boards | $R' = 1.081$ | $R' = 1.25$ | $R' = 1.5$ | $R' = 2$ |
|---|---|---|---|---|
| Scotland | 0.4528 | 0.9993 | 1 | 1 |
| Glasgow | 0.2169 | 0.8648 | 0.9999 | 1 |
| Lothian | 0.1487 | 0.6088 | 0.9906 | 1 |
| Grampian | 0.1243 | 0.4684 | 0.9465 | 0.9999 |
| Tayside | 0.1103 | 0.3798 | 0.8728 | 0.9999 |
| Fife | 0.0943 | 0.2781 | 0.7124 | 0.9987 |

$R' = 1.081$: Effect size, i.e. the ratio of the rate of CDI in patients aged 15-64 years in Scotland in 2010 and 2011. Also, $R'$ can be 1.25(25%), 1.50(50%) and 2(100%).

In conclusion, whenever an effect size increases, the power of the test increases for the same sample size. For a constant effect size, the power will

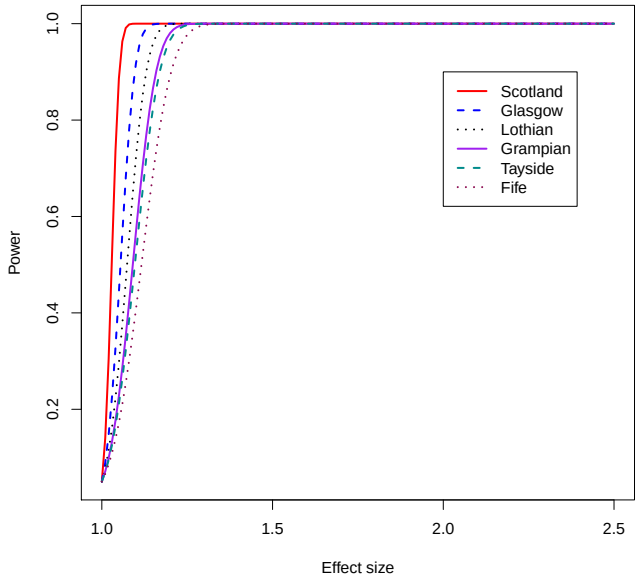



Figure 3.22: Relationship between the effect size and the power using CDI over 65.

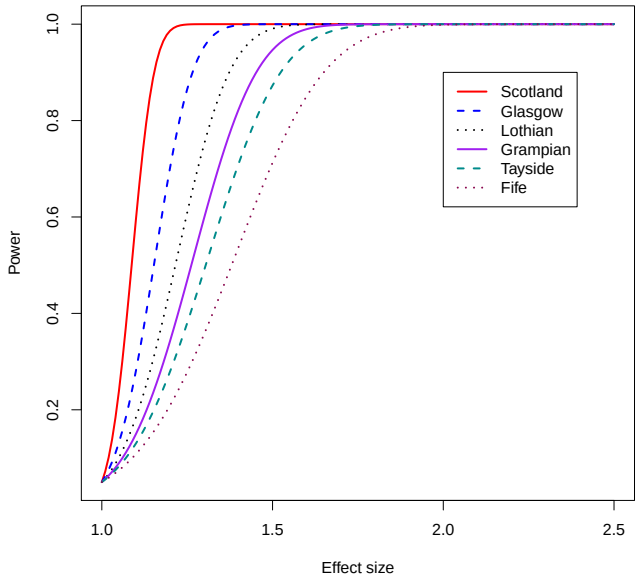



Figure 3.23: Relationship between the effect size and the power using CDI under 65.

increase if the sample size increases. Therefore, changes in MSSA from one year to the next year are easier to be detected because MSSA is more common and has greater number of events. Also, changes in CDI in patients over 65

years needs only a 20% effect size in small health boards and 10% effect size in large health boards to be detected at roughly 80% power. The next section will identify the detectible effect sizes for 80%, 90% and 95% power.

### 3.4.3 Effect size detection

In order to find the detectible effect sizes needed to achieve 80%, 90% and 95% power, Equation (3.18) was used with a significance level 5%. Given a sample size (cases of infections), the effect sizes ($R'$) are detected for 80%, 90% and 95% power. For example, if the sample size is 500, Figure 3.24 shows that the detectible effect size for 95% power is roughly 25% ($R' = 1.25$). The detectible effect size for 90% power is roughly 20% ($R' = 1.2$). This number of cases can be found in about two years of MRSA data in Scotland at the early years (when the data monitoring started) and in about three years at the later years (up to December 2013). However, 500 cases of MSSA in Scotland can be found in one quarter (three months) at early time points and roughly in four months at the later time points (up to December 2013).

The same number of cases (500) for CDI in patients over 65 years occurs roughly in one month at early time points (from October 2006) and in five months at later time points (up to September 2014). Similarly, 500 cases of CDI in patients aged 15-64 years occurs in five months at early time points (from April 2009) but roughly in one year at later time points (up to September 2014).

In conclusion, large sample sizes gives 80%- 95% power with small effect size (ratio of two rates). In contrast, only large effect sizes can be detected with 80%- 95% power in small sample sizes. Therefore, in small sample sizes (small health boards), the effect size cannot be detected between years and the more complicated polynomial model cannot be fitted. Models fitted for Scotland

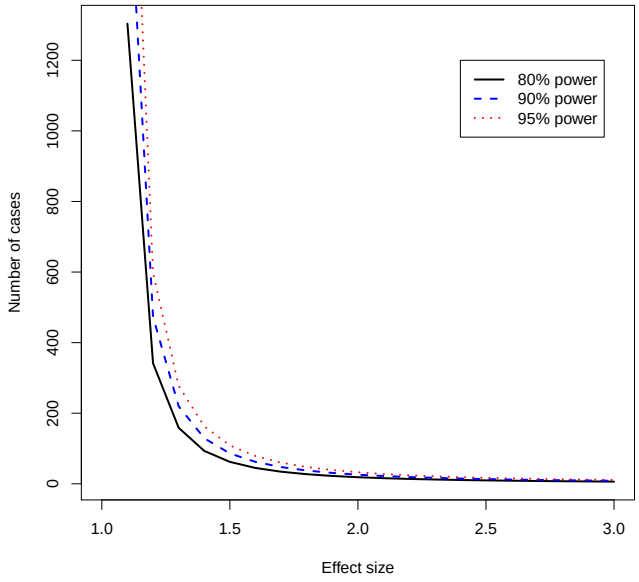



Figure 3.24: Relationship between the effect size and the sample size under 80%, 90% and 95% power.

might be not suitable for some individual health boards. Strong evidence of a change in trend can be found in Scotland but in some individual health boards there is no evidence of decreasing trend due to small numbers of infection cases.

## 3.5 Compare Scotland NHS health boards for HAI rates

### 3.5.1 Funnel Plot

A funnel plot is a graphical aid designed to look at the expected natural variation and is used to detect points which are further away from the average than expected. It is a type of scatter plot where the observed rates are plotted against their population sizes where this is called an un-adjustment funnel plot. The funnel plot has four components (see Figure 3.25); an indicator, a target,

a precision and control limits. The indicator is the observed value plotted on the Y-axis. The target is the overall expectation (mean) for the institutions considered and it appears as a horizontal line. The precision is a parameter to determine the accuracy of measuring the indicator where it can be the size of the population and is plotted on the X-axis. The control limits are overlaid on the scatter plot and represent the expected variation in the indicators (rates). The formula for control limits depends on the distribution of data and the funnel plots. In this research, Byar's method was used (see Section 3.1.5) to calculate control limits. Wide control limits relate to small precision and small variability is expected in large populations [Spiegelhalter (2005) and Dover and Schopflocher (2011)].

#### Unadjusted funnel plot

The observed rates are used to plot unadjusted funnel plot. This entails sorting the data in ascending order according to the population size and the observed rates of the institutions are then plotted against their population sizes. The target is the mean of observed rates and the control limits are calculated using Byar's method with the observed rates (unadjusted) [Morton et al. (2011)].

#### Risk adjustment funnel plot

The observed rates can be adjusted by some risk covariates using a regression model and this process is known as risk adjustment [Woodall (2006)]. Risk adjusted funnel plots are used to improve the analysis by adjusting the rates for known covariates which explain some of the variation among the data points. Having the precision (the size of population in each institution) and the indicator (observed rate in each institution) means that the risk adjusted rates can be obtained where it is calculated as [Dover and Schopflocher (2011)]:

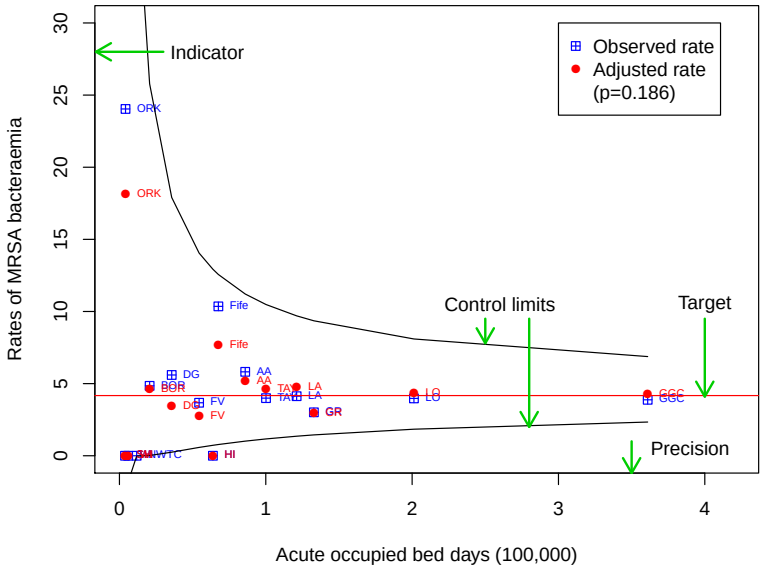



Figure 3.25: Funnel plot of adjusted and unadjusted MRSA bacteraemia rates in the period October-December 2013. The green arrows indicate each funnel plot components.

$$\text{Risk adjusted rate} = \frac{\text{Observed value}}{\text{Risk adjusted value}} \times \text{Average of observed rates.} \quad (3.19)$$

After sorting the data in ascending order according to the population sizes of health boards, the risk adjusted rate is plotted against institution. A smooth funnel shape with control limits depending on the observed rates (unadjusted rates) is plotted [Morton et al. (2011)].

### 3.5.2 Unadjusted funnel plot analysis

Funnel plot analysis was used to examine variations among the NHS boards in Scotland for the MRSA and MSSA bacteraemia rates during October-December 2013 which was the last current period of when this analysis was carried out, (see Appendix A.4.1). The funnel plot was also examined in all previous quarters. Since different health boards have got slightly different characteristics (e.g.

some of them are large and others are small), the aim of the funnel plot analysis is to establish if there are any health boards requiring further investigation to determine a cause of high rates. The rates of MRSA and MSSA bacteraemias for each health board occurring outside the 95% control limits may be associated with more variation than expected.

Figures 3.25 and 3.26 represent that the MRSA and MSSA bacteraemia rates respectively for individual NHS boards in Scotland in Qu4, 2013, where they are within or below the 95% control limits. The overall average of MRSA and MSSA bacteraemia rates for Scotland's NHS boards in the Qu4, 2013 is 4.2 and 29.2 per 100,000 acute occupied bed days, respectively. The rate of MRSA in NHS Highland is below the control limit which indicates that NHS Highland may provide an enhanced healthcare to control infection. Rates of MRSA in NHS Shetland, Western Isles and National Waiting Times Centre are approximately similar and the rates of MSSA in NHS Orkney and Western Isles are approximately the same. Therefore, most of NHS boards in Scotland do not have extremely high or low rates of MRSA and MSSA bacteraemias.

However, the funnel plot does not account for the risk factors or clinical procedures in different NHS boards. The next section will take into account the risk factors; surgical procedures and training courses for the trainer (teaching hospital).

### 3.5.3 Risk adjusted funnel plot analysis

In this section, a risk adjusted funnel plot was used to improve the analysis by adjusting the rates of MRSA and MSSA bacteraemias using surgical procedure data from April 2009. Teaching hospital was identified to adjust the rates of MRSA and MSSA bacteraemias and CDI in Great Glasgow and Clyde, Tayside,

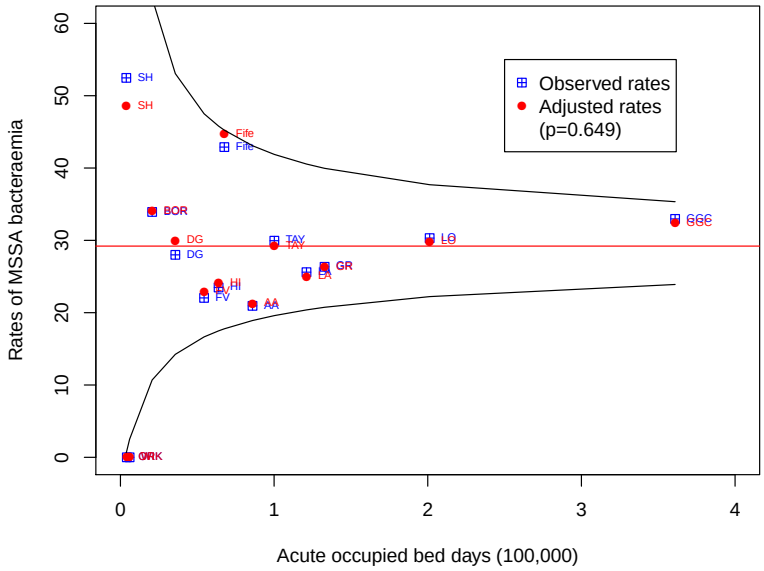



Figure 3.26: Funnel plot of adjusted and unadjusted MSSA bacteraemia rates in the period October-December 2013.

Grampian and Lothian (see Table 3.1). MRSA and MSSA bacteraemias are more common following a surgical procedure compared to only medical (non-surgical) risk factors. The rate of surgery varies over the health boards and this might explain some of the variation in the rates over the HBs. Information Services Division (ISD) has data on the number of patients who had surgery in NHS in Scotland [ISD (2014)]. The percentage of patients undergoing acute surgical procedure (*asp*) (see Table 3.1) was used to adjust the rates of MRSA and MSSA bacteraemias. The data from health boards for the Qu4, 2013 was used to fit a model by adding *asp* as a risk factor. Since data on *asp* is not provided for the National Waiting Times Centre (NWTC) health board, it is excluded from the analysis. Model (no.MRSA $\sim$ offset(log(AOBDs)) + *asp*) was fitted by using Poisson distribution to adjust the rate of MRSA bacteraemia where the *asp* coefficient was not significant in this quarter with p=0.186, see Figure 3.25.

Similarly, model (no.MSSA ~ offset(log(AOBDs)) + *asp*) was fitted using a Poisson distribution to adjust the rate of MSSA bacteraemia in the same quarter (Qu4, 2013) and the *asp* coefficient was also not significant in this quarter (p=0.649) (see Figure 3.26). As a result, surgical procedure does not affect the rates of MRSA and MSSA bacteraemias in the Qu4, 2013.

For more investigation about the impact of surgical procedure on the rate of MRSA and MSSA bacteraemias, different periods of time (different quarters) from April 2009 to December 2013 were considered to fit a model by adding ASP as a risk factor. The risk adjusted rates are obtained from Equation (3.19) where the risk adjusted value is the expected value from the model and the average of observed rates is the rate of Scotland overall. Figures A.1, A.2, A.3, A.4 and A.5 show the observed and adjusted rates of MRSA bacteraemia in different quarters and p-values of *asp* in each quarter. Clearly, in most quarters, the observed and adjusted rates of MRSA bacteraemia are approximately the same. The *asp* coefficient was significant in some quarters (Qu2, 2009, Qu1, 2011, Qu2, 2011 and Qu1, 2012) and the rate of MRSA bacteraemia changed significantly while in other quarters; the *asp* factor did not affect the rate of MRSA bacteraemia. Therefore, even though the observed and adjusted rates for MRSA are significantly different in some quarters, health boards are still less than the upper control limit. Figures A.6, A.7, A.8, A.9 and A.10 show that in all quarters, the observed and adjusted rates of MSSA bacteraemia are approximately the same. As a result, there is a weak evidence that surgical procedure affects the rates of MSSA bacteraemia. However, there are some health boards with rates above the 95% control limit in some quarters which may be a result of other unknown factors. In conclusion, risk adjustment by surgical procedure is not a very strongly associated factor to explain the rates and there could be other factors affecting the rates of MRSA and MSSA bacteraemias in

some health boards which contribute to explaining the variability. For the data of CDI, surgical procedure information is not recorded.

Using teaching hospital to adjust the rates of MRSA and MSSA bacteraemias does not affect the rates of MRSA and MSSA bacteraemias at any period of time. Teaching hospital was also used as a variable to adjust the rate of CDI in Qu3, 2014 and showed that this does not impact the rate.

## 3.6 Conclusion and discussion

Health Protection Scotland reported a general decrease in the rates over time of MRSA, MSSA until December 2013 [HPS (2014)] and of CDI until September 2014 [HPS (2015a)]. Polynomial regression models are fitted to the data for MRSA, MSSA and CDI and a significant pattern of change over time was observed. A quartic regression model describes the change in the rates of MRSA and CDI in patients over 65 years very well. A cubic regression model was fitted to the data for CDI in patients aged 15-64 years. MSSA does not show a significant change in pattern and there is only a general reduction during the period of study. However, polynomial regression models may not satisfy the prediction where the structure imposed cubic or quartic temporal trend is driven by the observed data and may not hold in the future.

The seasonal effect explains some of the change in the trend of some HAI rates. The models of MRSA and CDI in patients aged 15-64 year present strong evidence of seasonal effect. In contrast, the models of MSSA and CDI in patients over 65 years do not illustrate evidence of a seasonal effect. The seasonal effect is an important variable to explain the change in the rates of HAIs and give the best description of the change in the trend [Reil et al. (2012) and Rodriguez-Palacios et al. (2009)]. Additionally, the seasonal effect impacted

the rate when the models were adjusted by health boards which indicates that seasonality impacts individual health boards. Therefore, although seasonal effect does not impact some infections, the final selected models include the seasonal effect for all data because it describes the changes in the infection rates in individual health boards.

Polynomial regression models give the best description of the change in the trend of MRSA and CDI where all infections (apart from MSSA) have high rates initially then decrease, level off and increase again. This implies the occurrence of turning points which identify where the trend changes. These may then equate with interventions which were introduced to impact the rates of infections in Scotland. It is therefore of interest to estimate when changes took place and the identification of turning points and their confidence intervals from polynomial regression models which is investigated in Chapter 4.

Dealing with individual health boards showed that large population size health boards have similar trends to Scotland overall. It is difficult to observe nonlinear effects in small population size health boards because polynomial regression (nonlinear effects) will result in overprediction or underprediction in such health boards. The Scotland model should be adjusted by health board to better explain the trend of infections in each health board. The model adjusted by health board showed that each health board has a different shape of trend, but roughly all health boards have a similar pattern of increasing then decreasing rates of infection.

Differences between population sizes of health boards implies a power issue as some health boards are very small. The power analysis detects the effect size which is needed to fit polynomial models to Scotland overall data in order to

detect change points. However, the power analysis indicated that small health boards cannot detect small changes in rates (small effect size) with 80%-95% power and therefore complicated polynomial models cannot be fitted.

Funnel plots show no major differences between health boards in their overall rates at each period separately, with and without adjustment by surgical procedure or teaching hospitals. Funnel plots illustrate that although health board is a significant factor to explain the rate in individual health boards, the rates of infections stay within the control limits and all health boards indicate similar rates of infections. During the period of study very few quarters showed the impact of surgical procedure on the rate of infections. Therefore, there is no strong evidence of differences among health boards in trends and that implies that the changes occur at similar times (not identical but there are few different times) in each health board.

As a result, the interventions which are associated with changes in the rates are not restricted to some health boards as they are implemented in all health boards. This indicates that national policies may be associated with the changes and leads to investigating the change at the time when these interventions took place. This is discussed in Chapter 5.

In conclusion, the main aim of this chapter was to describe the trend of infections where two of three infections demonstrate non linear trends; MRSA and CDI. These trends have change points and in the following chapters these are going to be investigated. Chapter 4 involves methods identifying turning points within polynomial GLM regression models. Chapter 5 analyses segmented regression and joinpoint methods to detect change points.

# Chapter 4

# Estimation of Turning Points and Construction of their Confidence Intervals

Finding the points in time when the trend of healthcare associated infections (HAIs) changes is of a particular interest. The previous chapter (Chapter 3) determined the best fitting models for HAIs. These models showed that the rate of infection changes at particular time points and turning points were noted in a number of occasions. In this chapter, we shall use these models to estimate the time when the rates change (turning points) and their confidence intervals and try to determine which interventions had an impact on HAIs. In Section 4.1, turning points from polynomial models are estimated and constructing confidence intervals for estimated turning points is considered in Section 4.2 where bootstrap and delta methods are used. In Section 4.3, the methods of estimating turning points and confidence intervals are applied to data of HAIs. The associated interventions with these changes in trend are discussed in Section 4.4. Section 4.5 includes two parts of simulation studies. The first part includes a comparison between bootstrap and delta methods to construct

confidence intervals for estimated turning points from a quadratic model. The behaviour of those confidence intervals when the sample size changes is also explored. The second part uses a cubic model to investigate the performance of the bootstrap method to construct confidence intervals for two estimated turning points.

## 4.1 Estimating turning points from polynomial models

The turning points when the rate changes can be estimated by using polynomial models which is fitted by generalized linear model (GLM), (see Chapter 3). Considering the best fitted model is a cubic polynomial GLM model as:

$$\log(\text{rate}) = f(t) + Qu + \varepsilon. \tag{4.1}$$

Where $f(t)$ is the cubic polynomial function which is used to estimate the turning points; $f(t) = at^3 + bt^2 + ct + d$. Seasonal effect ($Qu$) is used to adjust the polynomial GLM model and gives the best fit for the data and $\varepsilon$ is the error of the model. Then, we use $f(t)$ to estimate the turning points as follow:

1. Find the first derivative of $f(t)$ in time $t$ where $f'(t) = \frac{df(t)}{dt} = 3at^2 + 2bt + c$.

2. Calculate roots of the $f'(t)$ (i.e. $3at^2 + 2bt + c = 0$) which gives two turning points.

$$t_{1,2} = \frac{-2b \pm \sqrt{(2b)^2 - 4(3a)c}}{2(3a)},$$

   $t_1$ and $t_2$ have a nonlinear function in the parameters of $f(t)$. They are the values of estimated turning points of trend (when the trend changes from a decrease to an increase or from an increase to a decrease). They may or

may not occur within the range of the data because they are calculated from predicted values.

3. Sort the roots into ascending order to get the estimated turning point in temporal order (the first location of estimated turning point is $t_{(1)}$ then the second one is $t_{(2)}$).

4. Use the values of the estimated turning points ($t_1$ and $t_2$) to calculate the value of $f(t)$ at each estimated turning point then use the values of $f(t)$ to calculate the rates when the trend changes as: $\exp(f(t)) \times 100,000$.

5. To find the inflection point, calculate the root of the second derivative of $f(t)$ as $f''(t) = 6at + 2b = 0$ and this implies one inflection point ($infp$). To calculate the rate on the inflection point use the expression $\exp(f(infp)) \times 100,000$.

6. Use $f''(t)$ in order to define whether the estimated turning point has maximum or minimum rate. If $f''(t) < 0$, the change occurs at the maximum rate and the rate starts to decrease. If $f''(t) > 0$, the change occurs at the minimum rate and the rate starts to increase.

## 4.2 Estimating confidence intervals for estimated turning points

Since the turning points are estimated from the fitted model, confidence intervals can give more information about where the change occurs and interventions which impact the rates can be determined. Bootstrap and delta methods are used to construct confidence intervals for the estimated turning points. Since the quadratic model implies one turning point which is the ratio of two random variables (two parameters from quadratic function), the delta method is used which can find the mean and variance of the ratio and use them to

calculate a confidence interval for the estimated turning point. However, for more than two random variables, the delta method is mathematically difficult (see Section 4.3.3) to calculate the variance of the estimated turning points and therefore the bootstrap method is used. The other reason of using the bootstrap method is because the exact distribution of the estimated turning point is unknown since the turning point is a nonlinear function in the parameters of the polynomial model.

### 4.2.1 Bootstrapping

A goal of statistical inference is to determine the value of a population parameter. In practice, if this is expensive or even impossible to measure directly then sampling is used to estimate the population parameter. Bootstrapping is an approach used in statistical inference to estimate population parameters based on random sampling with replacement from the original data [Efron (1979)]. Bootstrapping can be used to estimate standard errors, confidence intervals for parameters, perform hypothesis tests and improve prediction accuracy. It is often used as an alternative method to inference based on parametric approaches when those assumptions are dubious, or when parametric inference is impossible or requires very complicated formulas for the calculation of standard errors.

There are three types of Bootstrap which are parametric, non-parametric and semi-parametric. Parametric bootstrapping generates samples from known distributions using the estimated parameters, while non-parametric bootstrapping estimates unknown distributions from the empirical distribution obtained from the observed data. Semi-parametric bootstrapping is based on re-sampling the residuals of a regression model [Carpenter and Bithell (2000)].

The basic idea of bootstrapping is to take the original observed data (size $n$) from the population and re-sample the original data B times, treating each of these as a new sample from the population. The required statistic is then estimated from each sample. For example, assume the mean of body mass index (BMI) of people worldwide is of interest. The approach would be to take a sample of size $n$ from the population and record the individual BMIs where one estimate of the mean can be obtained in each single sample. In order to get a robust result about the population mean BMI, an estimate of the variability of the mean is required. The bootstrap sample of size $n$ is taken from the original data by sampling with replacement. Assuming $n$ is sufficiently large, the bootstrap sample has zero probability to be identical to the original sample. This process is repeated a large number of times (B times) and the mean for each of these bootstrap samples is computed (each of these is called a bootstrap estimate). A histogram of bootstrap means provides an estimate of the shape of the distribution of the mean [Davison (1997)].

The general bootstrap algorithm is as follows:

1. Get a sample of size $n$.
2. Re-sample with replacement size $n$ from original sample.
3. Calculate parameter of interest and save it.
4. Repeat steps 2 and 3, B times to get bootstrap estimates.
5. Use the bootstrap values in step 4 for statistical inference such as estimating confidence intervals for the mean.

### 4.2.2 Delta method

The delta method is a technique used to calculate confidence intervals for functions of maximum likelihood estimators where these functions have approximately normal distributions. Based on a truncated Taylor series, the delta method deals with a complicated and nonlinear function of one or more random variables to obtain the estimator of the variance of the nonlinear function. Using a first order Taylor expansion around the mean value of the variables and creating a linear approximation of that function, the variance of the simpler linear function is then estimated [Abramowitz and Stegun (1964), Casella and Berger (2002), Xu and Long (2005) and Hole (2007)].

If $X$ and $Y$ are random variables, $E(X) = \mu_X \neq 0, \quad E(Y) = \mu_Y \neq 0$ are their means respectively. If a function of a ratio of two random variables is assumed as $g(X, Y) = \frac{X}{Y}$, the approximation of the expected value $E\left(\frac{X}{Y}\right)$ and variance $Var\left(\frac{X}{Y}\right)$ of the function $g(X, Y)$ are given by [Casella and Berger (2002)]:

$$E\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}, \tag{4.2}$$

$$Var\left(\frac{X}{Y}\right) \approx \left(\frac{\mu_X}{\mu_Y}\right)^2 \times \left[\left(\frac{Var(X)}{\mu_X}\right)^2 + \left(\frac{Var(Y)}{\mu_Y}\right)^2 - 2\frac{Cov(X, Y)}{\mu_X \mu_Y}\right]. \tag{4.3}$$

Since the estimated turning point ($\hat{t} = \frac{-b}{2a}$) from a quadratic model is the ratio of two parameters ($a$ is a coefficient of quadratic term and $b$ is a coefficient of linear term) of the quadratic model, the approximate mean and variance of $\hat{t}$ are calculated and therefore the confidence interval of the $\hat{t}$ is constructed using the delta method where it is obtained as:

$$E(\hat{t}) \pm z_{1-(\alpha/2)} \sqrt{Var(\hat{t})}.$$

## 4.3 Estimation of turning points and confidence intervals for HAIs data

This section includes the estimating of turning points using HAIs data. Using HAIs data up to September 2014, turning points for the best fitting models of MRSA bacteraemia (quartic model) and CDI in patients over 65 years (quartic model) were estimated, and confidence intervals for these turning points were constructed using the bootstrap method. Since the MSSA bacteraemia model has a linear trend (see Chapter 3), there are no turning points. The estimated turning points for the cubic model of CDI in patients aged 15-64 years cannot be found because solving its quadratic function gives two complex numbers.

In addition, in order to compare delta and bootstrap methods for constructing a confidence interval of one turning point, a turning point from quadratic model of MRSA bacteraemia and its confidence interval are estimated. A quadratic model is used because the delta method deals easily with two random variables while it is computationally complicated when dealing with more than two random variables.

### 4.3.1 Estimated turning points in MRSA bacteraemia and CDI models

#### MRSA bacteraemia quartic model

The best fitting model to describe MRSA bacteraemia rates was a quartic model (3.7). Based on the method of estimating turning points (see Section 4.1), this model includes the quartic polynomial function $f(t)$ and is adjusted by seasonal effect ($Qu$) where the first derivative of quartic function of time is the cubic function. Solving the cubic function gives three estimated turning points. The

maximum predicted rate was 19.61 per 100,000 AOBDs when the rate starts to decrease at time point $\hat{t}$ = 2005.65 (August 2005). The minimum predicted rates were 17.13 per 100,000 AOBDs when the rates start to increase at time point $\hat{t}$ = 2002.63 (August 2002) and 2.83 per 100,000 AOBDs at time point $\hat{t}$ = 2013.80 (end of October 2013). The points of inflection (points of curve at which a change in the sign of curvature occurs) are at 2004 (January 2004) when the rate was 18.26 per 100,000 AOBDs and at 2010.75 (October 2010) when the rate was 6.40 per 100,000 AOBDs, see Table 4.1 and Figure 4.1.

Figure 4.1 did not show the first estimated minimum turning point (at 2002.6) as it occurs outside the range of data. Although a quartic model is the best fit to the data there are only two turning points within the range of the data and the confidence intervals will be constructed in the next section for these two points only. Figure 4.1 shows that there are some fitted rates greater than the rate at turning point 2005.65 (i.e. $\hat{y}(\hat{t}+1) > \hat{y}(\hat{t})$ or $\hat{y}(\hat{t}-1) > \hat{y}(\hat{t})$). This is due to the effect of seasonality (i.e. the fitted line adjusted by the quarterly effect) where such rise in the rate is associated with the seasonal effect as $\hat{t}$ is estimated from the polynomial function $f(t)$ only.

Table 4.1: Estimated turning points and their confidence intervals.

| Infection | Estimated turning points | Bootstrap confidence interval |
|---|---|---|
| MRSA | Min @ 2002.63 (August 2002) | (April 2000- October 2003) |
| bacteraemia | Max @ 2005.65 (August 2005) | (April 2005- December 2005) |
| (Quartic model) | Min @ 2013.80 (October 2013) | (June 2013- April 2014) |
| | inf1 @ 2004 (January 2004) | (June 2002- September 2004) |
| | inf2 @ 2010.75 (October 2010) | (June 2010- December 2010) |
| CDI in | Max @ 2007.64 (August 2007) | (July 2007- September 2007) |
| patients | Min @ 2012.53 (July 2012) | (February 2012- January 2013) |
| over 65 years | Max @ 2013.80 (October 2013) | (December 2012- January 2015) |
| (Quartic model) | inf1 @ 2009.44 (June 2009) | (April 2009- August 2009) |
| | inf2 @ 2013.20 (March 2013) | (October 2012- January 2014) |

**Min:** Minimum rate, **Max:** Maximum rate, **inf:** Inflection points.

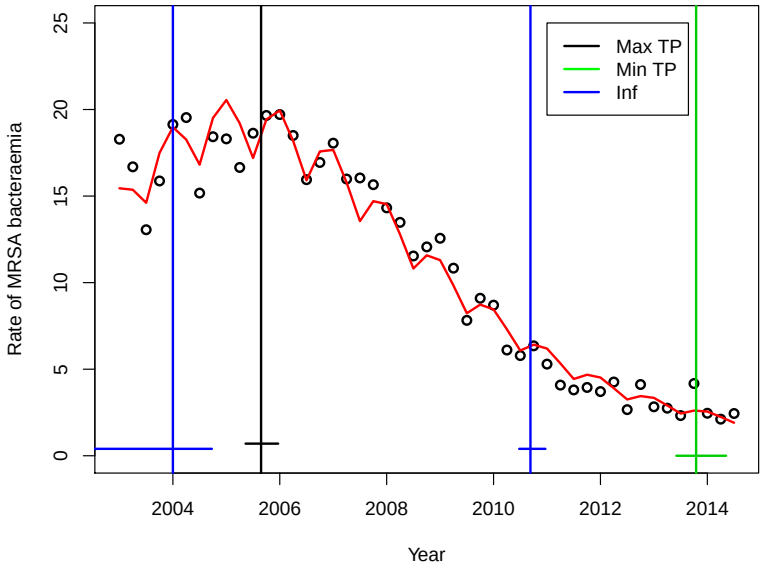



Figure 4.1: Estimated turning points and inflection points on MRSA fitted quartic model (3.7) (vertical line) with 95% confidence intervals (horizontal line). Black lines are estimated turning points when the rate start to decrease, green lines are estimated turning points when the rate start to increase and blue lines are inflection points. The red line is the predicted rates and the black circles are the observed rates.

### CDI in patients over 65 years quartic model

The quartic model (3.9) of CDI in patients over 65 years has three estimated turning points associated with a change in trend. There are two estimated turning points for CDI with maximum rates of 170.75 per 100,000 AOBDs at time point $\hat{t}$ = 2007.64 (i.e. the end of August 2007) and 36.81 per 100,000 AOBDs at time point $\hat{t}$ = 2013.80 (during October 2013). The minimum rate was 35.38 per 100,000 AOBDs at time $\hat{t}$ = 2012.53 (July 2012). The inflection points are at 2009.44 (June 2009) when the rate was 88.81 per 100,000 AOBDs and at 2013.20 (March 2013) when the rate was 36.11 per 100,000 AOBDs, see Table 4.1 and Figure 4.2.

Figure 4.2 shows that the local maximum rate at 2013.80 comes at a dip in the

observed rates which is a local dip associated with Qu4 and the calculation of the turning points is based upon the predicted trend adjusting for the quarterly pattern. Similarly, for the minimum rate at 2012.53, there is a local rise in the observed rates because it is associated with Qu3 while the calculation of the turning points comes from the predicted trend adjusting for the seasonal effect.

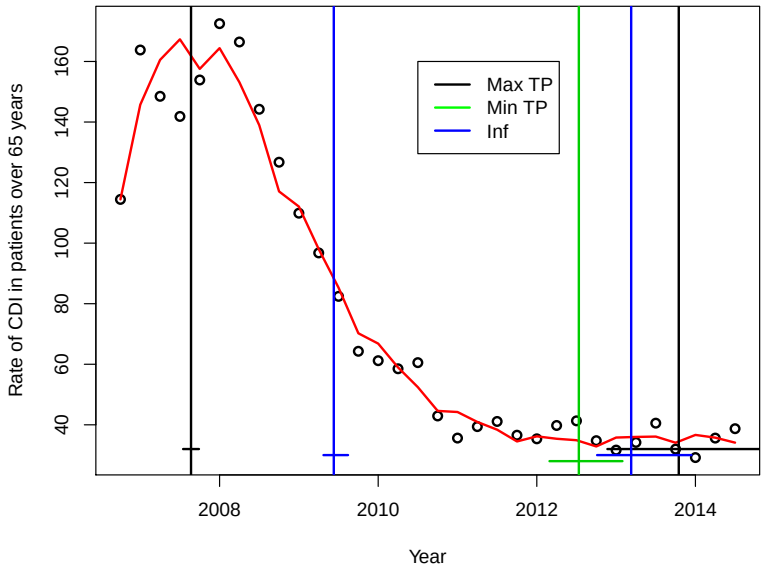



Figure 4.2: Estimated turning points and inflection points on CDI in patients over 65 years fitted model (3.9) (vertical line) with 95% confidence intervals (horizontal line). Black lines are estimated turning points when the rate start to decrease, green lines are estimated turning points when the rate start to increase and blue lines are inflection points. The red line is the predicted rates and the black circles are the observed rates.

### 4.3.2 Bootstrap confidence intervals

Since the appropriate probability model for the estimated turning points is unknown, the bootstrap method was used to generate bootstrap samples of estimated turning points by sampling directly from the data with replacement. Semi-parametric bootstrap method deals with regression models when the

residuals are re-sampled. The algorithm for constructing confidence intervals by bootstrapping semi-parametric re-sampling is as follows:

1. Use the original data sample to find the best fitted model of the rate of healthcare associated infections using **glm**.
2. Obtain the fitted values $\hat{y}_i$ and Pearson residuals $\hat{\varepsilon}_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$.
3. Re-sample the residuals and save a new response variable $y_i^*$ by adding re-sampled residuals to the fitted values as $y_i^* = \hat{y}_i + \hat{\varepsilon}_i \times \sqrt{\hat{y}_i}$.
4. Round $y_i^*$ to be integer values.
5. Use the new responses (bootstrap data) to fit the best fitted model by using **glm** function in **R**.
6. Use the coefficients from the bootstrap fitted model to calculate the estimated turning points of the bootstrap fitted model and save them, (see Section 4.1).
7. Repeat steps 3 to 6, 500 times to obtain bootstrap estimates of turning points.
8. Calculate 95% confidence intervals for the estimated turning point by using the bootstrap estimates of turning points conducted by the bootstrap percentile method using (**quantile**) at 0.025 and 0.975 in **R**.

This method was applied to construct confidence intervals for estimated turning points from the MRSA bacteraemia quartic model and CDI in patients over 65 years quartic model.

#### MRSA bacteraemia quartic model

The estimated turning point in the quartic model (3.7) has 95% confidence interval when the rate starts to decrease as (2005.36, 2005.97) (between April and

December 2005) with 95% confidence interval for the estimated rates (18.73, 20.88) per 100,000 AOBDs. The related interventions with this reduction are discussed in Section 4.4. The 95% confidence interval for the estimated turning points when the rate starts to increase is (2013.42, 2014.35) (between June 2013 and April 2014) with 95% confidence interval for the estimated rates (2.46, 3.17) per 100,000 AOBDs. The 95% confidence interval for the points of inflection are (2002.43, 2004.73) (between June 2002 and September 2004) with 95% confidence interval for the estimated rates (14.25, 19.32) per 100,000 AOBDs and (2010.48, 2010.97) (between June and December 2010) with 95% confidence interval for the estimated rates (5.86, 6.95) per 100,000 AOBDs. These are shown in Table 4.1 and Figure 4.1. Figure 4.1 shows that the width of the confidence intervals of the turning points are different due to the magnitude of change in the trend of MRSA bacteraemia.

#### CDI in patients over 65 years quartic model

The 95% confidence intervals for estimated locations of two estimated turning points where the rates of CDI in patients over 65 years start to decrease are (2007.54, 2007.73) (i.e. during Qu3, 2007) with 95% confidence interval (159.70, 182.59) per 100,000 AOBDs for the rates and (2012.90, 2015.10) (i.e. between of the end of 2012 and beginning of 2015) with 95% confidence interval (33.54, 42.85) per 100,000 AOBDs for the rates. The 95% confidence interval for the minimum rate was (31.57, 38.77) per 100,000 AOBDs at time (2012.16, 2013.10) (i.e between the beginning of 2012 and 2013). The 95% confidence intervals for the points of inflection are (2009.31, 2009.62) (between April and August 2009) with 95% confidence interval for the estimated CDI rates (81.94, 94.90) per 100,000 AOBDs and (2012.76, 2014) (between October 2012 and January 2014) with 95% confidence interval for the estimated rates (33.00, 40.17) per 100,000 AOBDs, see Table 4.1 and Figure 4.2. Figure 4.2 shows that the width of the

confidence intervals of the turning points are different and two confidence intervals are overlapping due to their relevant turning points occur close to each other at the end of the dataset.

### 4.3.3 Delta method of confidence intervals

The delta method is another approach to construct confidence intervals for the estimated turning points by finding means and variances of the estimated turning points. The best model for MRSA bacteraemia is the quartic model where the turning points are the roots of cubic polynomial function. The roots of cubic equation $at^3 + bt^2 + ct + d$ [Nickalls (1993), Cardano et al. (2007) and Neumark (2014)] are given as:

$$t_j = \frac{-1}{3a}(b + \eta^{j-1}\omega + \frac{\delta_0}{\eta^{j-1}\omega}), \quad j = 1, 2, 3,$$

where,

$$\omega = \sqrt[3]{\frac{\delta_1 \pm \sqrt{\delta_1^2 - 4\delta_0^3}}{2}}, \quad \omega \neq 0, \quad \eta = -\frac{1}{2} + \frac{1}{2}\sqrt{3}\,i,$$
$$\delta_0 = b^2 + 3ac, \quad \delta_1 = 2b^3 - 9abc + 27a^2d.$$

The estimated turning points $t_j$ are nonlinear functions of the parameters of the quartic model. Using the delta method for calculation of the means and variances of these complicated functions (including ratio, multiplication, square root and cube root) is mathematically difficult since $t_j$ has four random variables. Therefore, to investigate the performance of delta method, the quadratic model for MRSA bacteraemia is simply used.

Although the quartic model is the best fit to the data of MRSA bacteraemia, the quadratic model is the simplest model with a one turning point and the algebra of the delta method is relatively easy to calculate the mean and variance

of the estimated turning point ($\hat{t}$). $\hat{t}$ is a ratio of two random variables $a$ and $b$ (see the algorithm below) and this is the reason that the mean and variance of a ratio are calculated using the delta method which was presented in Section 4.2.2. Following this, the delta and bootstrap methods can be easily compared.

The algorithm is as follows:

1. Fit quadratic model with seasonality to the data of MRSA bacteraemia.

$$\log(\text{no.MRSA}) \sim \text{offset}(\log(\text{AOBDs})) + bt + at^2 + Qu$$

2. Save the coefficients of the quadratic and linear terms, where $a$ is a coefficient of quadratic term and $b$ is a coefficient of linear term.
3. Find the first derivative of quadratic equation which is a linear equation.
4. Calculate the estimated turning point from the linear equation as $\hat{t} = \frac{-b}{2a}$.
5. Calculate the mean and variance of $a$ and $b$ and the covariance of $a$ and $b$.
6. Use the delta method to calculate the expected value $E(\hat{t})$ and variance $Var(\hat{t})$ by using Equations (4.2) and (4.3).
7. Calculate 95% confidence interval for the estimated turning point using

$$E(\hat{t}) \pm 1.96 \times \sqrt{Var(\hat{t})}.$$

The estimated turning point for the quadratic model when the rates of MRSA bacteraemia start to decrease was at time 2004.79 (end of October 2004) and the 95% delta confidence interval is (2004.43, 2005.13) between the end of Qu2, 2004 and the beginning of 2005. The bootstrap method (1000 times) gave a 95% confidence interval of (2004.15, 2005.25) based on a quadratic model which is

a little bit wider than the 95% confidence interval where this was estimated by the delta method. Comparing delta and bootstrap methods of constructing confidence intervals is discussed in Section 4.5.

## 4.4 Associated interventions

The location of maximum estimated change points in the polynomial GLM models gives an idea of when rates start to decrease and may help identify clinical interventions which have impacted these rates. Some of the healthcare interventions listed in Table 1.2 are associated with a reduction in infection rates. Around the time of the drop (2005) in MRSA bacteraemia rates, the use of alcohol hand gel was advised by the chief nursing officer (CNO) [Martin (2015)]. In addition, a cleanliness course was given to all nurses regarding the general environment of patients such as beds, pillow, towels and chairs [RCN (2005)]. Intensive care medicine (ICM) funding in all boards in Scotland was also introduced as a new infection control initiative [ISD (2005)] and improvement of antimicrobial prescribing policy and practice in Scotland was recommended [SE (2005)].

The trend of CDI in patients over 65 years changed during 2007. In March 2007 the Scottish patient safety programme (SPSP) was introduced and aimed to improve safety and reliability in all health care settings [SPSP (2007)].

In conclusion, Table 4.2 summarizes the interventions which may have impacted the rates of HAIs in Scotland.

Table 4.2: Summary of the interventions detected by polynomial GLM which may impacted the rate of HAIs in Scotland.

| Point of change (95% CI) | Interventions | MRSA | CDI |
|---|---|---|---|
| 2005.65 ≈ August 2005 (April 2005- December 2005) | CNO letter on alcohol based hand rubs and infection control. | Yes | |
| | CNO requested that all G Grade Sisters/ Charge Nurses (Senior Charge Nurses) undertake the Cleanliness Champions Course commenced. | Yes | |
| | New IC structure in Boards, including ICM funding. | Yes | |
| | Antimicrobial Prescribing Policy and Practice in Scotland- Recommendations for good antimicrobial practice in acute hospitals. | Yes | |
| 2007.64 ≈ August 2007 (July 2007- September 2007) | Scottish Patient Safety Programme (SPSP) announced. | | Yes |

**CI:** Confidence interval, **MRSA:** Methicillin-resistant staphylococcus aureus, **CDI:** Clostridium difficile infection, **CNO:** Chief nursing officer, **IC:** Infection control, **ICM:** Intensive care medicine.

## 4.5 Simulation study

Since bootstrap and delta methods are two techniques used to construct confidence intervals for estimated turning points from a quadratic model, the comparison between them will be carried out (see Section 4.5.1). The performance of the bootstrap method to construct confidence intervals for two estimated turning points from a cubic model is also investigated, (see Section 4.5.2).

### 4.5.1 Comparing delta and bootstrap methods

A simulation study is carried out to investigate and compare the behavior and performance of bootstrap and delta methods to construct confidence intervals for the estimated turning point from the quadratic model (4.4) where the estimated turning point is a ratio of two parameters and the mean and variance

are easy to calculate using the delta method.

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2, \tag{4.4}$$

where $x_0$ is the true value of the turning point and $\lambda_n$ is the observed data.

The comparison of bootstrap and delta methods will involve the examination of the mean width of confidence intervals (WD) estimated by both methods and the percentage of confidence interval containing the true value of the turning point (CI.TP%).

Four considerations were assumed when carrying out the simulation study:

1. Different sample sizes (i.e. numbers of time points: large 50, moderate 35 and small 20) to fit the quadratic model according to specific values $\beta_0, \beta_1$ and $\beta_2$.

2. Since **glm** model with Poisson distribution is used, $exp(\beta_0)$ is the number of cases per time period which could be rare ($\beta_0 = 1.5$), small ($\beta_0 = 3$) or large ($\beta_0 = 5$).

3. The true turning point $x_0$ can occur in the beginning of the data, in the middle or at the end. For example, when the sample size is 20, the true turning point occurs at 5 (beginning), 10 (middle) or 15 (end).

4. Under different sample sizes, $\beta_1 = 0.001$ but $\beta_2$ often takes the value of -0.003 while when the sample size or the location of true turning point change, $\beta_2$ changes with very small differences (<0.05) (i.e. $\beta_2$s are similar but not identical).

By using **R** programming language (see Appendix B.1), the algorithm for simulation study is written as follows:

1. Generate data according to Poisson distribution,

$$Y_n \sim \text{Poisson}(\lambda_n),$$

where $Y_n$ is a simulated count data, $\lambda_n$ is the observed data of quadratic model (4.4) and $n$ is the sample size.

2. Use simulated data to fit **glm** quadratic model with Poisson distribution,

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{4.5}$$

3. Calculate simulated turning point ($\hat{t} = \frac{-\beta_1}{2\beta_2}$) from new fitted model (4.5).

4. Use the delta method (see Section 4.3.3) to construct a confidence interval for the simulated turning point ($\hat{t}$) and save the results.

5. Use the bootstrap method (500 times) to construct a confidence interval for the simulated turning point ($\hat{t}$) (see Section 4.3.2) and save the results.

6. Repeat the simulation (from step 1 to 5), 500 times to have sets of confidence intervals from both bootstrap and delta methods.

7. Use the mean width of the simulated confidence intervals (WD) and the percentage of simulated confidence intervals including the true turning point $x_0$ (CI.TP%) to compare methods.

## Results and conclusion

Tables 4.3, 4.4 and 4.5 summarise the results of the bootstrap and delta methods of constructing confidence intervals when the true turning point occurs in the middle of the data, beginning and at the end, respectively. $\beta_1 = 0.001$ in all sample sizes but $\beta_2$ (the parameter of curvature) are different when the sample size changes. When $n = 50$, $\beta_2 = -0.003$, when $n = 35$, $\beta_2 = -0.0035$

and when $n = 20$, $\beta_2 = -0.008$. The initial values of $\beta_1$ and $\beta_2$ were chosen to make $\log(\lambda_n)$ in Equation (4.4) positive as much as possible. The mean width of the confidence interval and the coverage of the true turning point are used to compare bootstrap and delta methods.

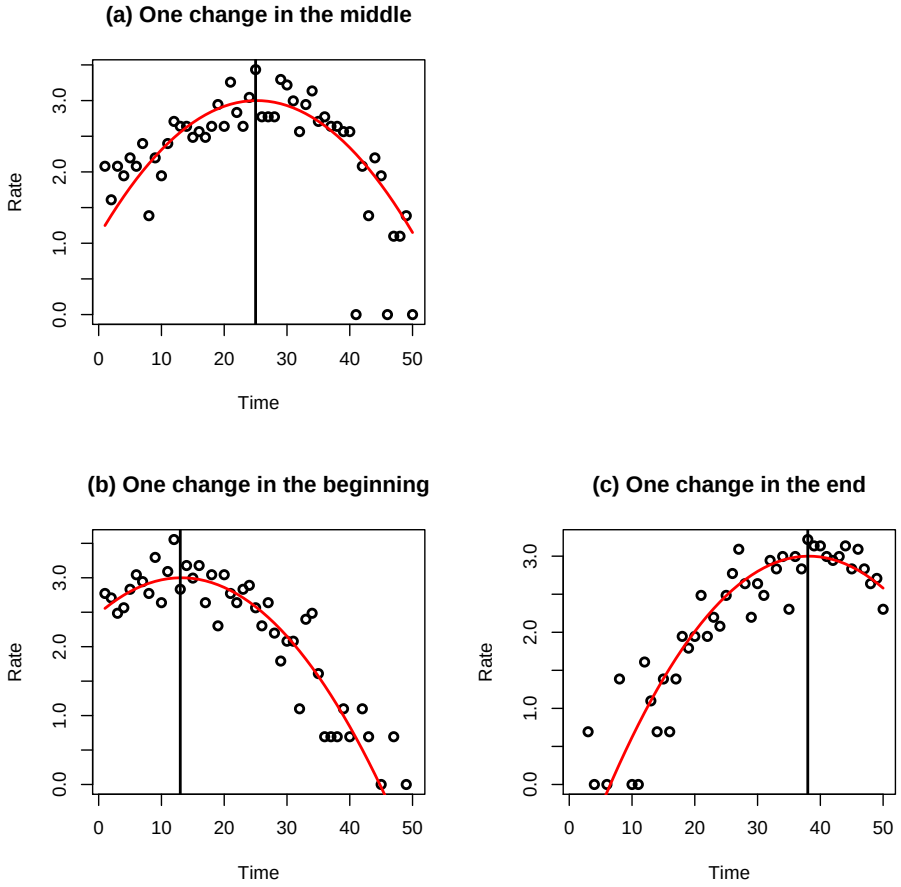



Figure 4.3: The simulated data (black circles) from the quadratic model (red line) when one turning point (black line) occurs in (a) the middle, (b) the beginning and (c) the end of data.

When the true turning point occurs in the middle of the data (i.e. at 25 when $n = 50$, at 18 when $n = 35$ and at 10 when $n = 20$) (see Figure 4.3(a)), the mean width of confidence interval (WD) gets wider when $\beta_0$ gets smaller in both bootstrap and delta methods, (see Table 4.3). The mean width of the confidence intervals by both methods are usually similar when $\beta_0 \geq 3$. In addition, the delta and bootstrap confidence intervals are roughly symmetric

around the true turning point. Based on 500 simulations and significance level of 5%, the range (93%- 97%) is consistent with 95% coverage (i.e. $(0.95 \pm z_{1-(\alpha/2)}\sqrt{0.95 \times 0.05/500}) \times 100 \approx (93\% - 97\%), \quad z_{1-(\alpha/2)} = 1.96$). Therefore, the confidence intervals constructed by the delta method often cover more true turning points (CI.TP%) within 95% coverage than the confidence intervals constructed by the bootstrap method. The bootstrap method covers at most 5% less than lower 95% coverage when $\beta_0 \geq 3$ because the mean width of confidence interval is roughly small but when $\beta_0 \leq 1.5$, it covers 95% because the mean width of confidence interval is large, (see Table 4.3). This low coverage may occur because the method of calculating the bootstrap confidence interval is percentiles. Different methods may give better coverage such as the bias corrected method.

Table 4.3: Comparing the delta and bootstrap methods of constructing confidence intervals for the estimated turning point from quadratic model with change in the middle.

| | | | Delta method | | | | Bootstrap method | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ (TTP) | $\beta_0$ | E(TP) | LCI | UCI | WD | CI.TP% | LCI | UCI | WD | CI.TP% |
| 50 (25) | 5.0 | 25.1 | 24.7 | 25.7 | 1.0 | 92.5 | 24.7 | 25.6 | 0.9 | 89.5 |
| | 3.0 | 26.4 | 23.9 | 26.5 | 2.6 | 93.5 | 23.9 | 26.5 | 2.6 | 90.5 |
| | 1.5 | 24.4 | 22.4 | 28.2 | 5.8 | 95.0 | 22.3 | 28.2 | 5.9 | 93.5 |
| 35 (18) | 5.0 | 17.6 | 17.6 | 18.7 | 1.1 | 92.5 | 17.6 | 18.7 | 1.1 | 87.5 |
| | 3.0 | 15.9 | 16.6 | 19.8 | 3.2 | 96.0 | 16.7 | 19.8 | 3.1 | 92.0 |
| | 1.5 | 19.3 | 13.9 | 22.6 | 8.7 | 96.0 | 10.8 | 26.2 | 15.4 | 94.5 |
| 20 (10) | 5.0 | 10.1 | 9.5 | 10.5 | 1.0 | 95.2 | 9.6 | 10.5 | 0.9 | 88.0 |
| | 3.0 | 9.1 | 8.6 | 11.6 | 3.0 | 96.5 | 8.2 | 11.7 | 3.5 | 90.5 |
| | 1.5 | 9.5 | -10.1 | 30.2 | 40.3 | 98.0 | -1.0 | 20.7 | 21.7 | 95.0 |

$n$**:** The sample size, **TTP:** The true turning point, $\beta_0$**:** The number of cases, **E(TP):** The mean of estimated turning points from the simulations, **LCI:** The mean of lower confidence levels, **UCI:** The mean of upper confidence levels, **WD:** The mean width of confidence intervals, **CI.TP%:** The percentage of confidence intervals that contains the true turning point.

When the true turning point occurs at the beginning of the data (i.e. at 13 when $n = 50$, at 9 when $n = 35$ and at 5 when $n = 20$) (see Figure 4.3(b)) and when $\beta_0$ gets smaller, the mean width of the confidence intervals gets larger for both methods. However, for $n \leq 35$ and $\beta_0 \leq 1.5$, the mean width of confidence intervals (WD) becomes greater than the range of data in both methods. The

percentage of confidence intervals containing the true turning point (CI.TP%) is within 95% coverage when $\beta_0 \leq 3$ using the bootstrap method while when $\beta_0 \geq 5$, the CI.TP% is at most 5% less than lower limit. Similarly, CI.TP% in delta method is sometimes within 95% coverage when $\beta_0 \leq 3$. However, when $\beta_0 \geq 5$, the CI.TP% is at most 3% less than lower limit. The mean width of confidence intervals in the bootstrap and the delta methods are approximately the same except when $\beta_0 \leq 1.5$ or when the mean width of confidence intervals is greater than the sample size, (see Table 4.4). The confidence intervals in both methods are asymmetric and are roughly close to the upper limit of the confidence interval.

Table 4.4: Comparing the delta and bootstrap methods of constructing confidence intervals for the estimated turning point from the quadratic model with change in the beginning.

| | | | Delta method | | | | Bootstrap method | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ (TTP) | $\beta_0$ | E(TP) | LCI | UCI | WD | CI.TP% | LCI | UCI | WD | CI.TP% |
| 50 (13) | 5.0 | 12.8 | 12.4 | 14.1 | 1.7 | 90.0 | 12.4 | 14.0 | 1.6 | 91.0 |
| | 3.0 | 11.8 | 10.7 | 15.5 | 4.8 | 92.5 | 10.3 | 15.1 | 4.8 | 94.0 |
| | 1.5 | 15.6 | 6.6 | 18.3 | 11.7 | 94.0 | -0.5 | 17.5 | 18.0 | 94.0 |
| 35 (9) | 5.0 | 9.0 | 7.9 | 10.2 | 2.3 | 92.5 | 7.9 | 10.1 | 2.2 | 89.0 |
| | 3.0 | 9.9 | 5.5 | 12.4 | 6.9 | 94.5 | 3.9 | 11.4 | 7.5 | 93.5 |
| | 1.5 | 7.2 | -21.5 | 36.4 | 57.9 | 90.0 | -36.2 | 43.4 | 79.6 | 93.0 |
| 20 (5) | 5.0 | 4.7 | 4.0 | 6.5 | 2.5 | 90.5 | 3.9 | 6.2 | 2.3 | 87.0 |
| | 3.0 | 4.3 | -0.2 | 8.8 | 9.0 | 94.5 | -8.7 | 12.2 | 20.9 | 94.5 |
| | 1.5 | 3.6 | -803.7 | 819.1 | >1000 | 90.5 | -34.0 | 39.8 | 73.8 | 94.5 |

See Table 4.3 for the definition of $n$, **TTP**, $\beta_0$, **E(TP), LCI, UCI, WD, CI.TP%**.

When the true turning point occurs at the end of the data (i.e. at 38 when $n = 50$, at 26 when the $n = 35$ and at 15 when $n = 20$) (see Figure 4.3(c)) and when $\beta_0$ gets smaller, the mean width of the confidence intervals gets wider in both methods. However, when $n \leq 20$ and $\beta_0 \leq 1.5$, the mean width of the confidence intervals becomes greater than the range of data for both methods. In addition, the mean width of the confidence intervals in the bootstrap and the delta methods are usually the same except when $\beta_0 \leq 1.5$. The confidence intervals constructed by the delta method have more true turning points within

95% coverage than the confidence intervals constructed using the bootstrap method. However, when the mean width of the confidence intervals is greater than the sample size in the delta method, the coverage is under 95%. The bootstrap confidence interval usually covers 95% when $\beta_0 \leq 3$ but at almost 4% less than 95% coverage when $\beta_0 \geq 5$, (see Table 4.5). Also, the confidence intervals in both methods are asymmetric and roughly close to the lower limit of the confidence interval. As a result, when the true turning point occurs roughly at the first quarter (1/4 i.e. at the beginning) or at the third quarter (3/4 i.e. at the end) of the data, the results are approximately similar.

Table 4.5: Comparing the delta and the bootstrap methods of constructing confidence intervals for the estimated turning point from the quadratic model with change in the end.

| | | | Delta method | | | | Bootstrap method | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ (TTP) | $\beta_0$ | E(TP) | LCI | UCI | WD | CI.TP% | LCI | UCI | WD | CI.TP% |
| 50 (38) | 5.0 | 38.7 | 37.3 | 39.1 | 1.8 | 94.0 | 37.4 | 39.1 | 1.7 | 91.5 |
| | 3.0 | 39.6 | 35.9 | 40.8 | 4.9 | 96.0 | 36.3 | 41.3 | 5.0 | 93.4 |
| | 1.5 | 35.4 | 32.6 | 45.4 | 12.8 | 96.0 | 32.4 | 54.3 | 21.9 | 95.5 |
| 35 (26) | 5.0 | 25.7. | 25.1 | 27.2 | 2.1 | 93.2 | 25.2 | 27.3 | 2.1 | 88.5 |
| | 3.0 | 25.1 | 23.2 | 29.7 | 6.5 | 95.5 | 24.1 | 31.2 | 7.1 | 93.5 |
| | 1.5 | 28.3 | 13.9 | 42.0 | 28.1 | 96.0 | -4.3 | 66.7 | 71.0 | 95.6 |
| 20 (15) | 5.0 | 16.1 | 14.2 | 16.5 | 2.3 | 95.0 | 14.4 | 16.6 | 2.2 | 89.0 |
| | 3.0 | 14.6 | 12.2 | 19.1 | 6.9 | 96.0 | 10.7 | 24.6 | 13.9 | 92.5 |
| | 1.5 | 13.8 | -199.5 | 227.3 | 426.8 | 91.5 | -13.5 | 47.1 | 60.6 | 96.6 |

See Table 4.3 for the definition of $n$, **TTP**, $\beta_0$, **E(TP), LCI, UCI, WD, CI.TP%**.

In conclusion, the delta method establishes good results for confidence intervals of estimated turning points in a quadratic model. The simulation study showed that both methods give roughly similar coverage for true turning points in quadratic models. Therefore, when dealing with polynomial models of an order of more than two, the delta method is mathematically complicated when there are two or more estimated turning points and as a result the bootstrap approximation is used in these cases. However, percentile bootstrap confidence intervals cannot easily cover 95% of turning points when $\beta_0 \geq 3$ for a turning point in the middle and when $\beta_0 \geq 5$ for turning points in the beginning or at

the end. This means that different bootstrap confidence interval approaches may give a better coverage.

### 4.5.2 Change in trend at two points

This section investigates the behaviour of bootstrap confidence intervals for two estimated turning points from a cubic polynomial model obtained as:

$$\log(\lambda_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \tag{4.6}$$

where $\beta_0$ takes values of 5, 3 and 1.5, $\beta_1 = (t_1 \times t_2)/10,000$, $\beta_2 = -0.5(t_1 + t_2)/10,000$, ($t_1$ and $t_2$ are the exact turning points) and $\beta_3 = 0.000033$. In this case, the $\beta_1$ and $\beta_2$ are dependent on the sample sizes and location of true turning points. Tables 4.6 and 4.7 summarise the results of estimating turning points from Equation (4.6) and their confidence intervals when two specified turning points occur roughly in the middle of the dataset and at the beginning and the end of the dataset, respectively. The initial values of $t_1$ and $t_2$ were chosen to be roughly in the middle of the dataset or roughly at the beginning and at the end of the dataset, (see Appendix B.1 for the simulation algorithm code).

When true turning points occur roughly in the middle of a dataset (i.e. at $t_1 = 16$ and $t_2 = 33$ when $n = 50$, at 11 and 24 when $n = 35$ and at 7 and 13 when $n = 20$) (see Figure 4.4(a)), the mean width of the confidence intervals of the first (WD1) and second (WD2) true turning points increase when $\beta_0$ decreases under the same sample size. However, there are no large differences between the mean widths of confidence intervals when $n = 20$. Here at all $\beta_0$ the mean widths of the confidence intervals are greater than the sample size because the simulated data are very random and the turning point can be estimated to occur anywhere on the fitted trend. The mean width of the confidence interval

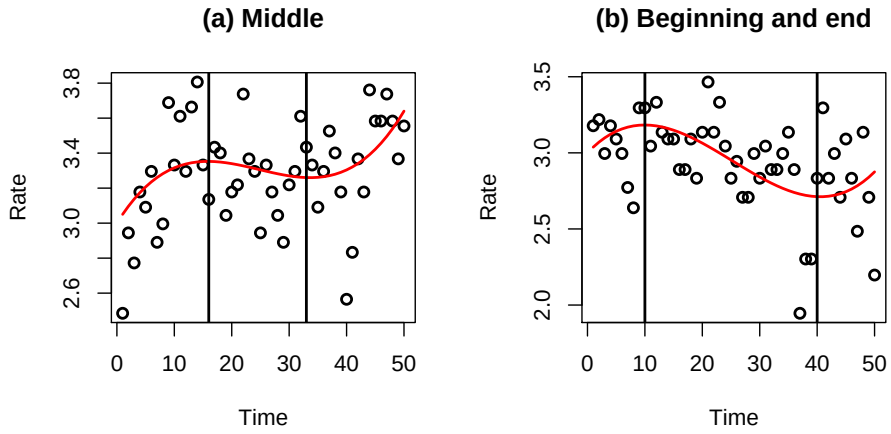



Figure 4.4: The simulated data (black circles) from the cubic model (red line) when two turning points (black lines) occur in (a) the middle, (b) the beginning and end of data.

becomes greater than the sample size when $\beta_0 \leq 1.5$ at $n = 50$ and when $\beta_0 \leq 3$ at $n = 35$ because the simulated data becomes more random when $\beta_0$ decreases which results in wide confidence intervals. The percentage of true turning points within 95% confidence intervals (CI.TP%) increases when $\beta_0$ decreases at sample sizes 50 and 35 however, for $n = 20$ the bootstrap method is not successful because the mean widths of confidence intervals are $> 20$. When the mean width of the confidence interval is less than the sample size, the confidence intervals are roughly symmetric around the true turning points when $n \geq 50$ and $\beta_0 \geq 5$. On the other hand, when $\beta_0 \leq 3$, the first true turning point is closer to the upper confidence level than it is to the lower confidence level and the second true turning point is closer to the lower confidence level than it is to the upper confidence level. When the mean widths of confidence intervals are less than the sample sizes, similar mean widths are observed for both turning points. At $n = 50$, the first true turning point is within 95% coverage where as the second true turning point is under the 95% coverage. At $n = 35$, the first and second true turning points are within 95% coverage, (see Table 4.6). This indicates that when the mean width of the confidence interval is small, the method of bootstrap confidence intervals can easily cover one true turning

point (the first one).

Table 4.6: The behaviour of confidence intervals of two estimated change points when the changes occur in the middle of the dataset.

| $n$(TTPs) | $\beta_0$ | E(TP1) | LCI1 | UCI1 | WD1 | CI.TP1% | E(TP2) | LCI2 | UCI2 | WD2 | CI.TP2% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 5.0 | 15.8 | 14.4 | 17.4 | 3.0 | 93.8 | 33.7 | 32.0 | 35.0 | 3.0 | 81.5 |
| (16, 33) | 3.0 | 16.4 | 0.4 | 23.5 | 23.1 | 95.7 | 33.2 | 25.5 | 47.4 | 21.9 | 91.2 |
| | 1.5 | 15.3 | -34.0 | 26.3 | 60.3 | 98.0 | 38.7 | 23.5 | 81.4 | 57.9 | 95.5 |
| 35 | 5.0 | 10.4 | -14.1 | 16.7 | 30.8 | 96.4 | 25.1 | 18.4 | 47.1 | 28.7 | 94.0 |
| (11, 24) | 3.0 | 10.7 | -44.2 | 19.2 | 63.4 | 98.3 | >35 | 16.2 | 78.6 | 62.4 | 98.1 |
| | 1.5 | 10.9 | -46.3 | 19.8 | 66.1 | 99.5 | 32.3 | 16.2 | 82.6 | 66.4 | 98.5 |
| 20 | 5.0 | 6.6 | -30.2 | 11.6 | 41.8 | 96.9 | >20 | 9.2 | 49.5 | 40.3 | 94.0 |
| (7, 13) | 3.0 | 6.5 | -32.9 | 11.9 | 44.8 | 96.6 | 21.5 | 9.4 | 51.4 | 42.0 | 91.3 |
| | 1.5 | 6.4 | -29.5 | 15.5 | 45.0 | 96.1 | 19.8 | 9.6 | 52.6 | 43.0 | 92.0 |

**TTPs:** Two true turning points, See Table 4.3 for the definition of $n$, $\beta_0$, **E(TP), LCI, UCI, WD, CI.TP%**.

When the true turning points occur roughly in the beginning and at the end of the dataset (i.e. at 10 and 40 when $n = 50$, at 7 and 27 when $n = 35$ and at 5 and 15 when $n = 20$) (see Figure 4.4(b)) and when $\beta_0$ decreases, the mean width of the confidence intervals (WD) increase for $n = 50$ and $n = 35$ but for $n = 20$ no significant increase is observed. The mean width of the confidence interval (WD) becomes greater than the sample size when $\beta_0 \leq 1.5$ for $n = 50$ and when $\beta_0 \leq 3$ for $n = 35$. The percentage of true turning points within 95% confidence intervals (CI.TP%) increases when $\beta_0$ decreases for $n = 50$ and $n = 35$ but for $n = 20$ the bootstrap method gives the mean widths of confidence intervals as > 20. When the mean width of the confidence interval is less than the sample size, the confidence intervals are asymmetric around true turning points. The first true turning point is closer to the upper confidence level than it is to the lower confidence level and the second true turning point is closer to the lower confidence level than it is to the upper confidence level. When the mean widths of confidence intervals are less than the sample size for $n = 50$, the first true turning point is within 95% coverage and the second true turning point is under the 95% coverage when the WD is small and is above the 95% coverage when the WD is large. For $n = 35$, the first and second true turning

points are within 95% coverage, (see Table 4.7).

Table 4.7: The behaviour of confidence intervals of two estimated change points when the changes occur at the beginning and the end of the dataset.

| $n$(TTPs) | $\beta_0$ | E(TP1) | LCI1 | UCI1 | WD1 | CI.TP1% | E(TP2) | LCI2 | UCI2 | WD2 | CI.TP2% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 5.0 | 9.9 | 7.8 | 11.3 | 3.5 | 95.0 | 40.6 | 39.2 | 42.9 | 3.7 | 87.0 |
| (10,40) | 3.0 | 9.9 | -14.0 | 17.7 | 31.7 | 96.3 | 42.4 | 30.9 | 71.1 | 40.2 | 97.4 |
| | 1.5 | 12.0 | -48.5 | 24.7 | 73.2 | 95.5 | 42.2 | 23.0 | 103.0 | 80.0 | 98.3 |
| 35 | 5.0 | 7.4 | -14.5 | 13.9 | 28.4 | 96.1 | 27.9 | 21.8 | 45.9 | 24.1 | 96.9 |
| (7, 27) | 3.0 | 9.5 | -42.5 | 19.4 | 61.9 | 96.2 | 32.6 | 16.1 | 78.2 | 62.1 | 97.1 |
| | 1.5 | 10.2 | -43.7 | 19.7 | 63.4 | 97.8 | 29.7 | 15.9 | 78.2 | 62.3 | 99.4 |
| 20 | 5.0 | 6.2 | -25.4 | 11.2 | 36.6 | 97.3 | >20 | 9.6 | 46.6 | 37.0 | 98.4 |
| (5, 15) | 3.0 | 6.5 | -30.9 | 11.5 | 42.4 | 98.9 | >20 | 9.6 | 51.8 | 42.2 | 98.9 |
| | 1.5 | 6.4 | -31.6 | 11.6 | 43.2 | 98.9 | 18.2 | 9.1 | 52.2 | 43.1 | 100.0 |

See Table 4.6 for the definition of $n$, **TTPs,** $\beta_0$, **E(TP), LCI, UCI, WD, CI.TP%**.

In conclusion, constructing accurate confidence intervals for two turning points from a cubic model using the bootstrap method (percentile confidence interval) involves that the mean width of the confidence interval should be less than the sample size and neither very narrow nor very wide which requires $35 \leq n < 50$ and $\beta_0 \geq 5$. However, when $n \geq 50$, the confidence interval often covers 95% of the first turning point only.

## 4.6 Conclusion and discussion

The main aim of this chapter is identifying changes in HAI rates and estimating when these changes occur which are the estimated turning points from polynomial GLMs. Confidence intervals for estimated turning points are constructed using the bootstrap method. Both delta and bootstrap methods showed similar finding when constructing confidence intervals for the turning point of a quadratic GLM of MRSA bacteraemia. Simulation studies were carried out to compare the delta and bootstrap methods on quadratic GLMs and showed that although the delta method was slightly better, the bootstrap method also gave reasonable results. The bootstrap method was used with a cubic GLM to con-

struct confidence intervals for estimated turning points in order to investigate the coverage and symmetrical properties of confidence intervals.

In polynomial GLMs of HAI rates, estimated turning points when the maximum estimated rates start to decrease are of interest to determine which intervention had an impact on reducing the rates of HAIs. For the MRSA bacteraemia model, the estimated turning point when the rate starts to decrease at time 2005.6 has a narrow confidence interval. Estimated turning points when the maximum rate starts to decrease are almost in the middle of the confidence interval (roughly symmetric confidence interval) because the model was fitted well and the data which was modelled showed a gradual change before and after estimated turning points (similar pattern before and after estimated turning points), (see Figure 4.1). This change is associated with some interventions which took place in Scotland such as a cleanliness course, improvement of antimicrobial prescribing policy and practice as well as other associated interventions (see Section 4.4).

Furthermore, in the quartic model of CDI in patients over 65 years, the first estimated turning point has a roughly symmetric confidence interval and it is very narrow because the trend changed steeply before and after the estimated turning point. The last two estimated turning points have asymmetric confidence intervals and were overlapping each other because the trend of the rates of CDI in patients over 65 years changes smoothly (flatter) around those two turning points. Also, there are not enough data points to estimate the turning points very precisely because the two turning points occur close to each other at the end of the dataset, (see Figure 4.2). The upper confidence level is wider than the lower confidence level because less data point are observed after the estimated turning point. The first change when the rate decreases is associated

with improving safety and reliability in health care settings under the Scottish patient safety programme (SPSP), (see Section 4.4).

The first simulation study used a quadratic GLM and compared the bootstrap and the delta methods on constructing confidence intervals. The confidence intervals for the estimated turning points of the quadratic GLM constructed by bootstrap and delta methods are similar. The application of this result was confirmed when the confidence interval was constructed for turning point of MRSA bacteraemia quadratic GLM.

From the simulation, if one change occurs roughly in the middle of the data-set, confidence intervals of estimated turning points constructed by the delta and the bootstrap methods are relatively narrow. The trend changes with the same pattern and there are roughly the same number of data points before and after the estimated turning point. However, they become larger when $\beta_0$ decreases. Confidence intervals are symmetric when $\beta_0 \geq 5$, otherwise they are asymmetric because the simulated data are more random when $\beta_0$ decreases. On the other hand, if one turning point exists in the beginning or at the end of the data and the number of data points are different before and after the estimated turning point, confidence intervals of estimated turning points are quite wide and become wider when $\beta_0$ decreases. Also, confidence intervals are asymmetric where the estimated turning points are often close to the upper confidence levels when true turning point occurs in the beginning of data because there are more data points after the change has occurred. They are usually close to the lower confidence levels when the true turning point occurs at the end of data because more data points exist before the change occurred. Therefore, both methods are good but the delta method gives enhanced 95% coverage than the bootstrap method in all cases. However, both methods are

not suitable to construct confidence intervals when $n \leq 20$ and $\beta_0 \leq 1.5$.

Since the delta method becomes more complicated when applied to more than two random variables, the bootstrap method is used to construct confidence intervals for estimated turning points from cubic and quartic GLMs. The performance of the bootstrap method with the cubic model was investigated in the second simulation study. If two changes occur roughly in the middle of the dataset or roughly at the beginning and the end, the confidence intervals for estimated turning points often cover 95% for the first turning point only when $n = 50$. These confidence intervals cover 95% for both turning points when $n = 35$ and $\beta_0 \geq 5$. However, when $\beta_0 \leq 3$ the mean width of the confidence interval becomes greater than the sample size and cannot give accurate 95% coverage because the mean width is in excess of the sample size. These confidence intervals are also asymmetric where the first turning point is close to the upper confidence level and the second turning point is close to the lower confidence level except when $n \geq 50$ and $\beta_0 \geq 5$, the confidence intervals of two turning points occurring in the middle are almost symmetric. Different number of data points before and after each turning point and the variability in simulated data almost result wide confidence interval for each estimated change point. Therefore, small sample sizes and rare numbers of cases will affect the mean width of the confidence intervals which may become larger than the range of the dataset.

## Technical issues of the bootstraps and simulations

After the bootstrapping and simulation study have been carried out and the results are interpreted, some technical issues were improved however, others need to be developed in further research.

In some simulated and bootstrapped samples, complex solutions of turning points were returned when finding the roots of quadratic and cubic equations (at most 5% of the time) where real parts from the complex numbers are the same for two turning points. In such cases, these samples were discarded and the results were calculated from the remaining samples.

Although the change occurring within the range of HAIs data is of interest, the method of estimating turning points and confidence intervals calculates all possible turning points at any time even outside the range of dataset. It is easy to pick up and present the turning points of interest and interpret their association with interventions. There is a technical issue when calculating confidence intervals of estimated turning points using the bootstrap method. In each loop, the procedure returns turning points where they are either inside or outside the range of data then the mean and standard error of turning point are calculated. This may estimate the mean and confidence levels to be outside the range of data and can affect the width of the confidence interval. Therefore, wide confidence intervals can be avoided in future research by discarding the bootstrapped samples which return at least one turning point outside the range of data.

In the simulation study when two turning points are assumed, the mean width of confidence intervals are always greater when the sample size is small at $n = 20$ compared to larger sample sizes and they are roughly similar whatever $\beta_0$ is. This happens because small sample sizes cannot fit the cubic model very well, (see Chapter 3). With poorly fitting models there is not enough power to estimate turning points from a quadratic polynomial function. This is one reason the confidence interval cannot be constructed in small sample sizes using the bootstrap method.

The percentile bootstrap confidence interval method was used to calculate confidence intervals from bootstrap samples however, this approach may not be accurate to construct such confidence intervals. This is because for one turning point and a large sample size or large $\beta_0$, the percentile bootstrap confidence interval gives narrow confidence intervals and covers less than 95%. However, with $\beta_0 \leq 3$ it often gives confidence intervals with 95% coverage of the true turning point but the mean width is wider. For two turning points, the percentile bootstrap confidence interval gives a very large mean width when $\beta_0 \leq 3$. The percentile bootstrap confidence interval usually gives 95% coverage for the first true turning point when $n \geq 50$ however, this is outwith 95% coverage for the second true turning point. For small sample sizes the percentile bootstrap confidence interval gives a mean width of greater than the sample size and this is another reason why a bootstrap confidence interval cannot be used with small sample sizes. The percentile bootstrap confidence interval may not particularly give the accurate coverage. Using different methods of bootstrap confidence intervals such as bias corrected, which takes into account the constant bias, may improve the bootstrap confidence interval method to provide 95% coverage in all different sample sizes.

In the simulation study the coefficients $\beta_1$ and $\beta_2$ of the quadratic and cubic models are different for different sample sizes or the location of true turning points. These coefficients were chosen to produce values of $\log(\lambda_n)$ to be positive as much as possible where a Poisson distribution can be used to fit a good model. Sometimes the simulation gives several zeros in simulated data because it is simulated randomly from a Poisson distribution. The zero-inflated distribution is then better to fit to such data. Our approach set a Poisson distribution to fit an appropriate model for each simulated set of data. Therefore, using the

wrong distribution to fit data with lots of zeros results in poorly fitted models and produces inaccurately estimated turning points. This often happen when $\beta_0 \leq 1.5$ and the turning point occurs in the beginning or at the end of the dataset. Choosing $\beta_1$s and $\beta_2$s to be slightly different (<0.05) in different sample sizes may give uncomparable results between different sample sizes. Different values of $\beta_2$ produce different curvature for the trend of data in each sample size. If the curvature is flat, the wide confidence interval of estimated turning point is expected. If the curvature is spiky, the narrow confidence interval of the estimated turning point is expected. The mean width of the confidence interval is associated with the curvature as well as the sample size thus it is unknown whether the sample size or curvature affects the confidence interval results. The estimated confidence intervals in different sample sizes can be compared when the coefficient of curvature is the same. To improve the procedure and have more accurate results of constructing confidence intervals using bootstrap and delta methods, identical values of $\beta_1$s and identical values of $\beta_2$s in different sample sizes should be chosen. Moreover, various values of $\beta_1$s and $\beta_2$s should be investigated.

Estimated turning points from fitted models of original data occur at minimum or maximum values but can be at different regions in the bootstrapped data. Constructing confidence intervals for the estimated turning points from HAIs data using the bootstrap method involves re-sampling the data, re-fitting the model and re-calculating the turning points. These turning points occur at minimum or maximum values but may or may not be in the same order of the original data. For example, MRSA bacteraemia fitted model (3.7) estimates three turning points where the first turning point with minimum rate occurs at 2002.63, the second turning point with maximum rate occurs at 2005.65 and the third turning point with minimum rate occurs at 2013.8. When constructing

confidence intervals for these turning points, the algorithm within the bootstrap method did not take into account that the first turning point must be a minimum value, the second turning point a maximum and the third turning point a minimum. This means that the first turning point may occur at a minimum or a maximum value which can affect the results of the confidence interval. This rarely happens in our results where MRSA bacteraemia showed that confidence intervals of the second turning point of MRSA bacteraemia has a narrow confidence interval which indicates that most of simulations get a second turning point at a maximum rate so the first and third turning points must be at minimum rates. This issue should be considered in a simulation study when small $\beta_0$ and sample size are assumed. In the simulation study the wide confidence intervals are observed with small $\beta_0$ (where the simulated data are more random) because the bootstrap method did not assume the same properties of the original turning points being at minimum or maximum values. The HAIs data were fitted well so this issue was not common. In a simulation study with small $\beta_0$ and models not fitting well, this issue can make confidence intervals inaccurate. Therefore, when the first original turning point occurs at a maximum value and bootstrap gives a turning point at a minimum value, this should be modified by rejecting samples when the wrong model is fitted. If discarding of some samples took place then confidence intervals are likely to be narrower because aberrant data gives a poorly fitted model. Our approach is more conservative and a greater width of the confidence interval is expected.

Estimated change points from polynomial models is not the only way to detect the change in the rate of HAIs. Other methods of detecting the change points in HAIs data will be discussed in the next chapter.

# Chapter 5

# Change Points Analysis

Change points analysis uses several methods to fit count data and to identify one or more statistically significant change points. It is possible to identify how many change points should be estimated. Change points statistical inference has two main issues where the first issue involves detecting the existence of change and the second involves estimating the number of change points and their locations [Chen and Gupta (2011)]. Estimated change points from polynomial GLM regression was discussed in the previous chapter. These points are located either at the time or slightly after the healthcare interventions took place. A list of interventions which took place is given in Table 1.2 and the times of these interventions are illustrated in Figure 5.1.

It is important to know which healthcare interventions impact healthcare associated infections (HAIs) and this can be established by fitting an appropriate change point model. This chapter aims to present a general approach to detect change points where infection rates change significantly and to determine which interventions are associated with these changes.

Segmented regression designs are used to discover change points where the

rates change significantly after specific interventions. The segmented regression in this research aims to find if all, some or none of these interventions have an impact on the rate of HAIs and to determine which of these interventions have had the most potential impact. Joinpoint analysis is used to estimate the existence of change points at unspecified times and estimate their location. The joinpoint method considers all data points and looks for change points whereas segmented regression only looks for the potential change points at data points that correspond to interventions.

This chapter explains segmented regression and joinpoint analyses where in Section 5.1, segmented regression analysis is explained. The method of segmented regression analysis is applied to HAI data in Section 5.1.3. Development of joinpoint analysis is illustrated in Section 5.2 where Section 5.2.3 includes the applications of the joinpoint method to HAI data. Profile likelihood and bootstrap confidence intervals are compared using a simulation study in Section 5.3. Finally, the discussion is presented in Section 5.4.

## 5.1 Segmented regression analysis

One of the methods of detecting change points is the segmented regression analysis. This method is used to estimate the impact of an intervention on the rates where they change significantly after a specific intervention [Wagner et al. (2002)]. Segmented regression analysis investigates the pattern of data before and after the intervention took place to determine whether the pattern changed significantly after the intervention. The idea of this analysis is to use several linear models to fit the data in order to find one or more statistically significant change points. Segmented regression is an approach where possible changes in trend are fixed at specific time points when interventions took place which means that a new trend (segment) will start from the same position of the end

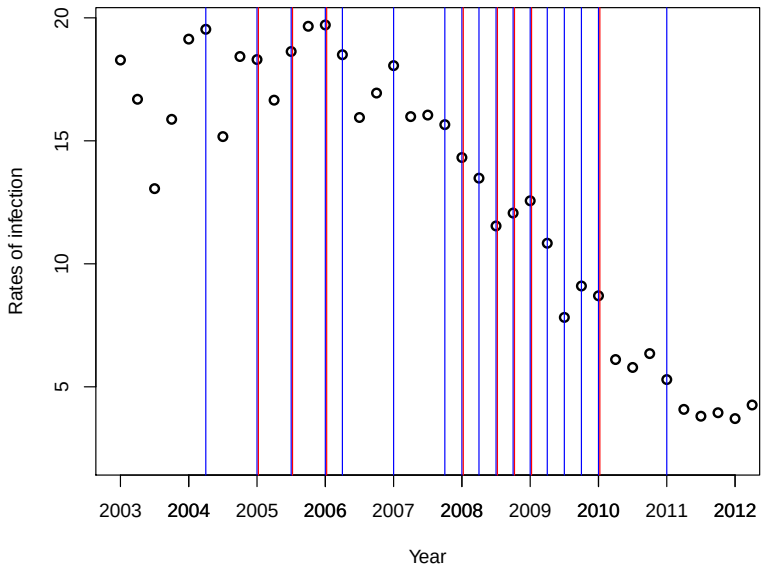



Figure 5.1: The time point of when interventions took place in Scotland. One line (blue) indicates that only one intervention took place in a quarter $i$ but two lines (blue and red) indicate that multiple interventions took place in a quarter $i$.

of the previous segment.

The Wagner et al. (2002) method is used to fit the segmented regression model. By using the Poisson model with seasonality, the general segmented regression model for the rates of HAIs is obtained as:

$$\begin{aligned} \log(\text{Cases}(t)) \sim \text{offset}(\log(\text{AOBDs}(t))) + \alpha_0 + \beta_0 t + \gamma_{Qu(t)} \\ + \alpha_1 \text{Int}(t,i) + \beta_1 t^*(i), \qquad t^*(i) = t - t(i), \end{aligned} \tag{5.1}$$

where $i$ indicates the quarter number when the intervention took place (see Table 1.2), Cases($t$) is the incidence of HAIs, AOBDs($t$) is the population at risk and $t$ is the year. Int$(t,i)$ is a dummy variable which indicates time $i$ occurring before the intervention at quarter $i$ (Int$(t,i) = 0, \quad t \leq i$) or after the intervention at quarter $i$ (Int$(i) = 1, \quad t > i$). $t^*(i)$ is a continuous variable counting the time

after the intervention at quarter $i$ (coded 0 before the intervention at quarter $i$ and $(t - t(i))$ after the intervention at quarter $i$). The coefficient $\alpha_0$ estimates the intercept (baseline level which is before the intervention at quarter $i$), the coefficient $\alpha_1$ estimates the change in the level after the intervention at quarter $i$ and $\alpha_0 + \alpha_1$ estimates the level after intervention at quarter $i$. The coefficient $\beta_0$ estimates the change in the slope of the rate of HAIs before the intervention at quarter $i$ (baseline slope), the coefficient $\beta_1$ estimates the change in the slope and the estimate of the slope after the intervention at quarter $i$ is $\beta_0 + \beta_1$. Finally, the coefficient $\gamma$ estimates the seasonal effect ($Qu$) and log(AOBDs) is an offset variable (the denominator of the rate) which has a coefficient of 1.

By fitting model (5.1) using a Poisson distribution to the MRSA bacteraemia data and choosing the intervention after two years of data (8 time points) which is at Qu1, 2005 (i.e $i = 9$ as the time of change), the level after intervention at $i = 9$ increased suddenly and the trend decreased (see Figure 5.2). This type of effect (an instantaneous increase or decrease in rates) may not be reasonable when considering interventions which potentially will modify the trend in the rates over time. Moreover, the aim of interventions is to investigate the reduction on the rate of MRSA bacteraemia, therefore a sudden rise of the rate after intervention followed by a reduction at the same time is not logical. As a result, the effect of change in the level (Int$(t, i)$) was omitted from model (5.1) to avoid the jump of the trend at the same time. The model is refitted as:

$$\log(\text{Cases}(t)) \sim \text{offset}(\log(\text{AOBDs}(t))) + \alpha_0 + \beta_0 t(i) + \gamma_{Qu(t)} + \beta_1 t^*(i). \qquad (5.2)$$

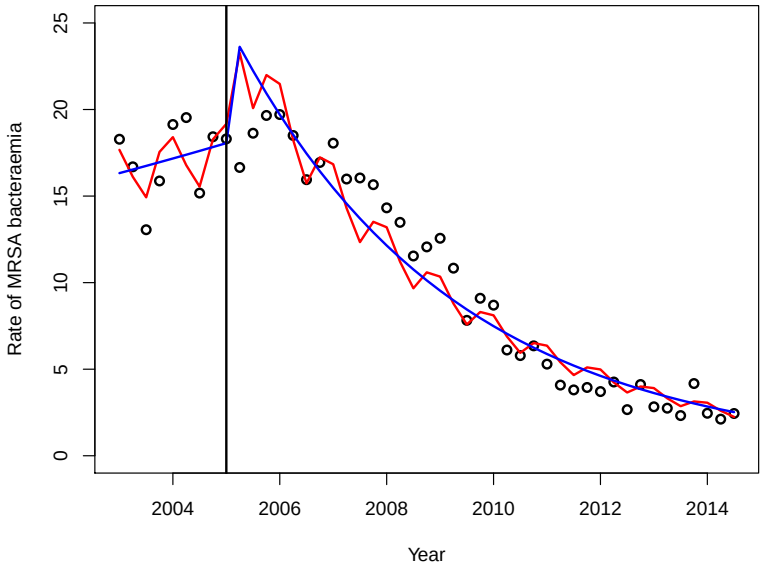



Figure 5.2: Segmented regression model (5.1) for the rate of MRSA bacteraemia with change in the slope and level at Qu1, 2005.

### 5.1.1 Determining the number of data points (time points)

If an intervention takes place towards the beginning of the data series (see Table 1.2 and Figure 5.1), the slope before the occurrence of the intervention is based upon few observations whereas the slope after the intervention has taken place is based on many observations. If the intervention is in the middle of the data, both slopes are based upon similar numbers of observations. To investigate the impact of the number of observations on the detection of a change point, three data sets are investigated.

To explain these three sets of data, MRSA data is used as an example and the number of data points is chosen in each part according to an intervention taking place roughly in the middle of all the datasets (i.e. at time point $i = 21$ (Qu1, 2008)).

### All data points before and after the intervention

The first approach is to use all data points before and after each intervention at quarter $i$ to fit the segmented regression model (5.2). The residual deviance is used as a measure of goodness of fit where the best fitted model has the lowest residual deviance. The residual deviance can be used to compare models because the same number of data points are used for each intervention to fit the model. However, fitting the linear trend cannot fit the data well over a long period of time (see Figure 5.3) and therefore a shorter period of data is investigated (see Section 5.4.1 for explanation). One or more interventions can be detected using all datasets because the same data are used each time when fitting a segmented regression model, therefore one or more interventions may impact the rate.

### Two years of data after the intervention

The second approach is to consider all data points before an intervention at quarter $i$ and eight points of data after each intervention at quarter $i$ (i.e. the data points after each intervention are always the same size (8 observations)) to fit model (5.2), (see Figure 5.3). The significance of using all data before the intervention is to make sure that there is enough data before the intervention to detect the change after the intervention. Eight points after the intervention are used to investigate the impact of an intervention on the rates of two years. Eight data points after intervention is also used to have sufficient data to detect the change and two years seems reasonable enough to detect the change. For example, when the intervention took place at time Qu1, 2008, the data from Qu1, 2003 to Qu1, 2010 is used to fit the segmented regression model. Residual deviance cannot be used to compare models as at each intervention at quarter $i$, different data points are used. The deviance is calculated based on the saturated model (a model with a parameter for every observation, see Equation 3.2) and

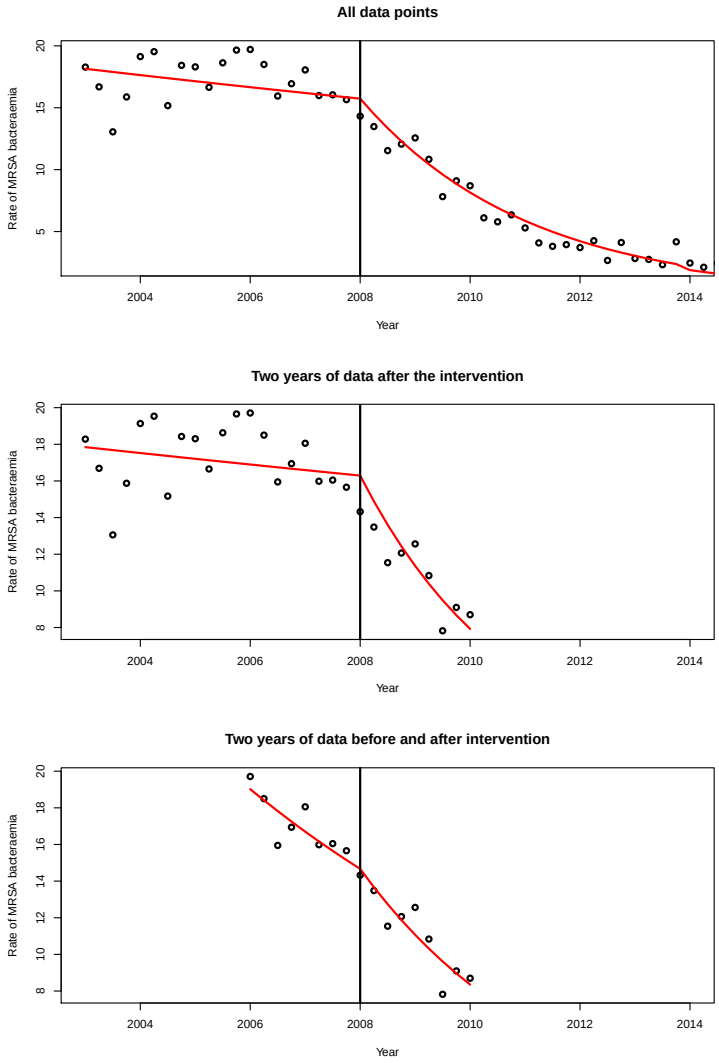



Figure 5.3: Different types of number of data points in segmented regression.

when using all data points before and after the intervention (observations are always the same in each model), the saturated model is always the same. In each model with a different point of intervention, the proposed models have the same number of parameters therefore deviances can be compared. However, when using all data before the intervention and 8 points after the intervention (observations are different in each model), the saturated model is different in each model where the intervention took place and the proposed models do not have the same number of parameters therefore, deviances are not comparable. Here the percentage of deviance explained (PDE) is used to

measure the goodness of fit.

### Percentage of deviance explained (PDE)

Percentage of deviance explained is a measure of goodness of fit of a model where this measure is analogous to $R^2$ in linear regression where the largest PDE indicates a good fit. The PDE is obtained as:

$$\text{PDE} = \left(1 - \left(\frac{\text{Deviance of proposed model}}{\text{Deviance of null model}}\right)\right) \times 100, \tag{5.3}$$

where the null model includes the intercept only which is obtained as:

$$\log(\text{rate}) = 1. \tag{5.4}$$

### Two years of data before and after intervention

The third approach uses four years of data (8 points of data before and 8 points of data after each intervention at quarter $i$) to fit the segmented regression model (5.2), (see Figure 5.3). The choice of two years of data before and two years after each intervention is to provide enough residual degrees of freedom where each model has a linear effect of time and a seasonal effect. For example, when the intervention took place at time Qu1, 2008, the data from Qu1, 2006 to Qu1, 2010 is used to fit the segmented regression model. Having the same number of data points avoids length data bias where the number of observations is equal before and after each intervention. However, the residual deviance cannot be used to compare models because at each intervention at quarter $i$ different observations are used. The saturated models are different and proposed models have the same number of parameters but the model is fitted at different observations. Again, the percentage of deviance explained (PDE) is used to measure the goodness of fit.

### 5.1.2 The algorithm of segmented regression analysis

To fit the segmented regression model to the data of HAIs using all data points, residual deviances are used as a measure of goodness of fit. The best fitted segmented model can have one or more change points, (see **R** code in Appendix C.1). The algorithm is:

1. Fit segmented regression models (5.2) with one intervention at each time when interventions have occurred and save the residual deviance of each model.

2. Choose the model with the least residual deviance and save it as the segmented regression model with one change point.

3. Fit segmented regression models with two interventions where the first intervention is fixed as the change point obtained from the previous step (2) and the second intervention is each intervention implemented after the fixed change point.

4. Choose the model with the least residual deviance and save it as the segmented regression model with two change points.

5. Use the likelihood ratio test (LRT) to test the differences between the model with one change point and model with two change points. If $p<0.05$ then the model with two change points is significantly better than the model with one change point, otherwise, the model with one change point is better.

6. Repeat the same process to fit segmented regression models with more interventions until the segmented model with $n$ interventions is not significantly different from the model with $n-1$ interventions. The model with $n-1$ interventions is the best model to detect the change in the trend of rate.

However, to fit the segmented regression model to the data of HAIs using all data points before the intervention and two years of data after intervention, the best fitted segmented model can have only one change point which has the largest percentage of deviance explained (PDE). The algorithm is:

1. Choose the subset of data according to when the intervention took place. If intervention took place at quarter $i$, then the subset data is from the first point of current data until the point $i + 8$.

2. Fit null model (5.4) based on the subset of data in step (1).

3. Fit segmented regression models (5.2) with one intervention at each time when interventions took place within the subset of data and calculate the PDE of each model as in Equation (5.3).

4. The model with the largest PDE is the best fitted segmented model.

The best fitted segmented model using two years of data before and after an intervention has a similar algorithm to the segmented regression model using all data points before the intervention and two years of data after an intervention. However, in this case if the intervention took place at quarter $i$, the subset of data is chosen within the interval $[i - 8, i + 8]$.

### 5.1.3 Segmented regression model of HAIs

In addition to modelling the rates of healthcare associated infections and estimating the change points from polynomial fitted models (see Chapter 4), change point analysis using segmented regression is another way to detect changes in the rates of HAIs. Given a series of interventions implemented from 2004 to 2011 (see Table 1.2), segmented regression models were fitted at each intervention in order to test which intervention leads to a significant change in the rate of healthcare associated infections.

### MRSA bacteraemia

In order to assess which intervention had an impact on the rate of MRSA bacteraemia, three different approaches were considered to fit model (5.2). The data was monitored quarterly as three month periods (i.e. four observed points in one year) from January 2003 to September 2014.

1. Consider all data before and after each intervention at quarter *i*. The residual deviance can be used to compare models because exactly the same number of data points were used for each intervention to fit the model. Using Poisson distribution to fit model (5.2) with each intervention found that the intervention at Qu2, 2006 gave the least residual deviance (83.60 on 41 degrees of freedom). However, given the time when the first intervention took place (Qu2, 2006), the second intervention at Qu2, 2008 gave a model with two interventions and the residual deviance is 57.68 on 40 degrees of freedom. The model with two interventions is significantly different from the model with one intervention ($p<0.001$). Also, the model with three interventions at Qu2, 2006, Qu2, 2008 and Qu1, 2011 is significantly better than the model with two interventions ($p<0.001$). The segmented model with three interventions is the best model (residual deviance is 47.94 on 39 degrees of freedom) to fit the data and to describe the change. (See Table 5.1 and Figure 5.4).

2. Consider all data points before an intervention at quarter *i* and two years of data points (8 observations) after each intervention at quarter *i*. Using a Poisson distribution, the fitted model (5.2) with intervention at Qu2, 2009 has the largest PDE=87.2%, (see Table 5.1). The change point then occurs at Qu2, 2009 when the rate of MRSA bacteraemia changed.

3. Consider two years of data (8 observations) before and two years of data after each intervention at quarter *i*. Percentage of deviance explained

(PDE) is used to measure the goodness of fit and by using a Poisson distribution, the fitted model (5.2) with intervention at Qu3, 2009 has the largest PDE=98.2% with significant change in slope, (see Table 5.1). The intervention at Qu3, 2009 was seen to have an impact on the rate of MRSA bacteraemia.

Table 5.1: Change points results of segmented regression analysis.

| Infection | Segmented points All data | Segmented point All B and 2 years A | Segmented point 2 years B and A |
|---|---|---|---|
| MRSA bacteraemia | Qu2, 2006<br>Qu2, 2008<br>Qu1, 2011<br>DV:47.94, DF:39 | Qu2, 2009<br><br><br>PDE=87.2% | Qu3, 2009<br><br><br>PDE=98.2% |
| MSSA bacteraemia | Qu2, 2006<br>DV:86.65, DF:32 | Qu2, 2006<br>PDE=48.1% | Qu2, 2009<br>PDE=63% |
| CDI in patients over 65 years | Qu4, 2007<br>Qu4, 2009<br>DV:460.16, DF:25 | Qu2, 2008<br><br>PDE=97.5% | Qu4, 2009<br><br>PDE=98.9% |
| CDI in patients aged 15-64 years | Qu4, 2009<br>DV:43.76, DF:16 | Qu4, 2009<br>PDE=86.3% | Qu4, 2009<br>PDE=86.3% |

**A:** After, **B:** Before, **Qu:** Quarter, **DV:** Residual deviance, **DF:** Degrees of freedom, **PDE:** Percentage of deviance explained.

### MSSA bacteraemia

Fitting segmented regression models using all data of MSSA bacteraemia from April 2005 to September 2014 before and after each intervention at quarter $i$ with quasi-Poisson distribution gave 86.6525 residual deviance on 32 degrees of freedom at Qu2, 2006, (see Figure 5.5). However, the coefficient of slope after the intervention is not significant (p=0.098) which indicates no evidence of change. If is considered there is an intervention at Qu2, 2006, there is no evidence of a second intervention (p>0.05). Poisson regression was used to fit all data before an intervention at quarter $i$ and 8 points of data after each intervention and 48.1% percentage of deviance was explained at Qu2, 2006 with significant change in slope after the intervention (p<0.05). However, the PDE=63% at

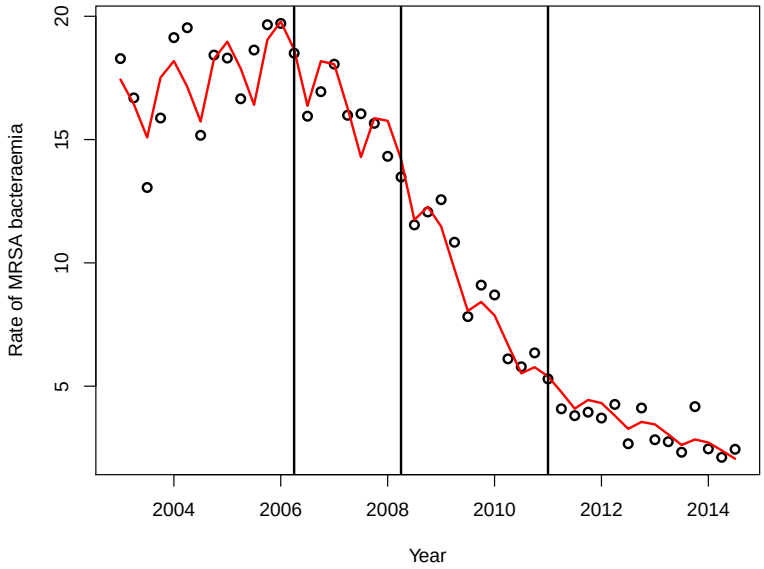



Figure 5.4: Segmented regression model for the rate of MRSA bacteraemia with three change points when using all data points.

Qu2, 2009 when two years of data before and after each intervention at quarter *i* was assumed but the coefficient of slope after intervention at Qu2, 2009 is not significant different from from the slope before the intervention, (see Table 5.1).

### CDI in patients over 65 years

Fitting segmented models with quasi-Poisson regression using data from October 2006 to September 2014 identifies the change points. Using all data before and after each intervention at quarter *i* gave two significant change points at Qu4, 2007 and at Qu4, 2009 with residual deviance of 460.16 on 25 degrees of freedom, (see Figure 5.6). However, using Poisson regression to fit all data before an intervention at quarter *i* and 8 points of data after each intervention at quarter *i* showed that Qu2, 2008 had the largest PDE=97.5%. Fitting segmented models with one change point and two years of data before and two years after each intervention at quarter *i* showed Qu4, 2009 had the largest PDE=98.9%, (see Table 5.1).

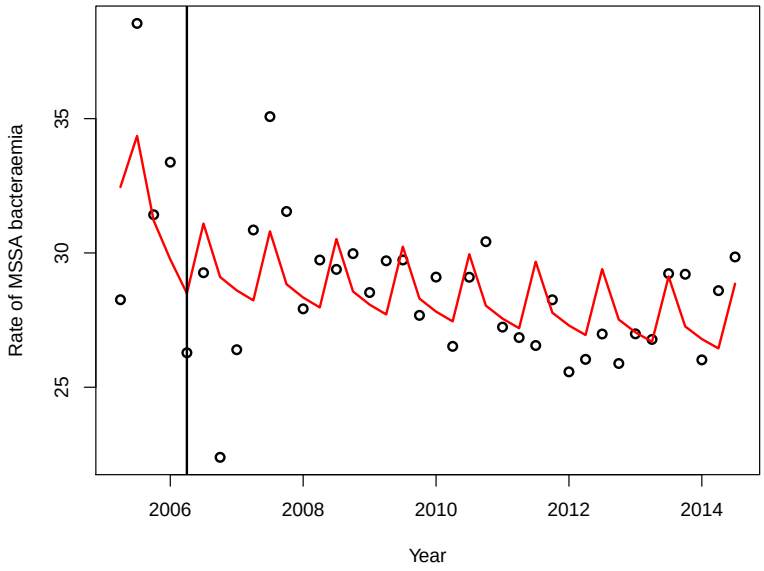



Figure 5.5: Segmented regression model for the rate of MSSA bacteraemia with one change point when using all data points.

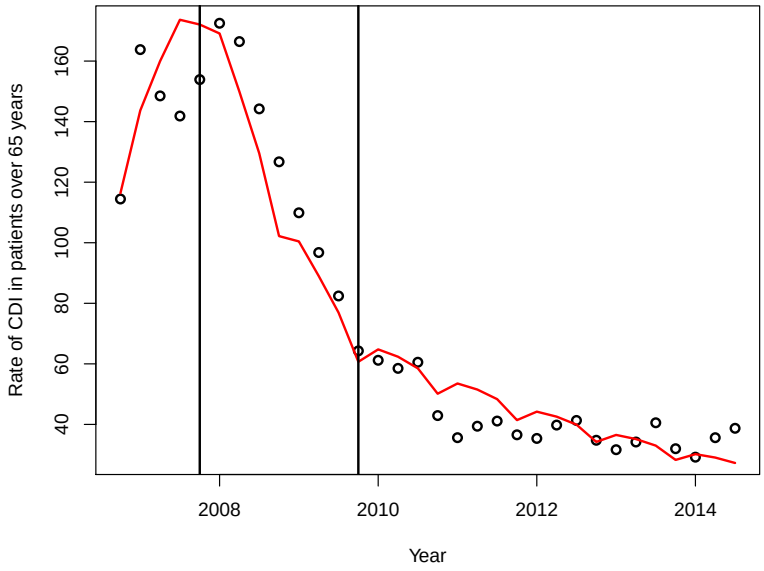



Figure 5.6: Segmented regression model for the rate of CDI in patients over 65 years with two change points when using full dataset.

### CDI in patients aged 15-64 years

Fitting segmented models with Poisson distribution to the data of CDI in patients aged 15-64 years gave the following results. Using all data from April 2009 to September 2014 before and after each intervention at quarter $i$ gave the model with Qu4, 2009 as the smallest residual deviance of 43.76 on 16 degrees of freedom was observed, (see Figure 5.7). Using all data before an intervention at quarter $i$ and 8 points of data after each intervention at quarter $i$ showed the same result when using two years of data before and after each intervention at quarter $i$ where Qu4, 2009 has the largest PDE=86.3% with significant parameter of change after the intervention, (see Table 5.1).

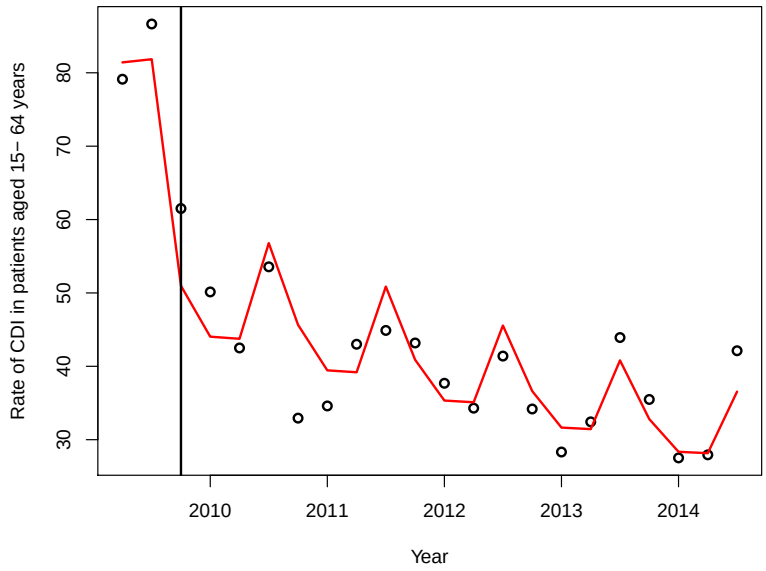



Figure 5.7: Segmented regression model for the rate of CDI in patients aged 15-64 years when using all data points.

In conclusion, segmented regression analysis describes and detects the change in trend of HAI rates when the interventions took place. Using the full dataset (all data points) to describe the trend shows that for long periods of time, more than one intervention had an impact on the rate of HAIs (see Section 5.4.5 which explains the detected change points in relation to the interventions). However,

choosing a subset of the data points (e.g. two years before and two years after the intervention) that is related to the intervention detects one change point because different subsets are used to detect different interventions. Small subsets show good fitted models (linear trends) but give less accuracy to compare models because different sets of data are used at each intervention.

If using all the data points detects more than one intervention, one of these interventions is likely to equate or is similar to an intervention identified by the segmented regression model from a subset of the data. For example, all data before and two years after the intervention detects a point close (one year after where this is a short time relative to the time that it may take for an intervention to have an impact) to the second change point from the full data for MRSA bacteraemia and detects the change point from the full dataset for MSSA bacteraemia. Also, it detects a point that is close (two quarters after) to the first change point from the full data for CDI in patients over 65 years and detects the change point from the full data for CDI in patients aged 15-64 years. Moreover, two years of data before and two years after the intervention detects a point close (five quarters after) to the second change point from the full data for MRSA bacteraemia. It also detects the second change point from the full dataset for CDI in patients over 65 years. In conclusion, the time when specific interventions took place may are not the time of change on the rate where the significant change may occur after the intervention had taken place. The next section is looking for the significant change at all possible data points and then attribute any change to the intervention had taken place.

## 5.2 Joinpoint analysis

Joinpoint analysis is used to estimate change points when the rate changes at unspecified times (i.e. at all possible points in the dataset). Using joinpoint

statistical software [NCISR (2017)] shows the trend of data and fits the simplest joinpoint model (includes year only) that the data allow. The minimum and maximum number of joinpoints is given and the program starts with the minimum number of joinpoints and tests whether more joinpoints are statistically significant and therefore can be added to the model (up to the maximum number). The permutation method (test of significance) is used to select the number of joinpoints. A grid search method considers every possible change point and searches for the minimum deviance. It has a discrete number of locations that are tested to find the best fitting model and joinpoints occur exactly at one or more times when the rates change [Kim et al. (2000)]. Using a similar idea to Kim et al. (2000), the joinpoint model was modified to include the seasonality as a factor and algorithm is rewritten in **R** programming language [R Core Team (2014)].

### 5.2.1 Joinpoint analysis algorithm

The algorithm of the joinpoint analysis is as follows, (see **R** code in Appendix C.2.1):

1. Set minimum (min) and maximum (max) number of joinpoints with assumption of max-min $\geq 2$.

2. Use Equation (5.2) with Poisson distribution to fit the null models with the minimum number of joinpoints and alternative models with the maximum number of joinpoints to the data. Here $i$ in Equation (5.2) indicates the number of time points (i.e. all possible time points during period of study except first and last points).

3. Find the best fitted model from null models and the best fitted model from alternative models by using a grid search method considering one point of data between every possible joinpoints.

4. Permute the residuals from the chosen null model (100 times) and use them to get permuted counts.

5. The permuted counts are calculated using Pearson residuals and are obtained from:

$$\text{Permuted counts} = \text{EN} + \text{Pearson residuals} \times \sqrt{\text{EN}}, \qquad (5.5)$$

   where EN is the expected values of the null model.

6. Use permuted counts to fit the alternative model and find the smallest permuted deviances.

7. The permutation test is calculated to find the p-value for accepting or rejecting the null model.

   a) Find the change in deviance from the original data ($X$= deviance of null model - deviance of alternative model).

   b) Find the change in deviance from permuted data ($Y$= deviance of null model - permuted deviances of alternative model).

   c) Calculate the p-value of the permutation test as $\sum I(Y > X)/(\text{length}(Y)+1)$ where $I(Y > X) = 1$; $Y > X$ and 0; otherwise.

8. Use Bonferroni correction to adjust the significance level of 5% [Kim et al. (2000)]. If p<0.05/(max-min), reject the null model and accept the alternative model then do the same analysis using the same alternative model and set the null model as (minimum number of joinpoints+1).

9. If p>0.05/(max-min), accept the null model and reject the alternative model then do the same analysis using the same null model and set the alternative model as (maximum number of joinpoints-1).

10. Repeat the analysis until the difference between the null model and alternative model becomes 1 (i.e. max-min=1).

11. The number of joinpoints and their location can then be estimated from the last accepted model.

## 5.2.2 Constructing confidence intervals for joinpoints

After detecting joinpoints using the above technique with a grid search method and permutation test, a confidence interval for the joinpoint is obtained. Two different methods were used to construct confidence intervals for joinpoints in count data; profile likelihood confidence intervals and bootstrap confidence intervals.

### 5.2.2.1 Profile likelihood confidence interval for one joinpoint

The profile likelihood confidence interval is based on the asymptotic chi-square distribution of the log-likelihood ratio test statistic. The 95% confidence interval for the joinpoint can then be computed by adding $\chi^2_{(0.95,1)} = 3.84$ to the minimum value of the residual deviance function which was calculated from each joinpoint model (residual deviance for each joinpoint model), where the deviance is defined in Equation (3.2) [Royston et al. (2007)]. **R** code was written to plot and calculate the confidence interval for estimated joinpoint (see Appendix C.2.3). Residual deviances from joinpoint models were calculated and the curve of the deviance function (Y axis is the residual deviances of joinpoint models and X axis is the set of estimated joinpoints) is plotted. The curve is roughly polynomial with a single smallest minimum where deviances start to decrease and at the smallest deviance (estimated joinpoint has the smallest residual deviance) start to increase again. A horizontal line *hl* is plotted based on the minimum deviance plus 3.84 and two points of intersection between the curve of deviance function and the horizontal line are identified. These two

points are the lower and upper confidence limits of the estimated joinpoint (see Figure 5.8). This method cannot be used for constructing confidence intervals for more than one joinpoint. It is based on the value of deviance and the estimated joinpoint from the joinpoint model. If the model has two estimated joinpoints associated with one value of deviance, the deviance function is then related to two joinpoints and the curve of the deviance function cannot be plotted.

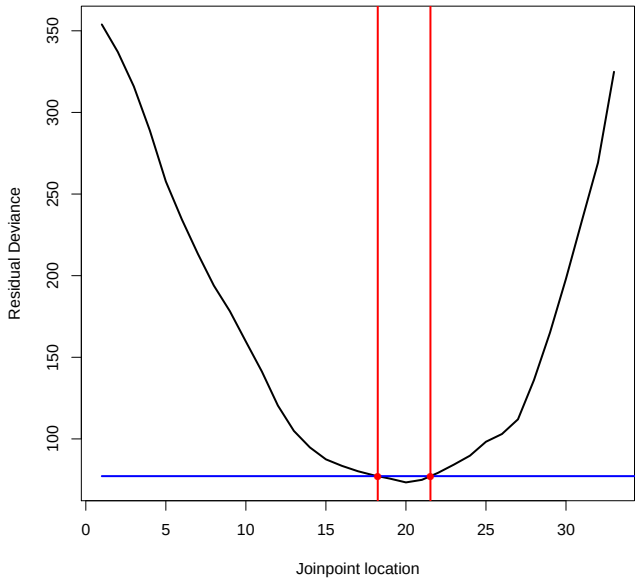



Figure 5.8: Profile likelihood confidence interval for the joinpoint. A black curve is based on the residual deviance of each joinpoint model, the blue line is a horizontal line and the confidence interval of estimated joinpoint in red lines. The estimated joinpoint is 20 with 95%CI (18.2389, 21.5311).

#### 5.2.2.2 Bootstrap confidence interval for joinpoints

Since the best joinpoint model is fitted and defines the number and location of joinpoints, the 95% confidence intervals for joinpoints can be constructed by bootstrapping. The 95% confidence intervals for joinpoints were calculated by using the algorithm of joinpoint within the bootstrap and then using the

function **quantile** at 0.025 and 0.975 in **R**. The algorithm for constructing confidence intervals by bootstrapping semi-parametric re-sampling is as follows (see Appendix C.2.3):

1. Use the last accepted joinpoint model using **glm** as explained in Section 5.2.1. This model is called the original joinpoint model.

2. Obtain the fitted values $\hat{y}_i$ (using function **predict** in **R**) and Pearson residuals $\hat{\varepsilon}_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$ (using function **residuals** in **R**).

3. For one bootstrap, re-sample the residuals and save a new response variable $y_i^*$ by adding the re-sampled residuals to the fitted values as $y_i^* = \hat{y}_i + \hat{\varepsilon}_i \times \sqrt{\hat{y}_i}$.

4. Round $y_i^*$ to be integer values.

5. Use the new responses (bootstrap data) to fit the best fitted joinpoint model with the same number of joinpoints as in the original model by using **glm** function in **R**. This step is searching only for the location of the joinpoints where the number of joinpoints is fixed as in the original joinpoint model. For example, if the original model has two joinpoints, then the algorithm searches for the best fitted model (smallest deviance) with one joinpoint then given the first joinpoint, the algorithm searches for the best fitted model with the second joinpoint.

6. Save the location of joinpoints from the best fitted model. These points are called bootstrapped joinpoints.

7. Repeat steps 3 to 6, 1000 times to obtain bootstrapped joinpoints.

8. Calculate 95% confidence intervals for the original joinpoints by using the bootstrapped joinpoints to find (**quantile**) at 0.025 and 0.975 in **R**.

### 5.2.3 Joinpoint analysis of HAIs

Joinpoint regression analysis is used to estimate change points from full datasets. It is used to identify the best fitting model where a statistically significant change in the trend of rates over time occurred. The analysis starts with the minimum number of joinpoints (zero), and tests whether one or more joinpoints are statistically significant and should be added to the model (up to three joinpoints). Each joinpoint in the final model indicates a statistically significant change in the trend. Adding seasonality to the previous model of joinpoint [NCISR (2017)] may give different results on MRSA bacteraemia, MSSA bacteraemia, CDI in patients over 65 years and CDI in patients aged 15-64 years.

#### MRSA bacteraemia

Using joinpoint software [NCISR (2017)] to analyse the rate of MRSA bacteraemia in Scotland from January 2003 to September 2014, the joinpoint model will be linear trends on the log rate of MRSA bacteraemia. The minimum number of joinpoints was fixed at 0 and the maximum at 3. The maximum number is chosen to be three where short period of data points does not expect more than three and previous method (polynomial method) detected three turning points only. Figure 5.9 shows the fitted joinpoint model which used a grid search method with uncorrelated errors. There is one significant change point at Qu1, 2007.

The joinpoint model was developed by adding the seasonality factor to the previous model (model with year only). By applying the method in Section 5.2, one significant joinpoint at Qu2, 2007 (residual deviance is 64.352 on 41 degrees of freedom) was observed in the trend of MRSA bacteraemia, see Table

5.2 and Figure 5.10.

A 95% profile likelihood confidence interval for MRSA bacteraemia joinpoint is (16.21, 19.22) ≈ (Qu4, 2006- Qu3, 2007) and 95% bootstrap confidence interval is (17, 19) ≈ (Qu1, 2007- Qu3, 2007). These are approximately similar however the bootstrap confidence interval is slightly narrower and needs much more computation than the profile likelihood confidence interval.

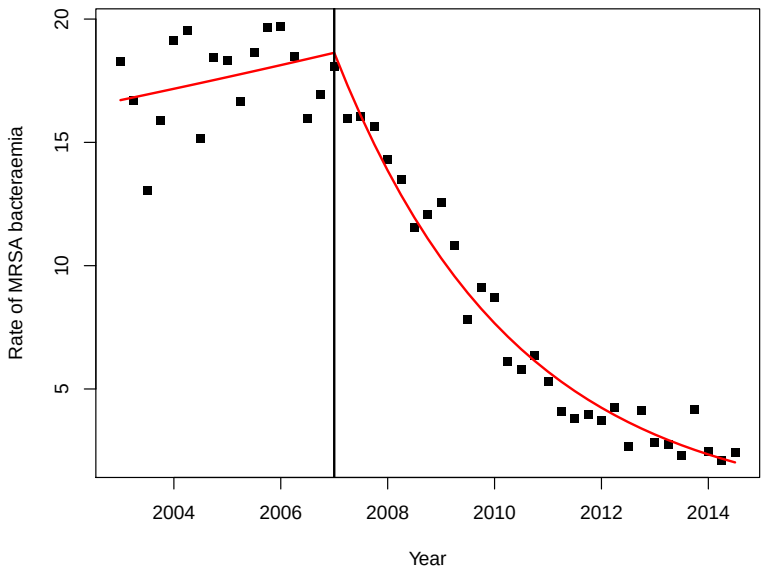



Figure 5.9: Joinpoint model for the rate of MRSA bacteraemia using joinpoint software ([NCISR (2017)]). The best fitted line (red line) with joinpoint at Qu1, 2007 (black line) and black squares are the observed rate.

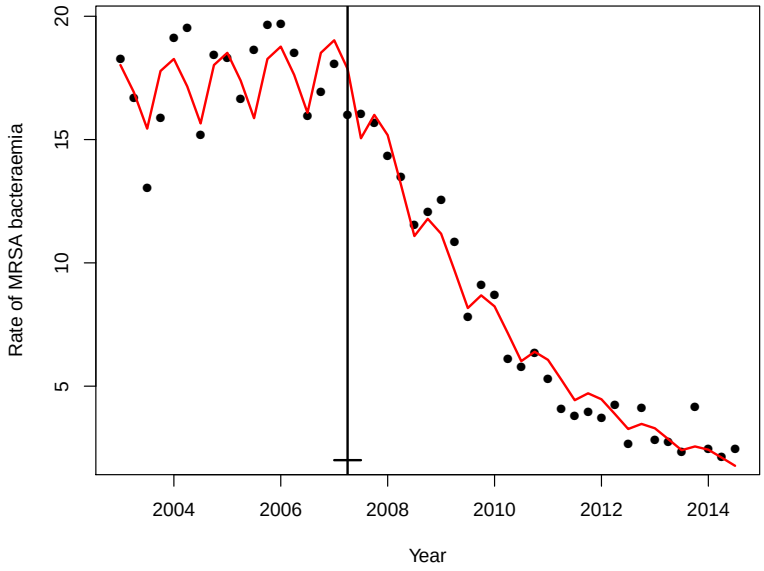



Figure 5.10: Joinpoint model for the rate of MRSA bacteraemia using joinpoint developed method. The best fitted line with seasonal effect (red line) with joinpoint at Qu2, 2007 and 95% bootstrap confidence interval (black lines) and black circles are the observed rate.

Table 5.2: Change points results of joinpoint analysis.

| Infection | JPs by NCISR | JPs | Confidence interval for joinpoint | |
|---|---|---|---|---|
| | | | Bootstrap | Profile likelihood |
| MRSA bacteraemia | Qu1, 2007 | Qu2, 2007 | (Qu1, 2007 - Qu3, 2007) | (Qu4, 2006 - Qu3, 2007) |
| CDI in patients over 65 years | Qu2, 2008<br>Qu1, 2011 | Qu2, 2008<br>Qu1, 2011 | (Qu1, 2008 - Qu3, 2008)<br>(Qu4, 2010 - Qu2, 2011) | |
| CDI in patients aged 15-64 years | Qu2, 2010 | Qu3, 2010 | (Qu2, 2010 - Qu1, 2011) | (Qu2, 2010 - Qu4, 2010) |

**JPs:** Joinpoints, **Qu:** Quarter.

### MSSA bacteraemia

For MSSA bacteraemia, using the joinpoint software model and the developed joinpoint model with seasonality did not identify any joinpoints between April 2005 and September 2014. This result is equivalent to the fitted model (3.8) where the log rate of MSSA bacteraemia has a linear trend.

### CDI in patients over 65 years

Applying joinpoint software and the developed joinpoint model with seasonal effect to the data of CDI in patients over 65 year where the data ranged from October 2006 to September 2014, two joinpoints were found. The first joinpoint is at Qu2, 2008 and the second is at Qu1, 2011 with residual deviance 138.86 on 25 degrees of freedom, (see Figures 5.11 and 5.12). A 95% bootstrap confidence intervals for the first joinpoint is (Qu1, 2008 - Qu3, 2008) and for the second joinpoint is (Qu4, 2010 - Qu2, 2011), (see Table 5.2). The profile likelihood confidence interval cannot be used here because two joinpoints are detected.

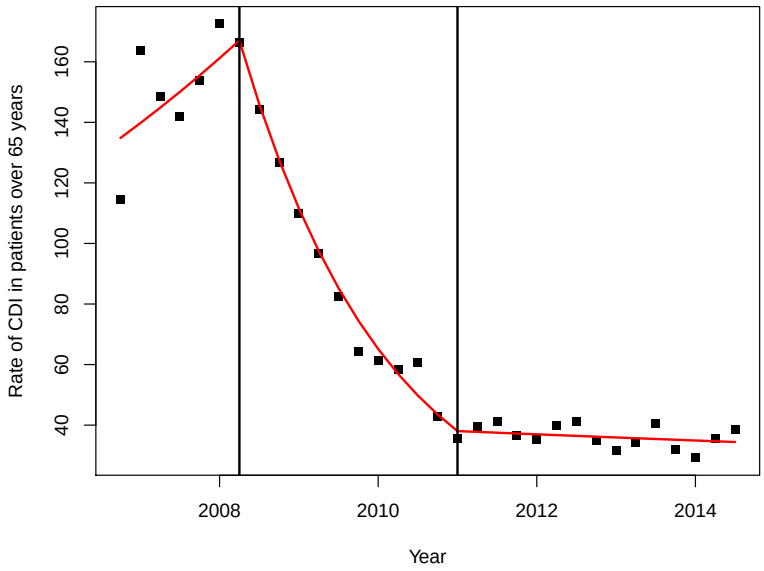



Figure 5.11: Joinpoint model for the rate of CDI in patients over 65 year using joinpoint software ([NCISR (2017)]. The best fitted line (red line) with joinpoint at Qu2, 2008 and Qu1, 2011 (black lines) and black squares are the observed rate.

### CDI in patients aged 15-64 years

For data of CDI in patients aged 15-64 years from April 2009 to September 2014, the joinpoint software [NCISR (2017)] found one joinpoint at Qu2, 2010, (see Figure 5.13). Using the joinpoint model with seasonal effect gave one joinpoint

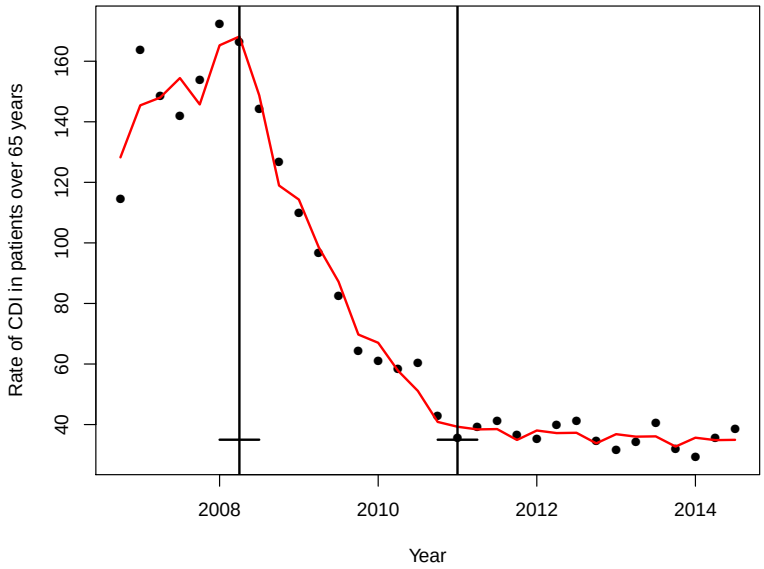



Figure 5.12: Joinpoint model for the rate of CDI in patients over 65 year using joinpoint developed method. The best fitted line with seasonal effect (red line) with joinpoints at Qu2, 2008 and Qu1, 2011 and 95% bootstrap confidence interval (black lines) and black circles are the observed rate.

at Qu3, 2010 with residual deviance 20.57 on 16 degrees of freedom, see Figure 5.14 and Table 5.2.

A 95% bootstrap confidence interval for the joinpoint is (Qu2, 2010- Qu1, 2011) and 95% profile likelihood confidence interval is (29.74, 32.30) ≈ (Qu2, 2010- Qu4, 2010). The profile likelihood confidence interval is slightly narrower than the bootstrap confidence interval.

In conclusion, seasonality affects the location of joinpoints where the rates of MRSA bacteraemia and CDI in patients aged 15-64 years changed significantly. The developed joinpoint model shows the importance of seasonality on the rate of infections. The detected joinpoints in relation to the interventions are explained in Section 5.4.5.

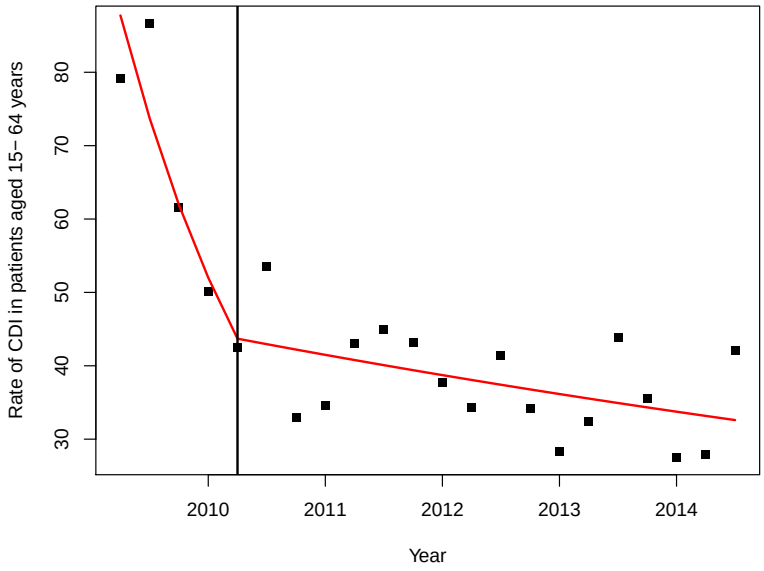



Figure 5.13: Joinpoint model for the rate of CDI in patients aged 15-64 years using joinpoint software ([NCISR (2017)] ). The best fitted line (red line) with joinpoint at Qu2, 2010 (black line) and black squares are the observed rate.

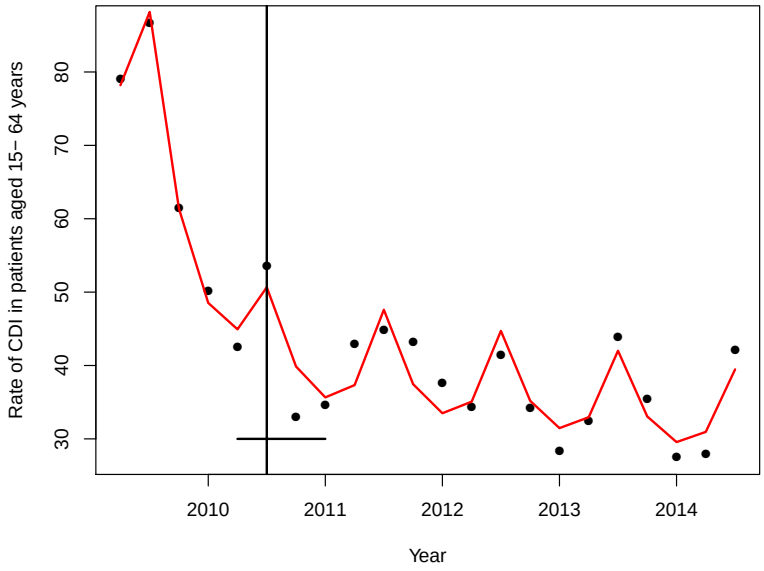



Figure 5.14: Joinpoint model for the rate of CDI in patients aged 15-64 years using joinpoint developed method. The best fitted line with seasonal effect (red line) with joinpoint at Qu3, 2010 and 95% bootstrap confidence interval (black lines) and black circles are the observed rate.

## 5.3 Comparing profile likelihood and bootstrap methods for confidence intervals of one joinpoint

Profile likelihood and bootstrap methods are used to construct confidence intervals for change points from joinpoint models. In the results of HAIs, although both methods showed similar results, one method gives wider confidence interval than the other method with different infections. A simulation study was carried out to determine which method is better to use (see algorithm code in Appendix C.2.4). Based on 200 simulations and a significance level of 5%, the range (92%- 98%) is consistent with 95% coverage (i.e. $(0.95 \pm z_{1-(\alpha/2)} \sqrt{0.95 \times 0.05/200}) \times 100 \approx (92\% - 98\%), \quad z_{1-(\alpha/2)} = 1.96$).

Different sample sizes (50, 35 and 20) were assumed to simulate data according to the initial model. The initial models are considered with one true turning point ($x_0$) where $x_0$ can occur in the beginning, at the end or in the middle (see Figure 5.15). The original models are as follows:

1. Quadratic polynomial model where the pattern of the data is curved with one true turning point,

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2,$$

   where $\lambda_n$ is the observed data, $\beta_0$ is the intercept which is assumed to be 5, $\beta_1$ is the coefficient of the slope (linear term) and $\beta_2$ is the coefficient of quadratic term.

2. Segmented regression model,

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)I(x),$$

where $\beta_1$ is the coefficient of the slope before the change, $\beta_2$ is the coefficient of change in the slope and $I(x)$ is an indicator function of $x$ where,

$$I(x) = \begin{cases} 1 & x \geq x_0, \\ 0 & x < x_0. \end{cases}$$

3. Combined model (quadratic and segmented with one turning point),

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \beta_3(x - x_0)I(x),$$

where $\beta_1$ is the coefficient of the slope (linear term) , $\beta_2$ is the coefficient of quadratic term and $\beta_3$ is the coefficient of change in the slope.

Different values of $\beta_i$s are assumed. $\beta_0 = 5$ is chosen in all cases of sample size, original model and the location of $x_0$. $\beta_1$s are identical and $\beta_2$s are identical when original models are quadratic and combined in all cases of sample size and location of $x_0$ (i.e. $\beta_1 = 0.001$ and $\beta_2 = -0.003$). However, when original model is segmented, $\beta_1 = 0.005$ and $\beta_2 = -0.03$. In a combined model, the coefficient of slope after change ($\beta_3 = -0.01$) is slightly different from the coefficient of slope after change in segmented regression ($\beta_2 = -0.03$).

The mean width of the confidence interval (WD) and the percentage of true turning points within the confidence interval (CI.TP%) were used to determine the best method of constructing confidence interval of one joinpoint. By using **R** programming language, the algorithm for simulation study is as follows:

1. Generate data according to Poisson distribution,

$$Y_n \sim \text{Poisson}(\lambda_n),$$

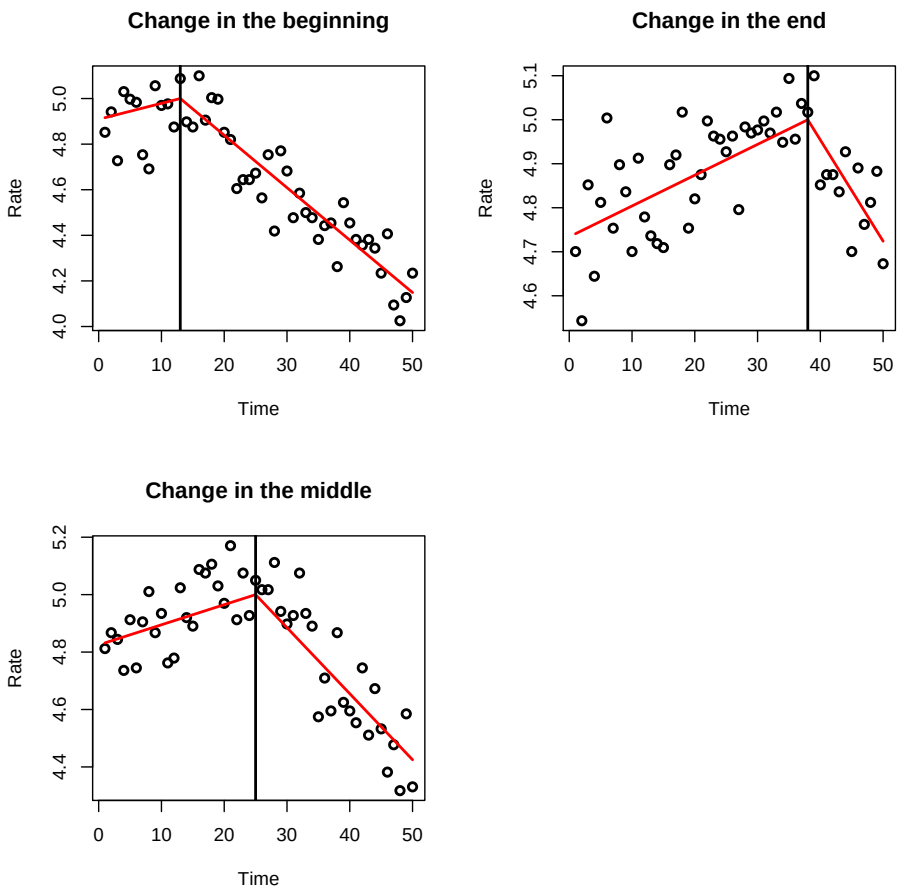



Figure 5.15: The locations of true turning point when the data is simulated (black circles) from segmented original model (red line).

where $Y_n$ is a simulated count data, $\lambda_n$ is the observed data of the original model (quadratic, segmented or combined) and $n$ is the sample size.

2. Use simulated data to estimate the best joinpoint model with one joinpoint which has the smallest deviance (see Section 5.2.1).

3. Save the simulated value of joinpoint from the simulated joinpoint model in step 2.

4. Use the profile likelihood method to construct a confidence interval for the simulated value of joinpoint (see Section 5.2.2.1) from step 3 and save the results.

5. Use the bootstrap method (500 times) to construct a confidence interval for the simulated value of joinpoint (see Section 5.2.2.2) from step 3 and save the results.

6. Repeat the simulation (from step 1 to step 5) 200 times to have sets of confidence intervals from both profile likelihood and bootstrap methods.

7. Use the mean width of the simulated confidence intervals (WD) and the percentage of simulated confidence intervals including the true turning point $x_0$ to compare two methods.

## Results of the simulation

Table 5.3 shows the results of profile likelihood and bootstrap confidence intervals for the joinpoint modelling when the true turning point occurs in the middle, beginning and at the end of the dataset. Not applicable (NA) is recorded in Table 5.3 for the profile likelihood method in the case of a small sample size ($n < 35$), (see Section 5.4.3 for explanation). The results of the profile likelihood method are also not applicable in small number of cases ($\beta_0 < 5$) therefore, $\beta_0$ is chosen to take the value of 5 only.

When the true turning point occurs in the middle, beginning or at the end of data and if $n \geq 35$ and the data is simulated from a quadratic or combined models, the mean width (WD) of the bootstrap confidence interval is less (with small differences) than the mean width of profile likelihood confidence interval. When $n \leq 20$, the profile likelihood method cannot construct a confidence interval (see Section 5.4.3) so a comparison was not made between profile and bootstrap methods for small sample sizes ($n \leq 20$). However, if the data is simulated from a segmented model, the mean width of the profile likelihood confidence interval is less than the mean width of the bootstrap confidence

Table 5.3: Profile likelihood and bootstrap confidence intervals for estimated change points from the joinpoint model.

| Change in the middle | | Profile likelihood method | | Bootstrap method | |
|---|---|---|---|---|---|
| O.M | Sample size (TTP) | WD | CI.TP% | WD | CI.TP% |
| Quadratic | 50 (25) | 2.5 | 63 | 2.2 | 44 |
| | 35 (18) | 3.6 | 72 | 3.0 | 46 |
| | 20 (10) | NA | NA | 5.9 | 59 |
| Segmented | 50 (25) | 6.8 | 93 | 7.8 | 90 |
| | 35 (18) | NA | NA | 11.6 | 94 |
| | 20 (10) | NA | NA | 13.2 | 93 |
| Combined | 50 (25) | 2.3 | 70 | 2.0 | 33 |
| | 35 (18) | 3.2 | 80 | 2.7 | 44 |
| | 20 (10) | NA | NA | 5.3 | 63 |
| Change in the beginning | | | | | |
| Quadratic | 50 (13) | 3.0 | 0 | 2.4 | 0 |
| | 35 (9) | 3.6 | 0 | 3.4 | 0 |
| | 20 (5) | NA | NA | 6.7 | 31 |
| Segmented | 50 (13) | 8.9 | 89 | 11.3 | 86 |
| | 35 (9) | NA | NA | 19.2 | 93 |
| | 20 (5) | NA | NA | 15.4 | 95 |
| Combined | 50 (13) | 2.9 | 0 | 2.3 | 0 |
| | 35 (9) | 3.6 | 0 | 3.4 | 0 |
| | 20 (5) | NA | NA | 6.6 | 25 |
| Change at the end | | | | | |
| Quadratic | 50 (38) | 2.9 | 0 | 2.3 | 0 |
| | 35 (26) | 4.0 | 0 | 3.3 | 0 |
| | 20 (15) | NA | NA | 6.5 | 32 |
| Segmented | 50 (38) | 8.4 | 90 | 10.1 | 92 |
| | 35 (26) | NA | NA | 15.8 | 93 |
| | 20 (15) | NA | NA | 14.7 | 92 |
| Combined | 50 (38) | 2.9 | 0 | 2.2 | 0 |
| | 35 (26) | 3.6 | 0 | 3.3 | 0 |
| | 20 (15) | NA | NA | 6.1 | 34 |

**O.M:** Original model, **TTP:** True turning point, **WD:** Mean width of confidence interval, **CI.TP%:** Percentage of true change points within the confidence interval, **NA:** Not applicable (no confidence interval is constructed by profile likelihood method).

interval when $n \geq 50$. However when $n \leq 35$, no confidence interval is constructed by the profile likelihood (see Section 5.4.3).

When the true turning point occurs in the middle and the original model is segmented, the profile likelihood covers 95% (i.e. within interval (92%- 98%)) of true turning point when $n \geq 50$ but the bootstrap covers 95% when $n \leq 35$. In contrast, when the original model is quadratic or combined, both methods are under 95% coverage because joinpoint is estimated from a curvature pattern and the width of confidence intervals are small. On the other hand, when the true turning point occurs in the beginning or at the end and the original model is segmented, the bootstrap covers 95% of true turning point when $n \leq 35$. However, when the original model is quadratic or combined, both methods are under coverage and tend to be zero when $n \geq 35$, (see Section 5.4.3). This result occurs because the joinpoint is estimated from a curvature trend and the mean of simulated joinpoints is in the middle so with small width of confidence interval around the middle, the true turning point cannot be covered.

In conclusion, when $n \geq 50$ and there is only one joinpoint in the middle of data, profile likelihood is used to construct a confidence interval for the change point from joinpoint model. Otherwise, the bootstrap method can be used. On the other hand, when the true turning point occurs in the beginning or at the end, neither methods are useful to construct confidence intervals for joinpoints when the data is simulated from a curvature pattern. However, when the data is simulated from a straight lines pattern, the bootstrap method can be used.

## 5.4 Discussion and conclusion

Change point analysis was used to detect significant interventions which impact the rates of healthcare associated infections. Two main methods were used

to investigate change points; segmented regression and joinpoint analysis.

### 5.4.1 Segmented regression discussion

Segmented regression was used on different sets of data to detect change points since the period of time over which HAIs (MRSA bacteraemia, MSSA bacteraemia, CDI in patients over 65 years and CDI in patients aged 15-64 years) is monitored by HPS is arbitrary (i.e. HAIs occurred before the year when the data collection started). Segmented regression was used to identify the change points at the time when interventions took place. Using the full dataset (all data points) to fit segmented regression has some advantages. An advantage is that the segmented model can be estimated more accurately because the full dataset was collected over a long period of time which makes the variance of the slope small so the best model has the least variance. Additionally, because the same data was used each time when fitting different models, the model with more than one intervention can be fitted and the likelihood ratio test can be used to test the significance of nested models. The residual deviance was also used to compare different models at different interventions. The limitation of using all data points is that the segmented regression fits linear trends before and after an intervention and the linear trend cannot appropriately describe the data well over a long period of time. If there are subsequent change points after the first one, the pattern after the first change point will not be accurately described by a linear model. There is also length of data bias where the number of observations before and after the intervention at quarter $i$ and intervention at quarter $j$ are different, therefore different information about the data is provided before and after the intervention.

Having all data before the intervention and two years after the intervention gives an indication about the strength (power) of the intervention which im-

pacts the rates within two years only. Residual deviance cannot then be used to compare fitted models because the data are different in each model and the saturated model is different each time. Instead of residual deviance, percentage of deviance explained (PDE) was used to measure the goodness of fit which compares each model with its null model where the model with the largest PDE is the best.

Specifying the number of data before and after each intervention as two years of data (8 points) ensures that there is enough residual degrees of freedom since each model has a linear effect of time and seasonal effect. This has some advantages. Approximately equal variances of the slopes before and after the intervention can be observed because there are the same number of data points before and after the intervention. The smallest variance can estimate the best fitted model. Since segmented regression is fitting linear trends before and after an intervention, the model is more likely to be linear and symmetric over this short period of time. The segmented model can be less accurate with less power and precision because a short period of time was used and this may not give a clear vision of the impact of the interventions. PDE was used to measure the goodness of fit.

### 5.4.2 Joinpoint analysis discussion

Standard joinpoint software [NCISR (2017)] does not account for seasonality where it uses Kim et al. (2000) methods to estimate joinpoints from the joinpoint model and this does not account for other variability. The algorithm was modified in our research to include seasonality and modifies grid search method to consider one point of data between every possible joinpoints then the algorithm is completely rewritten in **R**.

By adding seasonality to CDI in patients over 65 years, joinpoint models do not change the numbers and locations of joinpoints. The result of joinpoint for CDI in patients over 65 years agrees with the result showing in Figure 3.12 where the fitted line of rates of CDI increased up to 2008 then decreased dramatically up to 2011 and subsequently slightly decreased up to September 2014. Adding seasonality to MRSA bacteraemia and CDI in patients aged 15-64 years, the joinpoint models give the same numbers of joinpoints but in different locations indicating that season affects the rates. These change points are associated with some interventions which took place in Scotland (see Section 5.4.5).

Joinpoint analysis on MSSA bacteraemia showed zero change points as the method can not detect any change in the data. Segmented regression showed lack of fit so none of the interventions had a significant effect on the rates of MSSA bacteraemia where the rates of MSSA bacteraemia remain high. These results concur with the result of the polynomial regression fitted model of MSSA (3.8), where the polynomial regression model was linear in the log rates, (see Figure 3.11). Therefore, there is no intervention which has had significant impact on MSSA bacteraemia rates in Scotland.

Confidence intervals of change points can be constructed using the limiting distribution of change points with large sample size [Bai (1997)]. However, profile likelihood and bootstrap methods are non-parametric approaches to construct confidence intervals for change points when the distribution of change point is unknown. This research shows that confidence intervals for one change point from the joinpoint model which was constructed by profile likelihood and bootstrap are approximately similar. If the joinpoint model identified two or more change points, profile likelihood cannot be applied because

it is based on the value of the deviance of the joinpoint model and the estimated joinpoint from that model. If the model has two estimated joinpoints, one value of deviance describes these points and the curve of the deviance function cannot be plotted. Thus, profile likelihood method cannot construct confidence intervals for more than one change point from the joinpoint model so the bootstrap method was used. In contrast, Lerman (1980) approach to construct confidence intervals for the change point from joinpoint model is similar to our approach however he used a function depending on residual sum of squares which constructs confidence intervals for more than one change point while our method uses the residual deviance of the models.

### 5.4.3 Simulation study discussion

Comparing bootstrap and profile likelihood methods to construct confidence intervals using a simulation study shows interesting results. When the sample size is small ($n \leq 20$ in simulated data from quadratic polynomial and combined models and $n \leq 35$ in simulated data from segmented model), the profile likelihood method does not work. Also, when $\beta_0 < 5$, the profile likelihood methods does not work. The profile likelihood method cannot construct confidence intervals for change points from joinpoint models (about 5% of number of simulations) due to the following:

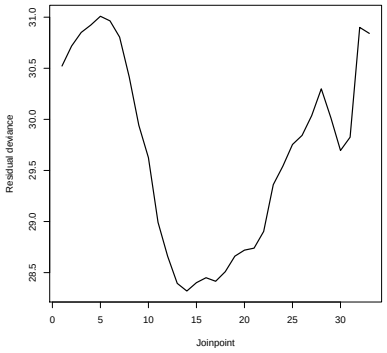
1. The variance of simulated data may increase so the residual deviance of all joinpoint models becomes close to each other. The difference between the minimum residual deviance and the maximum residual deviance becomes less than 3.84. Therefore, there is no intersection between the residual deviance function (curve) and the horizontal line at *hl*=the minimum residual deviance +3.84, (see Figures 5.16(a) and 5.16(b)). As a result, a confidence interval cannot be constructed.

2. The minimum residual deviance may occur at the first point in the residual deviance curve (i.e. first joinpoint model gives the smallest residual deviance) where there are no points before that which can intersect with the horizontal line (at *hl*= the minimum residual deviance +3.84) (see Figure 5.16(c)). Similarly, if the minimum residual deviance occurs at the last point in the residual deviance curve (i.e. last joinpoint model gives the smallest residual deviance). In this case, there are no points after that to intersect with the horizontal line (see Figures 5.16(d)). As a result, lower or upper confidence levels cannot be constructed.

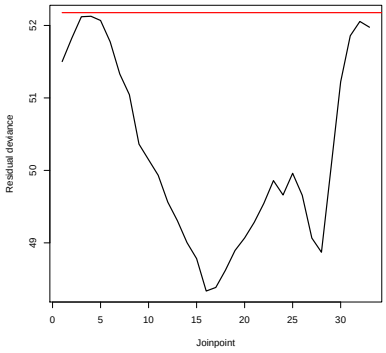
3. When all residual deviances before the minimum residual deviance are less than the horizontal line, there are no points to intersect with the horizontal line before the minimum residual deviance (see Figure 5.16(e)). Also, if all residual deviances after the minimum residual deviance are less than the horizontal line, there are no points to intersect with the horizontal line after the minimum residual deviance (see Figure 5.16(f)). Therefore, there is no lower or upper confidence level and the profile likelihood method cannot construct confidence intervals for change points from the joinpoint model.

The above problems do not influence the bootstrapping as it has a different algorithm from the profile likelihood method in the construction of a confidence interval for the joinpoint (see Section 5.2.2). The profile likelihood method depends on all joinpoints and the deviances while the bootstrap method depends on the bootstrap samples from the best joinpoint model.

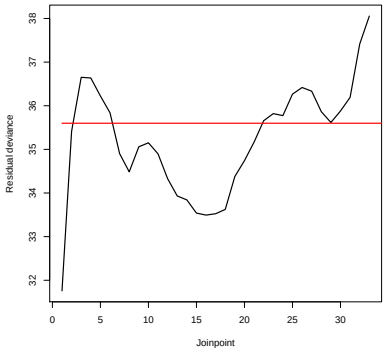
When the sample size $n \leq 35$ and the data is simulated from the segmented model, profile likelihood confidence intervals cannot be constructed. Simulating data from the segmented model and small sample sizes give large variation among simulated data. Therefore, the residual deviance of joinpoint models



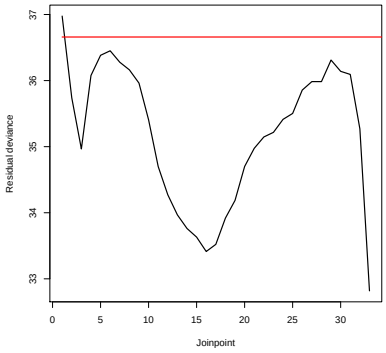

(a) Difference between minimum and maximum deviances is < 3.84.



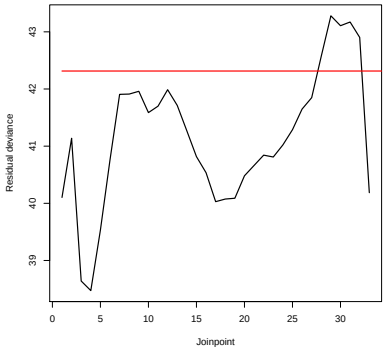

(b) Difference between minimum and maximum deviances is < 3.84



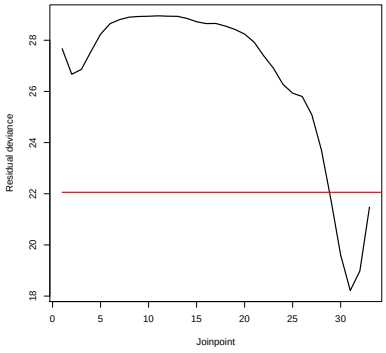

(c) No lower CI because joinpoint occurs at point 2




(d) No upper CI because joinpoint occurs at point 34




(e) No lower CI because there is no deviance > *hl* before joinpoint at point 4




(f) No upper CI because there is no deviance > *hl* after joinpoint at point 31

Figure 5.16: Types of problems that occur with profile likelihood confidence intervals which does not allow for the construction of confidence interval (CI) for joinpoint. Note that these figures are plotted based on the sample size of 35.

will be close to each other. As a result of one or more of the three cases above taking place; upper, lower or both confidence levels cannot be calculated. However, bootstrap confidence intervals can be used in this case and it will give wide intervals which covers 95% of the true turning points when $n \leq 35$. When the variation of the simulated data is large, a weak joinpoint model is fitted. Simulated change points from joinpoint models are variant so confidence intervals of these change points from joinpoint models are wide.

When the data is simulated from a quadratic model or combined model and the true turning point occurs in the middle, the percentage of true turning points within the estimated confidence intervals increases because the mean width increases as the sample size decreases. When the sample size decreases, the variation among simulated data increases and change points from joinpoint models will be located over a wide range. There is not 95% coverage because the data is simulated from a quadratic or combined models (the data are curved) but the straight lines model is fitted to the data (i.e. the wrong model is being fitted to the data). Furthermore, when the true turning point occurs in the beginning or at the end of the dataset, the percentage of true turning points within the estimated confidence intervals is 0. This is because the mean of simulated change points from joinpoint models tends to be roughly in the middle of the data and its confidence interval width is small. Also, the joinpoint model is being fitted to the data which is simulated from a quadratic or combined models.

Profile likelihood is a good method when the sample size $n \geq 50$. When the true turning point occurs in the middle and the data is simulated from a segmented model, the mean width of profile likelihood confidence intervals is less than the mean width of bootstrap confidence intervals. When the data is simu-

lated from quadratic or combined models, the mean width of profile likelihood confidence intervals is about 0.3 larger than the mean width of bootstrap confidence intervals (i.e. approximately similar). The percentage of actual turning points within the confidence interval is larger in profile likelihood confidence intervals than bootstrap confidence intervals but both are under 95% coverage. As a result, the profile likelihood method is a good method to construct confidence intervals for change points from joinpoint models when change occurs in the middle of data and $n \geq 50$. On the other hand, the bootstrap method can be used to construct confidence intervals for change points from joinpoint models for any sample size but needs much more computation than profile likelihood confidence intervals.

In conclusion, when data is simulated from segmented regression, the profile likelihood method does not construct confidence intervals when $n \leq 35$ but bootstrap constructs a confidence interval with 95% coverage of large width. When the data is simulated from a quadratic or combined models, the coverage is very low and tends to be zero when $n \geq 35$ where the true turning point occurs in the beginning or at the end of the data. These zero values are not too precise because the joinpoint model is being fitted to the the data which is simulated from quadratic or combined models. When the sample size is large ($n \geq 50$) and there is only one change point in the middle, profile likelihood is used to construct confidence intervals for the change point. Otherwise, the bootstrap method is not bad but better methods need to be derived.

### 5.4.4 Segmented regression vs joinpoint analysis

From the results of HAIs we can compare change points from joinpoint analysis with change points from segmented regression models which are fitted using full data points. As the segmented regression model detects the change points

at the time when interventions took place, this may not be the real time of change in the trend of rates of HAIs because the practical effect of intervention takes time to impact the infection. However, joinpoint analysis detects change points at any time during the period of study. The results on HAIs showed that the change point detected from joinpoint comes after the change point detected from segmented regression. For example, the change point from the joinpoint model for MRSA bacteraemia comes after the first change point from the segmented regression model and there are four quarters (four data points) between them. Similarly, there are only two data points between the first change point from the joinpoint model and the first change point from the segmented regression model for CDI in patients over 65 years and five data points between the second change points. There are three data points between change points for CDI in patients aged 15-64 years. This indicates that the intervention which is detected from the segmented regression model may have impacted HAIs where this impact is considered as a real change in the rates after a small period of time from that intervention. Therefore, joinpoint analysis showed the time when the trend changed significantly and segmented regression showed which intervention impacted the rate and helped to reduce the rates after a short time (about one year). This interpretation seems plausible because it is rare to have a reduction in the rate at the same time when an intervention took place. Joinpoint analysis detects the time of a significant change while segmented regression detects interventions which may impact the rates.

### 5.4.5 The associated interventions

This research aims to detect the time at which interventions reduce the rate of healthcare associated infections. Some interventions impact the rates of MRSA and MSSA bacteraemias and CDI.

### MRSA and MSSA bacteraemias

Segmented regression analysis showed that the change in rate of MRSA and MSSA bacteraemias occurred in 2006. Some interventions in 2006 may be attributed to this change. The first intervention was to introduce a hand hygiene policy which involved the following [RCN (2005)]:

1. Hands should be cleaned before and after contact with patients.
2. Use soap and water for dirty hands and dry hands thoroughly.
3. Use regular alcohol gel even when hands are clean.

The second intervention was standard infection control precautions initiated and involved the following [RCN (2005)]:

1. Cover all cuts with clean waterproof dressing.
2. Wash hands at regular times throughout the day.
3. Dispose of any waste safely.
4. Do not transfer patients unnecessarily between wards.
5. Arrange time for patient appointments to avoid crowding.
6. Isolate patients who have infections.

The third intervention is that health improvement, efficiency, access and treatment (HEAT) was introduced and it was expected that there would be 30% reduction in staphylococcus aureus bacteraemia (SAB) by 2010 [HPS (2006)].

Some interventions that took place in the beginning of 2008 may have impacted the rate of MRSA bacteraemia. The Scottish patient safety programme [SPSP (2008)] was initiated in January 2008 and aims to develop the safety of

health care in all Scottish hospitals through implementing some bundles related to interventions [Daniel et al. (2015)]. In addition, quality improvement Scotland (QIS) was launched. QIS covers the issues relating to provision of patient-focused care and treatment [HIS (2008)]. Also, in 2008, transmission based precautions (TBPs) commenced and was required to be used by staff. TBPs are control measures that should be implemented in addition to standard infection control precautions for infected patients [HPS (2008a) and ICT (2015)]. Finally, in 2011, MRSA screening practices changed and screening should include all patients at admission, discharge and transfer [RCN (2005)]. The rate of MRSA bacteraemia decreased when MRSA screening changed to clinical risk assessment (CRA) in 2011.

Joinpoint analysis showed a change in the rate of MRSA bacteraemia that happened in 2007 when the Scottish patient safety programme (SPSP) was declared. SPSP improves healthcare safety and reliability in all health care settings [SPSP (2007)].

### CDI

Segmented regression analysis showed that the rate of CDI in patients over 65 years changed two times during the period of study. The first change was in 2007 when the first hand hygiene report was published [SPSP (2007)]. The second change was in 2009 where several interventions took place. First healthcare environment inspectorate (HEI) was carried out across Scottish NHS hospitals and services [HIS (2009)]. Also, the hospital infection incident assessment tool (HIIAT) began in 2009. It provides all information regarding hospital infection [HPS (2009, 2016b)].

Joinpoint analysis showed that a change occurred in 2008 when quality im-

provement Scotland (QIS) was launched [HIS (2008)]. Also, in March 2008, Health Protection Scotland was authorized by the healthcare associated infection Task Force to start developing CDI cross-transmission prevention and control bundles. One of these bundles is the prudent prescribing of antibiotics [SGHAI (2008)]. In addition, transmission based precautions (TBPs) commenced in 2008 [HPS (2008a) and ICT (2015)].

In conclusion, Table 5.4 summarizes the interventions which may have impacted the rates of HAIs in Scotland.

Table 5.4: Summary of the interventions that have been detected by segmented and joinpoint analysis which may have impacted the rate of HAIs in Scotland.

| Point of change | Interventions | MRSA | CDI |
|---|---|---|---|
| Qu2, 2006 | Hand hygiene national campaign announced/launched. | Yes | |
| | Standard Infection Control Precaution model policies first launched. | Yes | |
| | HEAT targets introduced. | Yes | |
| Qu2, 2007 | Scottish Patient Safety Programme (SPSP) announced. | Yes | |
| Qu4, 2007 | First national hand hygiene compliance report issued. | | Yes |
| Qu2, 2008 | HPS care bundles related to interventions first launched (SPSP). | Yes | |
| | Launch of QIS standards. | Yes | Yes |
| | HPS CDI bundle launched. | | Yes |
| | Transmission Based Precaution model policies first launched. | Yes | Yes |
| Qu4, 2009 | First HEI inspection carried out. | | Yes |
| | HIIAT issued for managing outbreaks. | | Yes |
| Qu1, 2011 | MRSA screening changes to CRA. | Yes | |

**MRSA:** Methicillin-resistant staphylococcus aureus, **CDI:** Clostridium difficile infection, **Qu:** Quarter, **HEAT:** Health improvement, efficiency, access and treatment, **SPSP:** Scottish patients safety program, **HPS:** Health Protection Scotland, **QIS:** Quality improvement Scotland, **HEI:** Healthcare environment inspectorate, **HIIAT:** Hospital infection incident assessment tools, **CRA:** Clinical risk assessment.

### 5.4.6 Conclusion

In conclusion, the change points issue is an important analysis to detect the existence of change points and their locations. Segmented regression analysis depends on specific (known) time points (exact time when the interventions took place) and looks for the time when the pattern changes significantly after them. Joinpoint analysis confirmed the importance of seasonality to describe the rates. In addition, joinpoint analysis detects changes at unknown times and looks for change over a period of time and gives the most significant change points among all data. Joinpoint analysis is more accurate than segmented regression analysis because the most significant change point occurs at any time point even if there was no intervention at that time. This suggests that one or more interventions which took place before the time of change may impact the rates where usually the intervention will take time to impact the rate of infections. The aim of this study was to detect the time and the associated intervention that reduces the rate of healthcare associated infections. The research recommends improving the implementation of healthcare interventions such as hand hygiene, giving training courses for hospital staff to deal with infection, screening MRSA in patients prior to hospital admission and applying antibiotic policy to reduce and prevent the occurrence of infection in hospitals and healthcare systems.

# Chapter 6

# Comparing Change Points Methods

In the last two chapters, different methods of detecting change points in the trend of count data were discussed. This chapter investigates change points methods through a simulation study by comparing and finding particular changes in trends that some methods detect more easily and accurately than others. Polynomial generalized linear models (GLM) including quadratic and cubic models estimate the change in trend where the maximum rates start to decrease and minimum rates start to increase (see Chapter 4). However, segmented regression detects the change at particular times and joinpoint analysis estimates the number of changes and their location at all possible data points (see Chapter 5). Although all methods detect similar change points based on different algorithms, the most precise detection method is of interest.

Simulation studies are carried out to compare change points methods. This chapter is divided into three parts according to the number of change points; 0, 1 and 2. The aim of a simulation study when there is no change in trend is to investigate whether polynomial or joinpoint methods detect changes in

the trend. If a method detects a change point, this indicates that the method may be too sensitive. The aim of a simulation study when there are one or two changes in trend is to investigate whether the sample size, the location of turning points, the frequency of the response (number of cases) and the pattern of data affect the detection of change points using different methods. This leads us to investigate whether there is any bias in the estimated change points and to decide which method detects the smallest mean width of confidence intervals around change point, covering 95% of the actual turning points.

Scope of this chapter is as follows: Section 6.1 illustrates the general technique to compare change point methods which includes the assumptions of the original models and simulated data. It also includes the simulation procedure to fit different models using the simulated data and the criteria of choosing the best method. The algorithm for the simulation study is then described. Section 6.2 includes a simulation study based on linear models with no change points. Section 6.3 explains the simulation study on models with one change point in the slope. Section 6.4 includes the results of a simulation study based on two change points. Finally, discussion and conclusion are presented in Section 6.5.

## 6.1 General algorithm to compare change point methods

The simulation study includes four steps; (1) model assumptions (Section 6.1.1), (2) data generation (Section 6.1.2), (3) simulation procedure (Section 6.1.3) and (4) the decision (Section 6.1.4). The process and technical details of the simulation study is described in Section 6.1.5.

### 6.1.1 Model assumptions

Starting with specific values of change points based on the model, different assumptions can be considered:

1. Different numbers of changes can occur (0, 1 or 2).
2. Change points can occur at different places (in the beginning, in the middle or at the end) of the dataset.
3. The model coefficient $\beta_0$ is the number of cases (the intercept of the linear regression in the log of rate). This is assumed to be large where $\beta_0 = 5$, small where $\beta_0 = 3$ or rare where $\beta_0 = 1.5$. These values are chosen to reflect the variety of the number of cases that are observed in HAIs data.
4. The trend of the initial (original) model could be a linear model, curved (polynomial models), straight lines (segmented regression model) with a change in the slope and a combination of polynomial and segmented models. These patterns are chosen to reflect the sorts of changes that can be seen in datasets and also reflect the impact of various interventions.

Different scenarios are considered for the original models with specific (true) turning points where $x_0$ and $x_1$ are the true turning points, these models are as follows:

1. Linear model,

$$\log(\lambda_n) = \beta_0 + \beta_1 x, \tag{6.1}$$

where $\lambda_n$ is the observed data and $\beta_1$ is the coefficient of the slope and can take various values as $\beta_1$ =0, 0.001 and 0.008.

2. Quadratic model where the pattern of the data is curved with one true turning point $x_0$,

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2, \tag{6.2}$$

where $\beta_1 = 0.001$ is the coefficient of the slope (linear term) and it is fixed in all cases of different sample sizes, $\beta_0$s and location of $x_0$. $\beta_2 = -0.003$ is the coefficient of quadratic term and it changes with very small differences (<0.05) (i.e. $\beta_2$s are similar but not identical) when sample size, $\beta_0$s or location of $x_0$ change.

3. Segmented regression model with one true turning point $x_0$,

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)I(x \geq x_0), \tag{6.3}$$

where $\beta_1 = 0.005$ is the coefficient of the slope before change and $\beta_2 = -0.03$ is the coefficient of change in the slope. However, $\beta_1$ and $\beta_2$ change with very small differences (<0.05) (i.e. $\beta_i$s are similar but not identical, $i = 1, 2$) when sample size, $\beta_0$s or location of $x_0$ change. $I(x \geq x_0)$ is an indicator function of $x$ where,

$$I(x \geq x_0) = \begin{cases} 1 & x \geq x_0, \\ 0 & x < x_0. \end{cases} \tag{6.4}$$

4. Combined model (quadratic and segmented with one true turning point ($x_0$)),

$$\log(\lambda_n) = \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \beta_3(x - x_0)I(x \geq x_0), \tag{6.5}$$

where $\beta_1 = 0.001$ is the coefficient of the slope (linear term) and it is fixed in all cases of different sample sizes, $\beta_0$s and location of $x_0$. $\beta_2 = -0.003$ is the coefficient of quadratic term and $\beta_3 = -0.05$ is the coefficient of change in the slope. However, $\beta_2$ and $\beta_3$ change with very small differences (<0.05) (i.e. $\beta_i$s are similar but not identical, $i = 2, 3$) when sample size, $\beta_0$s or location of $x_0$ change.

5. Cubic model where the pattern of the data is curved with two true turning points; $x_0$ and $x_1$,

$$\log(\lambda_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, \tag{6.6}$$

where $\beta_1$ is the coefficient of linear term and $\beta_2$ is the coefficient of quadratic term and their values depend on the sample size and the location of two true turning points, (see Section 4.5.2). $\beta_3 = 0.000033$ is the coefficient of the cubic term and it is fixed in all cases.

6. Segmented regression model (two true turning points in the slope; $x_0$ and $x_1$),

$$\log(\lambda_n) = \beta_0 + \beta_1 x + \beta_2 (x - x_0) I(x \geq x_0) + \beta_3 (x - x_1) I(x \geq x_1), \tag{6.7}$$

where $\beta_1 = 0.005$ is the coefficient of the slope before the change and is fixed in all cases. $\beta_2 = -0.005$ is the coefficient of change in the slope after the first true turning point ($x_0$) but it changes with very small differences (<0.05) (i.e. $\beta_2$s are similar but not identical) when the location of $x_0$ and $x_1$ change. $\beta_3 = 0.01$ is the coefficient of the slope after the second true turning point ($x_1$) and is fixed in all cases. $I(x \geq x_i), \quad i = 0, 1$ as defined in Equation (6.4).

7. Combined model (cubic and segmented with two true turning points; $x_0$ and $x_1$),

$$\log(\lambda_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - x_0) I(x \geq x_0) + \beta_5 (x - x_1) I(x \geq x_1), \tag{6.8}$$

where $\beta_1$ is the coefficient of the linear term and $\beta_2$ is the coefficient of the quadratic term and their values depend on the sample size and the location of $x_0$ and $x_1$. $\beta_3 = 0.000033$ is the coefficient of the cubic term and

it is fixed in all cases. $\beta_4 = -0.005$ is the coefficient of change in the slope after first true turning point ($x_0$) but it changes with very small differences (<0.05) (i.e. $\beta_4$s are similar but not identical) when the location of $x_0$ and $x_1$ change. $\beta_5 = 0.01$ is the coefficient of change in the slope after the second true turning point ($x_1$) and it is fixed in all cases. $I(x \geq x_i), \quad i = 0, 1$ as defined in Equation (6.4).

### 6.1.2 Data generation

The assumptions of the simulated data are:

1. Different sample sizes (number of data points); small (20), moderate (35) or large (50). These values are chosen to reflect the variety of the number of data points that are observed in HAIs data.

2. Data generated according to a Poisson distribution (since the data is count/rate per unit)

### 6.1.3 Simulation procedure

Data from the original model (see Section 6.1.1) were used to generate data according to Poisson distribution,

$$Y_n \sim \text{Poisson}(\lambda_n), \tag{6.9}$$

where $Y_n$ is simulated Poisson random variables, $\lambda_n$ is observed data from the original model and $n$ is the sample size. Simulated data $Y_n$ were then used to fit different models (polynomial, segmented and joinpoint models) and change points from these were calculated. Results are then compared.

### Procedure for one turning point

The following fitted models are compared in order to detect one change point from the original model with one true turning point.

1. The quadratic GLM model,

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{6.10}$$

   Estimate the change point and find the standard error for the estimator, find the confidence intervals for the estimated change point using bootstrapping as in Chapter 4. Calculate the significance of $\beta_2$ (the coefficient of quadratic term which identifies the change point) and the residual deviance of the model.

2. The cubic GLM model,

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon. \tag{6.11}$$

   Doing the same procedure as in the quadratic model then, extract the results. In this case, $\beta_3$ is the coefficient of the cubic term which identifies two change points.

3. The segmented GLM model with one change point ($s_0$) is fitted to investigate whether the true turning point can be detected.

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_{2(s_0)}(x - s_0)I(x \geq s_0) + \varepsilon, \tag{6.12}$$

   where $s_0$ is the true turning point ($x_0$) from the original model and $I(x \geq s_0)$ as defined in Equation (6.4). Also, $s_0$ is chosen to be the true turning point$\pm 3$ ($x_0 \pm 3$) because the segmented model is fitted when the true

turning point is misspecified. This explains how the segmented regression works when the true turning point is not specified properly. In such cases, the residual deviances and the significance of the parameters of changes in the slope $\beta_{2(s_0)}$ evaluate the segmented regression model.

4. The joinpoint GLM model,

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_{2(j_0)}(x - j_0)I(x \geq j_0) + \varepsilon, \tag{6.13}$$

where $j_0$ is all possible data points except the first and the last points and $I(x \geq j_0)$ as defined in Equation (6.4). Find the change point which is from the model with the smallest residual deviance, construct the confidence intervals for the joinpoints using the bootstrap method and test the significance of the coefficient of change $\beta_{2(j_0)}$.

Moreover, in the case of no change occurring in the original model, all fitted models above are compared except the segmented model because there is no specific turning point.

### Procedure for two turning points

In order to detect two true turning points, the following fitted models are compared.

1. The quadratic GLM model (6.10).
2. The cubic GLM model (6.11).
3. The segmented GLM model with two change points; $s_0$ and $s_1$,

$$\begin{aligned}\log(Y_n) = \beta_0 + \beta_1 x + \beta_{2(s_0)}(x - s_0)I(x \geq s_0) \\ + \beta_{2(s_1)}(x - s_1)I(x \geq s_1) + \varepsilon,\end{aligned} \tag{6.14}$$

where $s_0$ and $s_1$ are the true turning points ($x_0$ and $x_1$) from the original model or true turning points±3 ($x_0 \pm 3$ and $x_1 \pm 3$) and $I(x \geq s_i)$ as defined

in Equation (6.4). Calculate the residual deviances and the significance of the parameters of change in the slope $\beta_{2(s_i)}$, where $i = 0, 1$.

4. The joinpoint GLM model,

$$\log(Y_n) = \beta_0 + \beta_1 x + \beta_{2(j_0)}(x - j_0)I(x \geq j_0) + \beta_{2(j_1)}(x - j_1)I(x \geq j_1) + \varepsilon, \quad (6.15)$$

where $j_0$ is all possible data points except the first point and the last two points and $j_1$ is all possible data points after $j_0$ except the last point. $I(x \geq j_0)$ as defined in Equation (6.4). Find the change points which are the joinpoints from the model with the smallest residual deviance, construct the confidence intervals for the joinpoints using the bootstrap method and calculate the significance of the coefficients of change $\beta_{2(j_i)}$, $\quad i = 0, 1$.

### 6.1.4 The decision

To investigate and decide which method is the best to detect change points, the following criteria will be examined:

1. Residual deviance of the fitted model. Calculate the residual deviance for polynomial, segmented and joinpoint simulated models then, choose the best model which has the smallest residual deviance.

2. The estimate of the change point and its standard error are applied in polynomial and joinpoints models. The method giving the closest change points to the true turning points is chosen as the best method.

3. Confidence intervals for change points. In polynomial and joinpoint models, we are looking at whether confidence intervals cover 95% of the true turning points. Based on 200 simulations with significance level of 5%, the range (92%- 98%) is consistent with 95% coverage (i.e. $(0.95 \pm z_{1-(\alpha/2)}\sqrt{0.95 \times 0.05/200}) \times 100 \approx (92\% - 98\%)$, $\quad z_{1-(\alpha/2)} = 1.96$). The width of the confidence intervals is also calculated.

4. The significance of the parameter of change in the model. In the quadratic model (6.10), $\beta_2$ is the parameter to determine one change point. In the cubic model (6.11), $\beta_3$ is the parameter to identify two change points. In the segmented models (6.12) and (6.14), $\beta_{2(s_i)}$ are the parameters of change in the slope. Finally, in the joinpoint models (6.13) and (6.15), $\beta_{2(j_i)}$ are the parameters of the change in the slope, $i = 0, 1$. 200 simulations justify the large difference in the percentage of significance of the parameter of change between different sample sizes or different $\beta_0$s. A standard error ($SD$) of sample proportion (i.e. percentage ($\hat{p}$)) is calculated using normal approximation as $SD = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ where $N$ is the number of simulation. Thus, the standard error for the difference of two independent proportions is given as $SD(\hat{p_1} - \hat{p_2}) = \sqrt{SD_1^2 + SD_2^2}$. For example, if $\hat{p_1} = 25\%, \quad \hat{p_2} = 50\%$, the $SD(\hat{p_1} - \hat{p_2}) = 9.4\%$ which indicates the difference of less than 9.4% is not important so the difference of 25% is significant.

5. The percentage of the coefficient of change parameter estimated in the polynomial and joinpoint models that are not equal to zero.

6. The percentage of the estimated change points from the polynomial and joinpoint models that occur within the range of dataset.

### 6.1.5 The algorithm of simulation study

1. Set up the sample size; $n$=50, 35 or 20.

2. Set up the number of cases where $\beta_0$ = log( number of case ) = 5, 3 or 1.5.

3. Identify the number and location of change points then based on these, identify the original model and its coefficients $\beta_i$, $i$ = 1, 2, 3, 4 and 5.

4. Calculate $\log(\lambda_n)$ from the original models (Section 6.1.1) then calculate $\lambda_n$.

5. Use Poisson distribution to generate the data $Y_n$ from the original model, (see Equation (6.9)).

6. Use $Y_n$ to fit different models; quadratic, cubic, segmented and joinpoint.

7. Calculate change points and their confidence intervals (CI based on 500 bootstrap samples) from simulated model in step 6; use the method in Chapter 4 for change points from polynomial models and the method in Chapter 5 for change points from joinpoint models.

8. Save the results of the estimated change points, their confidence intervals and all criteria in Section 6.1.4.

9. Do the simulation 200 times in each situation and save the results. The outputs are the residual deviance of the estimated model, the significant of the parameter of change within the estimated model, the percentage of parameter of change values are equal to zero, the estimated change points and the percentage of their confidence intervals containing the true turning points. Finally the percentage of estimated change points occurring within the range of dataset.

10. According to the results and criteria, the methods of detecting change points are compared.

## 6.2 Models with no change points

These models show one segment without any change and the pattern of data is linear. Three different original models are considered; first model with no change points and the slope of the trend is zero ($\beta_1 = 0$), (see the first row in

Figure 6.1). The second model with the slope of the trend as $\beta_1 = 0.001$, (see the second row in Figure 6.1) and the final model with the slope of the trend as $\beta_1 = 0.008$, (see the third row in Figure 6.1). The simulation under each model shows the results of different sample sizes (50, 35 and 20).

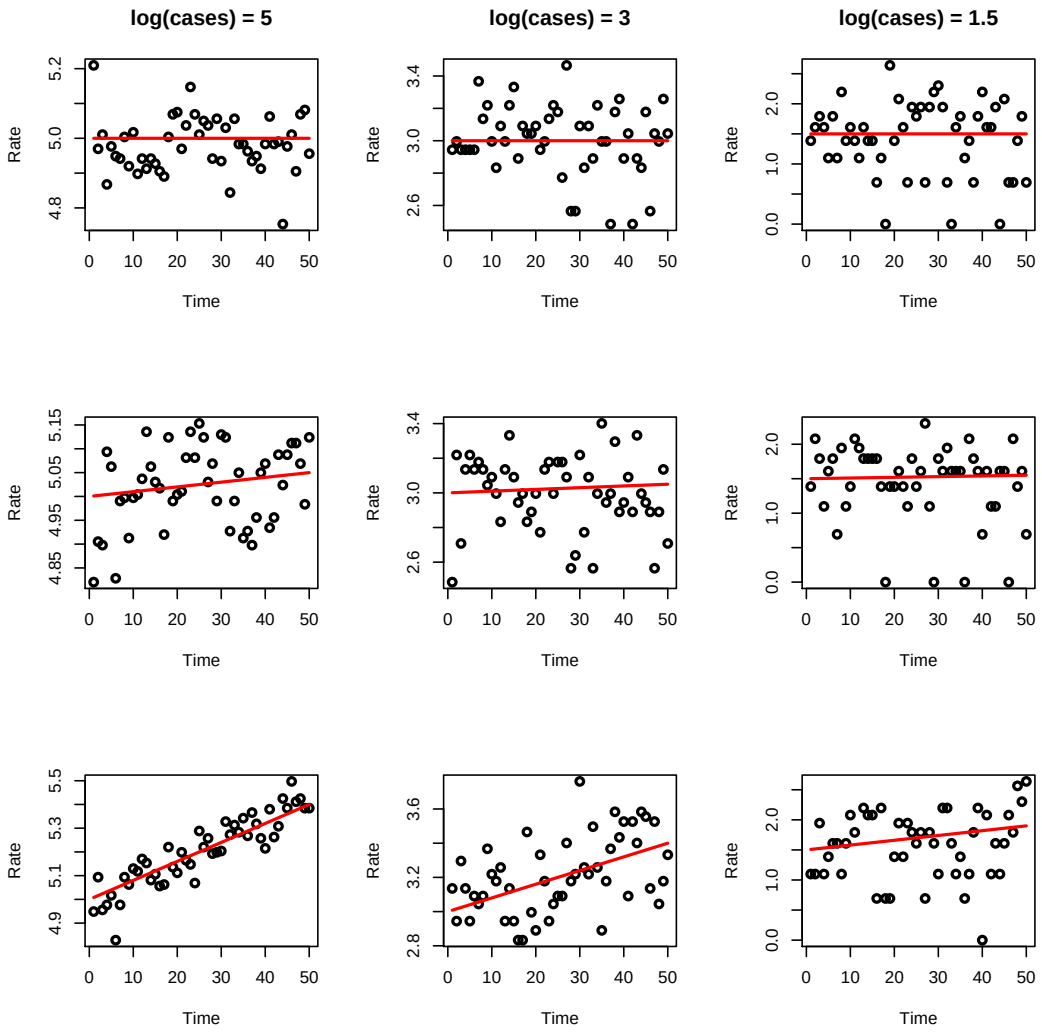



Figure 6.1: Original model (red line) and simulated data (black circles) in different cases of $\beta_0 = \log(\text{cases})$ and $\beta_1$. The slope $\beta_1 = 0$ in the first row, $\beta_1 = 0.001$ in the second row and $\beta_1 = 0.008$ in the third row.

This part of the simulation study is designed to investigate the proportions of false positivity, (i.e. is there any method more likely to report a significant change point when no change in slope exists?). Section 6.2.1 includes the simulation results when the original model has no change point and no slope.

Section 6.2.2 observes the results of no change point with increasing slope. Summary and conclusion are discussed in Section 6.2.3.

### 6.2.1 Original model with no change and no slope

Setting $\beta_1 = 0$ in model (6.1) shows that the trend of the model is linear as in the first row in Figure 6.1. Table 6.1 shows the results of different sample sizes; $n$ =50, 35 and 20. The terms observed in the Table 6.1 are explained as:

1. $\beta_0$ is the constant related to the number of cases where 1.5 is rare, 3 is small and 5 is large.

2. $\beta_1$ is the coefficient of the slope in linear models (6.1).

3. S.M is the type of change point fitted models to the simulated data (for simplicity, simulated models) where Q.TP is a quadratic model, C.TP is a cubic model (C.TP1 and C.TP2 are related to the first and the second estimated change points from a cubic model) and JP is a joinpoint model.

4. DV is the mean of residual deviances of simulated models.

5. SG.CH% is the percentage of significant parameters of change over 200 simulations (i.e. number of times when the coefficient of change is significant).

6. TP.ES is the mean of estimated change points from polynomial (quadratic and cubic) and joinpoint models.

7. TP.SD is the mean of standard errors of estimated change points from polynomial and joinpoint models.

8. CI.WD is a mean width of confidence intervals. This is calculated as a mean of the difference between upper and lower levels from confid-

ence intervals of estimated change points from polynomial and joinpoint models.

9. NO.TP% is a percentage of times that the coefficients of change in the polynomial and joinpoint models are zero (i.e. coefficient of change=0).

10. TP.IN% is a percentage of times where estimated change points are inside the range of data for polynomial and joinpoint models.

### 6.2.2 Original model with no change and increasing slope

Setting $\beta_1 = 0.001$ in model (6.1) shows a linear trend with slightly increasing slope as shown in the second row in Figure 6.1. Also, putting $\beta_1 = 0.008$ in model (6.1) shows a linear trend with a large increase in the slope, (see the third row in Figure 6.1). The results of both cases are obtained in Tables D.1 and D.2.

### 6.2.3 Summary and conclusion

Tables 6.1, D.1 and D.2 showed the results of the simulation when there is no turning points and the slope $\beta_1 = 0$, 0.001 and 0.008, respectively. The polynomial and joinpoint models are fitted to the simulated data. However, segmented regression models do not fit the simulated data because there is no true turning point to specify the change.

The residual deviances (DV) for polynomial and joinpoint models are similar. The percentage of significant parameters of change (SG.CH%) in polynomial models is often 5% of the time. However, the SG.CH% in joinpoint models are about 25% which is larger than polynomial models. SG.CH% in joinpoint models often decreases when $\beta_0$ decreases but this difference is not usually important as it is less than 10% and according to 200 simulations this is not considered an important difference. The results of SG.CH% of polynomial

Table 6.1: Simulation study on linear model with no slope ($\beta_1 = 0$).

| Number of data points $n = 50$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 46.1 | 4.5 | 31.1 | 535.1 | 169.0 | 2.0 | 83.0 |
| | C.TP1 | 45.0 | 4.5 | 8.0 | 158.0 | 122.5 | 16.5 | 85.0 |
| | C.TP2 | | | >50 | 278.5 | 132.9 | | 86.5 |
| | JP | 44.1 | 27.5 | 26.4 | 17.8 | 46.3 | 0.0 | 100.0 |
| 3 | Q.TP | 47.0 | 9.5 | 10.7 | >1000 | 167.4 | 0.5 | 85.0 |
| | C.TP1 | 46.0 | 5.0 | 8.1 | 119.9 | 110.5 | 4.0 | 84.5 |
| | C.TP2 | | | 40.3 | 125.9 | 113.9 | | 89.5 |
| | JP | 44.9 | 24.0 | 24.7 | 18.3 | 46.2 | 0.0 | 100.0 |
| 1.5 | Q.TP | 50.7 | 3.0 | <1 | 500.0 | 166.5 | 0.5 | 84.5 |
| | C.TP1 | 49.7 | 5.0 | 6.1 | >1000 | 105.2 | 4.0 | 90.5 |
| | C.TP2 | | | 40.1 | 283.6 | 114.8 | | 89.5 |
| | JP | 48.2 | 23.5 | 25.6 | 19.2 | 44.9 | 0.0 | 100.0 |
| Number of data points $n = 35$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 31.5 | 7.5 | 22.2 | 586.4 | 109.5 | 0.0 | 84.0 |
| | C.TP1 | 30.4 | 4.0 | 7.5 | 133.6 | 82.0 | 2.0 | 87.0 |
| | C.TP2 | | | 34.6 | 125.1 | 75.1 | | 91.0 |
| | JP | 29.4 | 25.5 | 16.8 | 12.5 | 30.9 | 0.0 | 100.0 |
| 3 | Q.TP | 32.8 | 6.5 | 15.9 | 316.5 | 113.6 | 0.0 | 86.0 |
| | C.TP1 | 31.7 | 7.5 | 9.5 | 92.9 | 81.4 | 0.0 | 89.5 |
| | C.TP2 | | | 31.3 | 478.7 | 74.1 | | 94.5 |
| | JP | 30.8 | 23.5 | 18.6 | 12.5 | 31.2 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.9 | 6.5 | 20.7 | 316.9 | 121.3 | 0.0 | 81.5 |
| | C.TP1 | 32.9 | 6.5 | 3.6 | 203.6 | 97.3 | 0.5 | 84.5 |
| | C.TP2 | | | 28.3 | 663.1 | 84.4 | | 95.0 |
| | JP | 31.9 | 19.5 | 17.8 | 12.1 | 31.1 | 0.0 | 100.0 |
| Number of data points $n = 20$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 17.1 | 6.5 | 9.7 | 486.9 | 64.5 | 0.0 | 86.0 |
| | C.TP1 | 15.9 | 7.5 | <1 | 50.4 | 41.7 | 0.0 | 81.0 |
| | C.TP2 | | | >20 | 88.6 | 39.1 | | 98.0 |
| | JP | 15.3 | 27.0 | 10.0 | 6.4 | 16.1 | 0.0 | 100.0 |
| 3 | Q.TP | 17.1 | 3.5 | 8.7 | 161.2 | 66.8 | 0.0 | 83.0 |
| | C.TP1 | 16.0 | 5.0 | <1 | 46.9 | 43.5 | 0.0 | 86.5 |
| | C.TP2 | | | >20 | 43.6 | 48.5 | | 98.0 |
| | JP | 15.4 | 22.5 | 10.1 | 6.4 | 16.2 | 0.0 | 100.0 |
| 1.5 | Q.TP | 17.5 | 6.5 | 10.6 | 319.0 | 63.8 | 0.0 | 80.5 |
| | C.TP1 | 16.6 | 3.0 | 5.2 | 41.7 | 43.7 | 0.0 | 89.0 |
| | C.TP2 | | | 15.3 | 39.7 | 45.7 | | 100.0 |
| | JP | 15.9 | 17.5 | 10.6 | 6.3 | 16.1 | 0.0 | 100.0 |

$\beta_1$**:** The coefficient of the slope in linear model, $\beta_0$**:** The log of number of cases, **S.M:** The simulated model, **Q.TP:** Quadratic model from simulated data, **C.TP:** Cubic model from simulated data, **JP:** Joinpoint model from simulated data, **DV:** The mean residual deviance of simulated model, **SG.CH%:** The percentage of significant parameters of change, **TP.ES:** The mean of estimated change points from the simulations, **TP.SD:** The mean of standard errors of estimated change points, **CI.WD:** The mean width of confidence intervals, **NO.TP%:** The percentage of times that the coefficients of change is zero, **TP.IN%:** The percentage of times where estimated change points are inside the range of data.

models are reasonable because there are no change points. However, SG.CH% of the joinpoint model shows that generally 25% of the time the change coeffi-

cient is significant but this is a poor conclusion since there is no change point yet the joinpoint model detects them roughly 25% of the time.

The estimated change points (TP.ES) from joinpoint models are roughly similar and approximately in the middle of the dataset. There should not be estimated change points because the original model is linear but the joinpoint model finds change points and picks the values randomly and therefore their average occurs roughly in the middle of the dataset. However, TP.ES of polynomial models can occur anywhere, even outwith the range of data. The standard errors of the estimated change points (TP.SD) from polynomial models are very large. However, TP.SD from joinpoint models are large and approximately the same whenever $\beta_0$ changes. This is about 1/3 of the sample size (about 18 when sample size is 50, 12 when sample size is 35 and 6 when sample size is 20). The mean width of the confidence intervals of estimated change points (CI.WD) from joinpoint models are large but less than the range of data however, the CI.WD from polynomial models are very wide and greater than the sample size so any detected change points are very imprecise. The CI.WD from joinpoint models are approximately similar whenever $\beta_0$ changes. This explains the centering of the mean of estimated change point and indicates that the joinpoint method is useless in detecting the change point accurately. The percentage of estimated change points within the range of data(TP.IN%) from the polynomial models decreases when $\beta_0$ increases and it is about 80%. However, estimated change points from joinpoint models occurs 100% within the range of data because it is estimated to be one of the data points.

In conclusion, the simulation studies are carried out for both a constant and an increasing slope where the same results were obtained for both. In the polynomial method, the value of TP.ES is not precise because roughly 5% of

the polynomial models have significant coefficients of change points (i.e. about $0.05\times200 = 10$ simulations) and the rest of values of TP.ES are estimated on non-significant coefficient of change points (i.e. approximately 190 change points). Therefore, in polynomial models, the values of TP.ES, TP.SD and CI.WD are not precise and therefore the polynomial method is in fact a good method since it did not detect change points when there were no change points in the data. However, the SG.CH% of joinpoint model is roughly 25% which indicates false positives. This shows that the estimated change point occurs roughly in the middle of the data but the mean width of confidence interval of the estimated change point is large. Therefore, this indicates that the joinpoint method is poor for the detection of change points.

## 6.3 Models with one change point

A simulation study was carried out with one true turning point to investigate change point methods, (see the algorithm in Appendix D.1). We considered three different sample sizes; 50, 35 and 20 and set up three different original models; the quadratic model (6.2), segmented model with one true turning point (6.3) and the combination between them (model (6.5)) with the true turning point in the middle, in the beginning or at the end. The simulated models based on the original model was explained in Section 6.1.3. Section 6.3.1 contains the results when one true turning point occurs in the middle of data. Section 6.3.2 includes the results when one true turning point is present in the beginning or at the end of data. The conclusion is in Section 6.3.3.

### 6.3.1 Change occurring in the middle of the dataset

We assume the true turning point occurs in the middle of the dataset for three different sample sizes; 50, 35 and 20 and three different numbers of cases where

$\beta_0$ = 5, 3 and 1.5. In each case, the simulation study produces results using different methods; polynomial models, segmented regression and joinpoint analysis to detect a change point when the data is simulated from quadratic, segmented and combined models (see Figure 6.2).

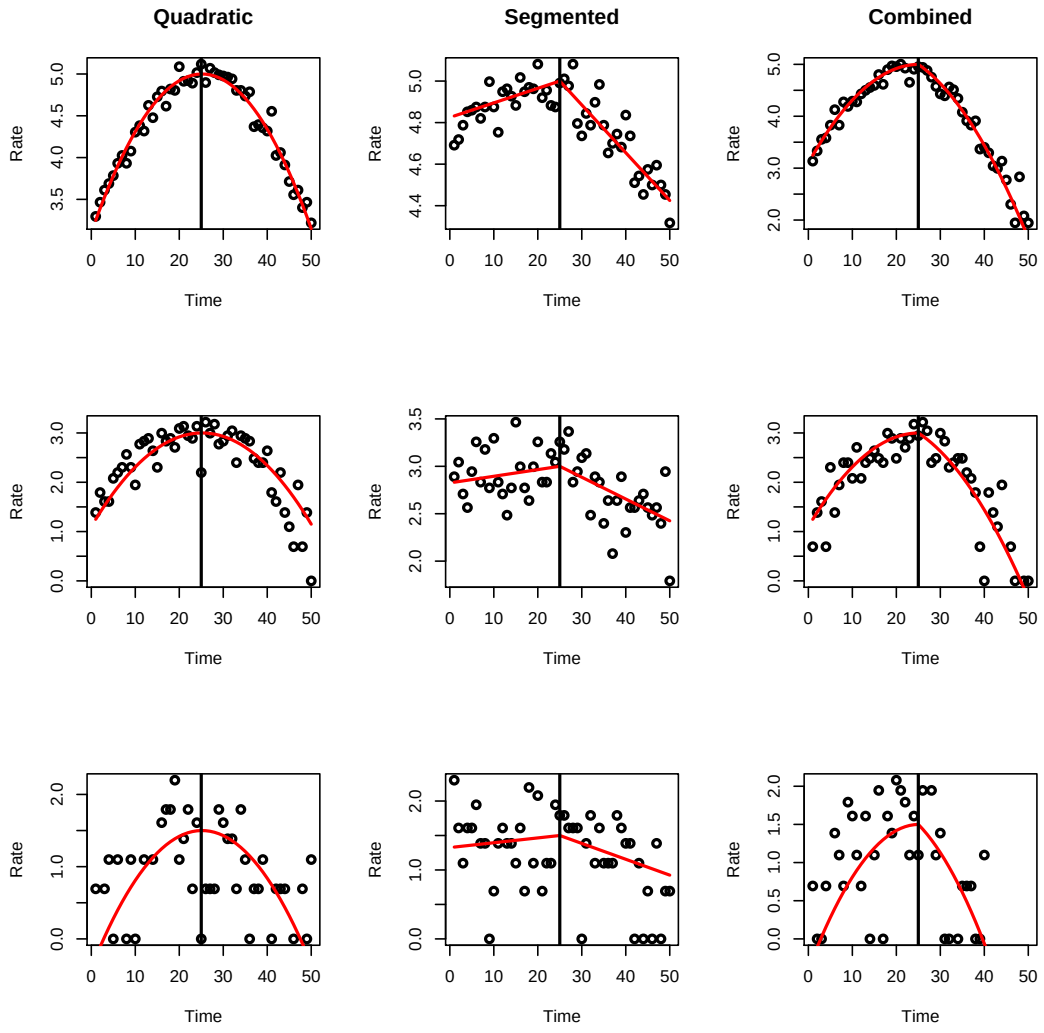



Figure 6.2: Original model (red line) and simulated data (black circles) from quadratic, segmented and combined models with one true turning point in the middle (black vertical line). Plots in the first row associated with $\beta_0$ = 5, the second row with $\beta_0$ = 3 and the third row with $\beta_0$ = 1.5.

Consider the number of data points as 50, and the true turning point occurs at point 25. Table 6.2 shows the results of simulation studies for different original

models and $\beta_0$s. The terms observed in the Table 6.2 are as follows:

1. O.M is the original model where the data comes from.

2. S.M is the type of change point fitted models (simulated models) where SR is a segmented regression model with specific change at true turning point and SR±3 are segmented regression models with specific points of change at true turning point±3. The other simulated models were explained in Section 6.2.

3. CI.TP% is a percentage of confidence intervals around estimated change points containing the true turning point. It is calculated for polynomial and joinpoint models.

Table 6.2: Number of data points=50 and the true turning point in the middle at 25.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 47.8 | 100.0 | 25.1 | 0.2 | 91.0 | 0.8 | 0.0 | 100.0 |
| ratic | C.TP1 | 46.8 | 6.0 | <1 | >1000 | 83.0 | >1000 | 12.0 | 58.0 |
| | C.TP2 | | | >50 | >1000 | 33.0 | >1000 | | 42.0 |
| | SR | 121.1 | 100.0 | | | | | | |
| | SR-3 | 132.7 | 100.0 | | | | | | |
| | SR+3 | 134.4 | 100.0 | | | | | | |
| | JP | 120.4 | 100.0 | 25.4 | 1.2 | 78.0 | 2.1 | 0.0 | 100.0 |
| Segm- | Q.TP | 49.7 | 100.0 | 13.0 | 4.7 | 0.0 | 11.4 | 0.0 | 100.0 |
| ented | C.TP1 | 48.7 | 3.0 | <1 | 510.4 | 0.0 | 520.3 | 16.0 | 63.0 |
| with | C.TP2 | | | >50 | >1000 | 93.0 | 548.0 | | 4.0 |
| one | SR | 46.5 | 100.0 | | | | | | |
| change | SR-3 | 46.5 | 100.0 | | | | | | |
| point | SR+3 | 46.9 | 100.0 | | | | | | |
| | JP | 46.3 | 100.0 | 25.3 | 3.4 | 96.0 | 13.9 | 0.0 | 100.0 |
| Comb- | Q.TP | 55.9 | 100.0 | 21.9 | 0.2 | 0.0 | 0.8 | 0.0 | 100.0 |
| ined | C.TP1 | 54.3 | 14.0 | <1 | >1000 | 0.0 | >1000 | 12.0 | 27.0 |
| | C.TP2 | | | >50 | >1000 | 85.0 | >1000 | | 73.0 |
| | SR | 99.3 | 100.0 | | | | | | |
| | SR-3 | 107.3 | 100.0 | | | | | | |
| | SR+3 | 143.8 | 100.0 | | | | | | |
| | JP | 97.5 | 100.0 | 24.5 | 0.7 | 91.0 | 2.0 | 0.0 | 100.0 |
| $\beta_0 = 3$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP-Q | 48.6 | 100.0 | 25.2 | 0.6 | 90.0 | 2.3 | 0.0 | 100.0 |
| ratic | C.TP1 | 47.6 | 7.0 | <1 | >1000 | 68.0 | 961.8 | 2.5 | 51.0 |
| | C.TP2 | | | >50 | >1000 | 39.0 | 959.1 | | 49.0 |

*Continued on next page*

Table 6.2 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| | SR | 58.6 | 100.0 | | | | | | |
| | SR-3 | 60.2 | 100.0 | | | | | | |
| | SR+3 | 60.3 | 100.0 | | | | | | |
| | JP | 56.3 | 100.0 | 25.5 | 3.6 | 61.0 | 5.2 | 0.0 | 100.0 |
| Segm- | Q.TP | 48.1 | 79.5 | 16.7 | 176.9 | 31.5 | 56.1 | 0.0 | 98.0 |
| ented | C.TP1 | 47.0 | 5.0 | <1 | 438.5 | 5.0 | 244.6 | 4.5 | 73.0 |
| with | C.TP2 | | | >50 | >1000 | 94.0 | 277.8 | | 58.0 |
| one | SR | 46.1 | 83.5 | | | | | | |
| change | SR-3 | 46.2 | 79.5 | | | | | | |
| point | SR+3 | 46.2 | 72.0 | | | | | | |
| | JP | 45.8 | 96.0 | 25.8 | 8.1 | 95.5 | 32.9 | 0.0 | 100.0 |
| Comb- | Q.TP | 51.8 | 100.0 | 21.8 | 0.5 | 0.0 | 2.1 | 0.0 | 100.0 |
| ined | C.TP1 | 50.7 | 6.5 | <1 | >1000 | 0.0 | >1000 | 2.5 | 41.0 |
| | C.TP2 | | | >50 | 841.8 | 95.0 | >1000 | | 59.0 |
| | SR | 57.5 | 100.0 | | | | | | |
| | SR-3 | 58.7 | 100.0 | | | | | | |
| | SR+3 | 63.4 | 100.0 | | | | | | |
| | JP | 56.3 | 100.0 | 24.5 | 2.1 | 78.5 | 4.1 | 0.0 | 100.0 |
| | | | | $\beta_0 = 1.5$ | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 52.6 | 100.0 | 25.2 | 1.2 | 91.0 | 4.6 | 0.0 | 100.0 |
| ratic | C.TP1 | 51.5 | 5.5 | <1 | 701.6 | 70.0 | 510.3 | 1.5 | 52.0 |
| | C.TP2 | | | >50 | 434.4 | 53.0 | 545.6 | | 51.0 |
| | SR | 54.7 | 100.0 | | | | | | |
| | SR-3 | 55.0 | 100.0 | | | | | | |
| | SR+3 | 55.1 | 100.0 | | | | | | |
| | JP | 52.6 | 99.5 | 25.6 | 5.4 | 73.5 | 10.7 | 0.0 | 100.0 |
| Segm- | Q.TP1 | 51.0 | 37.5 | 20.1 | >1000 | 80.5 | 152.6 | 0.0 | 87.0 |
| ented | C.TP1 | 50.1 | 2.5 | <1 | 749.7 | 29.0 | 180.7 | 1.5 | 81.0 |
| with | C.TP2 | | | >50 | 262.5 | 94.0 | 179.1 | | 69.0 |
| one | SR | 49.9 | 40.0 | | | | | | |
| change | SR-3 | 49.9 | 37.5 | | | | | | |
| point | SR+3 | 50.0 | 29.5 | | | | | | |
| | JP | 48.8 | 60.0 | 26.3 | 15.4 | 92.5 | 40.7 | 0.0 | 100.0 |
| Comb- | Q.TP2 | 50.2 | 100.0 | 21.8 | 1.2 | 17.5 | 4.6 | 0.0 | 100.0 |
| ined | C.TP1 | 49.0 | 6.0 | <1 | >1000 | 0.0 | 608.3 | 2.5 | 43.0 |
| | C.TP2 | | | >50 | 810.1 | 97.0 | 583.8 | | 60.0 |
| | SR | 50.7 | 100.0 | | | | | | |
| | SR-3 | 51.1 | 100.0 | | | | | | |
| | SR+3 | 52.0 | 100.0 | | | | | | |
| | JP | 49.4 | 98.5 | 24.4 | 4.7 | 75.0 | 9.8 | 0.0 | 100.0 |

**O.M:** Original model, **SR:** Segmented regression model from simulated data, **CI.TP%:** The percentage of confidence intervals that contains the true turning point. See Table 6.1 for the definition of $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

Tables D.3 and D.4 show the results of simulation studies when sample sizes are 35 and 20, respectively.

**Summary**

Tables 6.2, D.3 and D.4 show the results of the simulation studies carried out to compare change point methods and explain the best method to detect the change in trend of count data. This simulation is based on the assumption of where the data came from; data has a quadratic pattern or segmented pattern with one true turning point which occurs in the middle of the data. Also, the combination between quadratic and segmented models with the same true turning point was considered.

The residual deviance (DV) and the percentage of significance of change point parameter (SG.CH%) are important measurements to detect the best method of detecting the change point. When the data has a quadratic or combined (combination model of quadratic and segmented regression) pattern with the true turning point in the middle, the DV of polynomial models (quadratic and cubic) are less than DV of joinpoint and segmented models when the sample size is large ($n \geq 35$) and $\beta_0 \geq 5$. However, when the sample size decreases ($n \leq 20$) or $\beta_0 \leq 3$, the DV of all models is similar. On the other hand, when the data has segmented regression pattern with the true turning point in the middle, the DV of joinpoint, segmented regression and polynomial models are similar.

The SG.CH% in cubic models are about 5% because the data are simulated from models which assume that one change point has occurred. When the original models are quadratic or combined models, often 100% of SG.CH% have been detected for quadratic, segmented regression and joinpoint models when $n \geq 50$ and when $n \leq 35$ and $\beta_0 \geq 3$ which indicates that the change point is always found. However, whenever the sample size and $\beta_0$ decrease, the SG.CH% in quadratic and joinpoint models decrease. SG.CH% in the join-

point model becomes greater than the SG.CH% in the quadratic model when $n \leq 20$ and $\beta_0 \leq 1.5$. On the other hand, when the original model is a segmented regression model, usually 100% of the SG.CH% is detected for all models when $n \geq 35$ and $\beta_0 \geq 5$. The 100% SG.CH% indicates that a change point is found in each simulation however, it may be biased (i.e. it may not be in the right place). For example, when $n = 50$, $\beta_0 \geq 5$ and the true turning point occurs at 25, the quadratic model fitted to data which is simulated from the segmented regression showed 100% SG.CH% for the coefficient of change but the average of estimated change points (TP.ES) occurs at the wrong place; at 13 (i.e. not close to the middle at 25). However, the SG.CH% in quadratic, segmented regression and joinpoint models decreases when the sample size and $\beta_0$ decrease. In general, the SG.CH% in joinpoint models are greater than the SG.CH% in quadratic models and a large difference occurs when $n \leq 35$ and $\beta_0 \leq 3$ because joinpoint model fits well to the data from straight lines (segmented regression pattern). In addition, SG.CH% in the joinpoint model is greater than that in segmented regression models indicating that the joinpoint detects change points better than segmented regression but it has a wide confidence interval. In all original models, NO.TP% measures the number of times when the coefficient of the change point in different models is exactly equal to zero which explains how many times the change points cannot be estimated. This very rarely occurs where quadratic and joinpoint models report 0% of NO.TP% and the cubic models report that at most 15% of the models are fitted with the coefficient of change being exactly equal to zero.

The main important results are associated with the estimated change point. The estimated change points (TP.ES) from the cubic model is usually located outside the range of data therefore they have very wide confidence intervals. When the original model is quadratic and the sample size is 50, the quadratic

and joinpoint methods roughly locate estimated change points (TP.ES) at the position of the true turning point with small standard error (TP.SD). However, TP.SD increases as $\beta_0$ decreases. The mean width of confidence interval (CI.WD) of estimated change points from the quadratic method is less than the mean width of the confidence interval of the estimated change points from the joinpoint method. On the other hand, when the original models are segmented or combined, the joinpoint method locates TP.ES close to the true turning point with small TP.SD but wide CI.WD. However, TP.SD increases as $\beta_0$ decreases. In contrast, the quadratic method did not locate TP.ES at the true turning point position (at the middle) and CI.WD of estimated change point increases as $\beta_0$ decreases. CI.TP% measures the 95% coverage of true turning points where the 95% coverage with 200 simulations is equivalent to (92%- 98%). When the original model is quadratic, the joinpoint method did not cover 95% of true turning points however the quadratic method often covers 95% of the true turning points when $n \leq 35$ and $\beta_0 \leq 5$ with large mean width of confidence interval. When the original model is combined, all methods; polynomial and joinpoint report CI.TP% under 95% coverage. However, when the original model is segmented regression, the joinpoint method usually covers 95% of the true turning points but with wide confidence intervals. CI.TP% for estimated change points from polynomial methods are under 95% coverage. TP.IN% measures the number of times that estimated turning points occur within the range of data. TP.IN% from the joinpoint method is always within the range of data because the joinpoint method estimates the change point to be only within the range of the data. However, the TP.IN% from the quadratic method occurs almost 100% within the range of data except when the original model is segmented and $n \leq 35$ and $\beta_0 \leq 3$. The TP.IN% from the cubic method is often less than 50% because the cubic model fits the data that assumes one change point. A similar pattern of results is obtained when $n = 35$ and 20.

### 6.3.2 Change occurring in the beginning or at the end of dataset

We assume that the true turning point occurs in the beginning or at the end of data and three numbers of data points; 50, 35 and 20 with three different values of $\beta_0$; 5, 3 and 1.5. Figures 6.3 and D.1 explain the different scenarios and Tables 6.3, D.5, D.6, D.7, D.8 and D.9 show the results of simulation studies at different sample sizes; 50, 35 and 20 and different locations of the true turning point; in the beginning or at the end of data.

Table 6.3: Number of data points=50 and the true turning point in the beginning at 13.

| | | | | $\beta_0 = 5$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 46.7 | 100.0 | 13.2 | 0.4 | 85.0 | 1.4 | 0.0 | 100.0 |
| ratic | C.TP1 | 45.7 | 6.0 | <1 | >1000 | 90.0 | >1000 | 11.0 | 39.0 |
| | C.TP2 | | | >50 | >1000 | 56.0 | >1000 | | 61.0 |
| | SR | 161.5 | 100.0 | | | | | | |
| | SR-3 | 235.2 | 100.0 | | | | | | |
| | SR+3 | 118.4 | 100.0 | | | | | | |
| | JP | 96.0 | 100.0 | 21.1 | 1.4 | 0.0 | 2.3 | 0.0 | 100.0 |
| Segm- | Q.TP | 51.6 | 94.0 | <1 | 354.5 | 6.0 | 161.2 | 0.0 | 13.0 |
| ented | C.TP1 | 47.1 | 46.0 | 5.1 | 51.8 | 11.0 | 67.3 | 4.0 | 91.0 |
| with | C.TP2 | | | >50 | 190.2 | 35.0 | 148.1 | | 17.0 |
| one | SR | 46.8 | 89.0 | | | | | | |
| change | SR-3 | 47.0 | 64.0 | | | | | | |
| point | SR+3 | 47.0 | 96.0 | | | | | | |
| | JP | 46.0 | 98.0 | 14.5 | 7.9 | 92.0 | 21.4 | 0.0 | 100.0 |
| Comb- | Q.TP | 49.9 | 100.0 | 10.9 | 0.4 | 0.0 | 1.6 | 0.0 | 100.0 |
| ined | C.TP1 | 46.3 | 7.0 | <1 | >1000 | 0.0 | 803.2 | 2.0 | 97.0 |
| | C.TP2 | | | >50 | >1000 | 48.0 | >1000 | | 3.0 |
| | SR | 110.6 | 100.0 | | | | | | |
| | SR-3 | 161.7 | 100.0 | | | | | | |
| | SR+3 | 92.3 | 100.0 | | | | | | |
| | JP | 90.4 | 100.0 | 17.5 | 1.2 | 0.0 | 2.1 | 0.0 | 100.0 |
| | | | | $\beta_0 = 3$ | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 47.7 | 100.0 | 13.2 | 1.5 | 91.0 | 5.8 | 0.0 | 100.0 |
| ratic | C.TP1 | 46.5 | 7.0 | <1 | 778.0 | 96.0 | 705.4 | 5.0 | 50.0 |
| | C.TP2 | | | >50 | 815.7 | 71.0 | 843.3 | | 53.0 |
| | SR | 58.0 | 100.0 | | | | | | |
| | SR-3 | 64.2 | 92.5 | | | | | | |
| | SR+3 | 54.0 | 100.0 | | | | | | |
| | JP | 49.0 | 100.0 | 22.8 | 4.8 | 10.5 | 8.3 | 0.0 | 100.0 |
| Segm- | Q.TP | 49.5 | 43.0 | <1 | >1000 | 93.5 | 419.4 | 0.0 | 46.5 |
| ented | C.TP1 | 47.4 | 21.5 | 5.9 | 213.1 | 89.0 | 133.2 | 1.5 | 89.0 |

Continued on next page

Table 6.3 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| with | C.TP2 | | | >50 | 346.2 | 70.0 | 193.1 | | 49.0 |
| one | SR | 47.3 | 39.0 | | | | | | |
| change | SR-3 | 47.3 | 20.0 | | | | | | |
| point | SR+3 | 47.4 | 50.5 | | | | | | |
| | JP | 46.3 | 71.0 | 17.5 | 13.2 | 96.0 | 39.6 | 0.0 | 100.0 |
| Comb- | Q.TP | 47.4 | 100.0 | 11.6 | 1.1 | 71.5 | 4.4 | 0.0 | 100.0 |
| ined | C.TP1 | 46.3 | 6.5 | <1 | >1000 | 44.0 | 685.9 | 3.0 | 65.0 |
| | C.TP2 | | | >50 | 922.6 | 89.0 | 769.8 | | 36.0 |
| | SR | 57.4 | 100.0 | | | | | | |
| | SR-3 | 65.4 | 99.5 | | | | | | |
| | SR+3 | 53.7 | 100.0 | | | | | | |
| | JP | 50.9 | 100.0 | 19.2 | 3.5 | 20.5 | 5.8 | 0.0 | 100.0 |
| | | | | $\beta_0 = 1.5$ | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 49.2 | 99.0 | 12.7 | 99.8 | 90.5 | 21.2 | 0.0 | 98.5 |
| ratic | C.TP1 | 48.3 | 3.5 | <1 | >1000 | 99.0 | 432.1 | 3.5 | 61.0 |
| | C.TP2 | | | >50 | 754.4 | 79.0 | 427.8 | | 54.0 |
| | SR | 51.7 | 72.5 | | | | | | |
| | SR-3 | 53.6 | 37.0 | | | | | | |
| | SR+3 | 50.5 | 92.5 | | | | | | |
| | JP | 47.7 | 95.5 | 22.6 | 7.9 | 53.5 | 19.7 | 0.0 | 100.0 |
| Segm- | Q.TP | 53.6 | 20.0 | 18.5 | >1000 | 99.5 | 432.0 | 0.0 | 48.5 |
| ented | C.TP1 | 52.3 | 9.0 | <1 | 119.5 | 100.0 | 116.6 | 2.5 | 86.0 |
| with | C.TP2 | | | 47.5 | 258.0 | 65.0 | 168.5 | | 66.0 |
| one | SR | 52.2 | 18.0 | | | | | | |
| change | SR-3 | 52.3 | 8.0 | | | | | | |
| point | SR+3 | 52.1 | 22.5 | | | | | | |
| | JP | 51.1 | 35.0 | 21.8 | 16.4 | 94.5 | 43.4 | 0.0 | 100.0 |
| Comb- | Q.TP | 46.4 | 99.5 | 12.1 | 9.2 | 92.0 | 11.8 | 0.0 | 99.0 |
| ined | C.TP1 | 45.2 | 5.5 | <1 | 444.7 | 89.0 | 360.9 | 0.0 | 69.0 |
| | C.TP2 | | | >50 | > 1000 | 80.0 | 362.9 | | 47.0 |
| | SR | 48.7 | 88.0 | | | | | | |
| | SR-3 | 50.5 | 57.0 | | | | | | |
| | SR+3 | 47.6 | 98.5 | | | | | | |
| | JP | 45.0 | 93.0 | 20.6 | 7.5 | 59.5 | 15.9 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR** and **CI.TP%**. See Table 6.1 for the definition of $\beta_0$, $\beta_1$, **S.M, Q.TP, C.TP,JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

## Summary

When data is simulated from a quadratic or combined model (combination model of quadratic and segmented regression) with true turning point near the beginning, the residual deviance (DV) of polynomial models is less than the DV of segmented regression and joinpoint models when the sample size is large ($n \geq 35$) and $\beta_0 \geq 5$. However, when the sample size decreases ($n \leq 20$) and when $n \leq 50$ and $\beta_0 \leq 3$, the DV of all models are approximately similar.

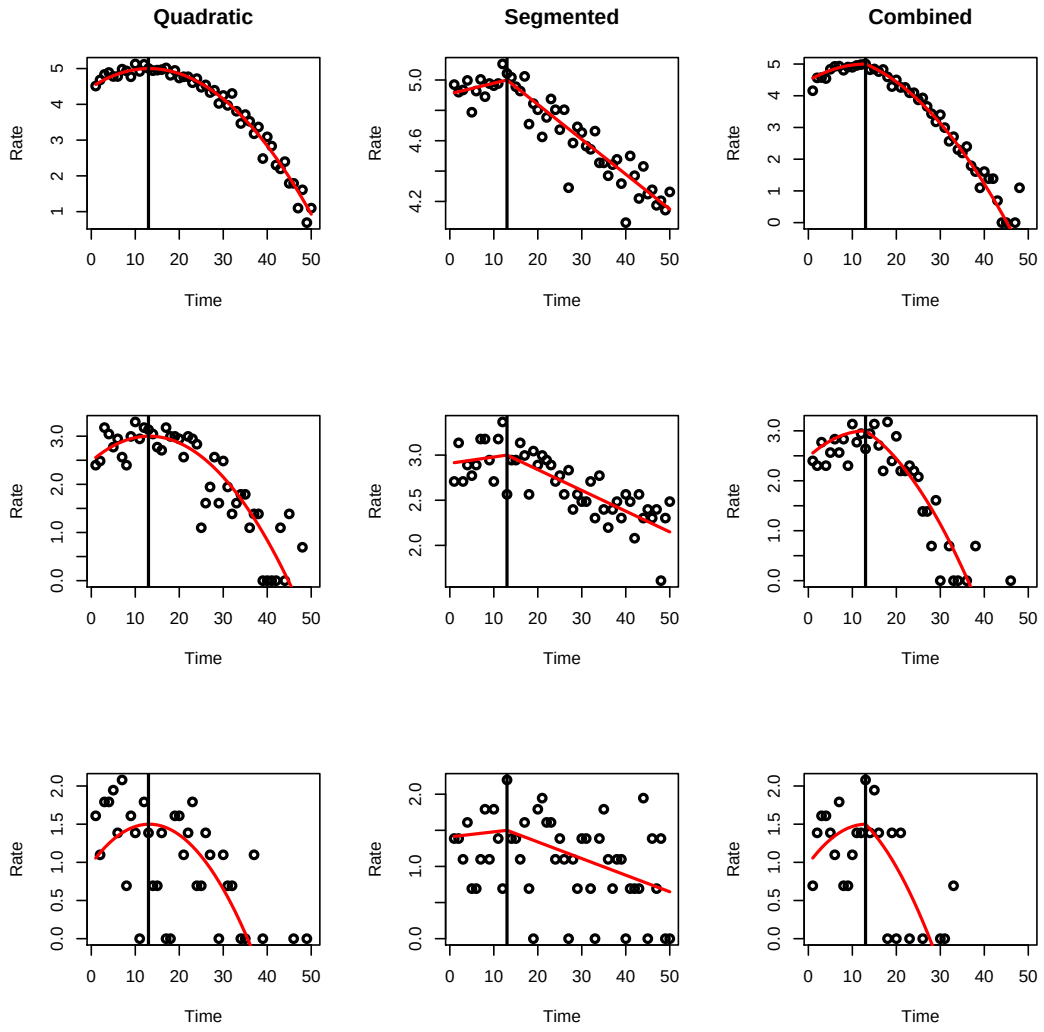



Figure 6.3: Original model (red line) and simulated data (black circles) from quadratic, segmented and combined models with one true turning point in the beginning (black vertical line). Plots in the first row associated with $\beta_0 = 5$, the second row with $\beta_0 = 3$ and the third row with $\beta_0 = 1.5$.

In contrast, when the original model is a segmented regression, the DV of all models has a similar pattern to the result of when true turning point occurs in the middle of the data.

When the original models are quadratic or combined, the result of SG.CH% in the cubic model is often the same when the true turning point occurs in the middle of the data. When the sample size is large ($n = 50$), roughly all

models have 100% of SG.CH% but the SG.CH% of the segmented regression model decreases as $\beta_0$ decreases. When $n \leq 35$ and $\beta_0 \leq 1.5$, the SG.CH% for all models decreases. On the other hand, when the data is simulated from segmented regression, the SG.CH% in the cubic model is less than 50% and decreases to 7% as the sample size and $\beta_0$ decrease. The SG.CH% in all other models is greater than 80% when $n = 50$ and $\beta_0 \geq 5$ otherwise, they are often less than 50%. In general, SG.CH% from the joinpoint model is greater than SG.CH% from the quadratic model particularly when $\beta_0$ decreases but a wide confidence interval is estimated from the joinpoint method. Similar results of NO.TP% when the true turning point occurs in the middle is reported in the case of the true turning point occurring at the beginning of the data.

The TP.ES from the cubic model has a similar result when the true turning point occurs in the middle. It is located far from the true turning point with a very large standard error and mean width of confidence interval. When the original model is quadratic, the quadratic method often locates the TP.ES very close to the true turning point and the standard error and the mean width of the confidence interval increases as $\beta_0$ decreases. However, the TP.ES from the joinpoint method tends to be in the middle of the data and as a result it is not located at the true turning point where it has a wide confidence interval as $\beta_0$ decreases. The TP.SD from the quadratic and joinpoint models increase as the sample size and $\beta_0$ decrease. Moreover, the location of the change point from the simulated model SR+3 (i.e. the simulated segmented regression with change point occurring at true turning point+3) is roughly better than the location of estimated change point from the joinpoint method. When the original model is segmented, the TP.ES from the quadratic method did not locate the true turning point and TP.ES occurs mostly outside the range of data. The TP.ES from the joinpoint method also did not locate the true turning point and TP.ES

tends to be in the middle of the data as $\beta_0$ and the sample size decrease. The TP.SD from the quadratic model is very large and the TP.SD from the joinpoint model increases when $\beta_0$ decreases. When the original model is combined, the TP.ES from the joinpoint and quadratic methods often did not locate the true turning point. However, for a small sample size ($n \leq 20$) the TP.ES from the quadratic method often locates the true turning point but has a very wide confidence interval. The TP.SD is the same when simulated from a quadratic model.

The CI.TP% has no clear pattern and the percentage coverage of the true turning point is often outwith the range (92%- 98%). This may be because of sampling error or because the confidence interval was calculated through an inaccurate method (percentile bootstrap method). The only exception is when the original model is quadratic. Here the CI.TP% of the estimated change point from the quadratic method is within 95% coverage when $n \leq 35$ and $\beta_0 \leq 3$. A wide confidence interval is observed in quadratic and joinpoint methods and the CI.WD increases as the sample size and $\beta_0$ decrease. Also, when the original model is combined, the quadratic method often covers 95% of the true turning points when $\beta_0 \leq 1.5$ with a very wide confidence interval. The percentage of estimated change point occurring within the range of data (TP.IN%) has a similar result as when the true turning point is present in the middle of the data. However, when the original model is segmented regression, the quadratic method reports that less than 50% of estimated change points are in the range of data. A similar pattern is obtained when $n$ =35 and 20.

When the true turning point occurs near the end of the data, the same conclusion to the result of the true turning point being near to the beginning is obtained. The only difference when the true turning point is at the end is when

the original model is segmented regression, the quadratic method reports that more than 85% of the estimated change points are in the range of data (TP.IN%). A similar pattern of results is obtained when sample sizes are 35 and 20.

### 6.3.3 Conclusion

When one change point occurs in the middle of the data and describes the slope of trend, all methods; quadratic polynomial, segmented regression and joinpoint analysis often detect change point when the pattern of the data seems quadratic polynomial. However, when $n \geq 50$ and $\beta_0 \geq 5$, the quadratic polynomial method is better to detect change points. When $n \leq 20$ and $\beta_0 \leq 1.5$, nothing can detect the change in the middle. In contrast, when the pattern of the data seems as straight lines or mixed between the polynomial and segmented, segmented and joinpoint methods are better to detect change points roughly in the middle when $n \geq 35$ and $\beta_0 \geq 3$. Otherwise, no method detects change in the middle.

On the other hand, when one change point occurs in the beginning or at the end of the data and the original model is quadratic, the quadratic polynomial method can detect change points when $n \geq 20$ and $\beta_0 \geq 3$. Otherwise, different methods need to be investigated to detect change points.

## 6.4 Models with two change points

As in the one change point simulation study, three different sample sizes; 50, 35 and 20 and three numbers of cases, where $\beta_0 = 5, 3$ and 1.5, are considered. Three different original models; the cubic model, segmented model with two true turning points in the slope and the combination between them with two true turning points occurring roughly in the middle of the dataset, (see Section 6.4.1)

or one in the beginning and one at the end, (see Section 6.4.2) are investigated. The conclusion is in Section 6.4.3. See the algorithm in Appendix D.2.

### 6.4.1 Two change points occurring approximately in the middle of data

This section includes the results of simulation studies when the true turning points occur roughly in the middle of the dataset. When the number of data points is 50, the true turning points are assumed to be at 16 and 33, (see Figure 6.4). When the sample sizes are $n = 35$ and 20, the true turning points occur at (11 and 24) and (7 and 13), respectively.

Tables 6.4, D.10 and D.11 show the results of simulation studies when $n = 50, 35$ and 20, respectively. The terms in these tables are explained as follow:

1. S.M is the type of change point fitted models (simulated models) where Q.TP and C.TP are explained in Section 6.2. SR is a segmented regression model (SR1 and SR2 are related to the first and the second true turning points) and SR ± 3 are segmented regression models (SR1 ± 3 and SR2 ± 3 are related to the first and the second true turning points ± 3). JP is a joinpoint model (JP1 and JP2 are related to the first and the second estimated change points from joinpoint model).

2. LCL and UCL are the lower and upper limits, respectively of the confidence interval of the estimated change point.

3. CI% is the percentage of confidence intervals of the estimated change points containing the true turning point, (CI1% and CI2% are related to the first and the second estimated change points, respectively).

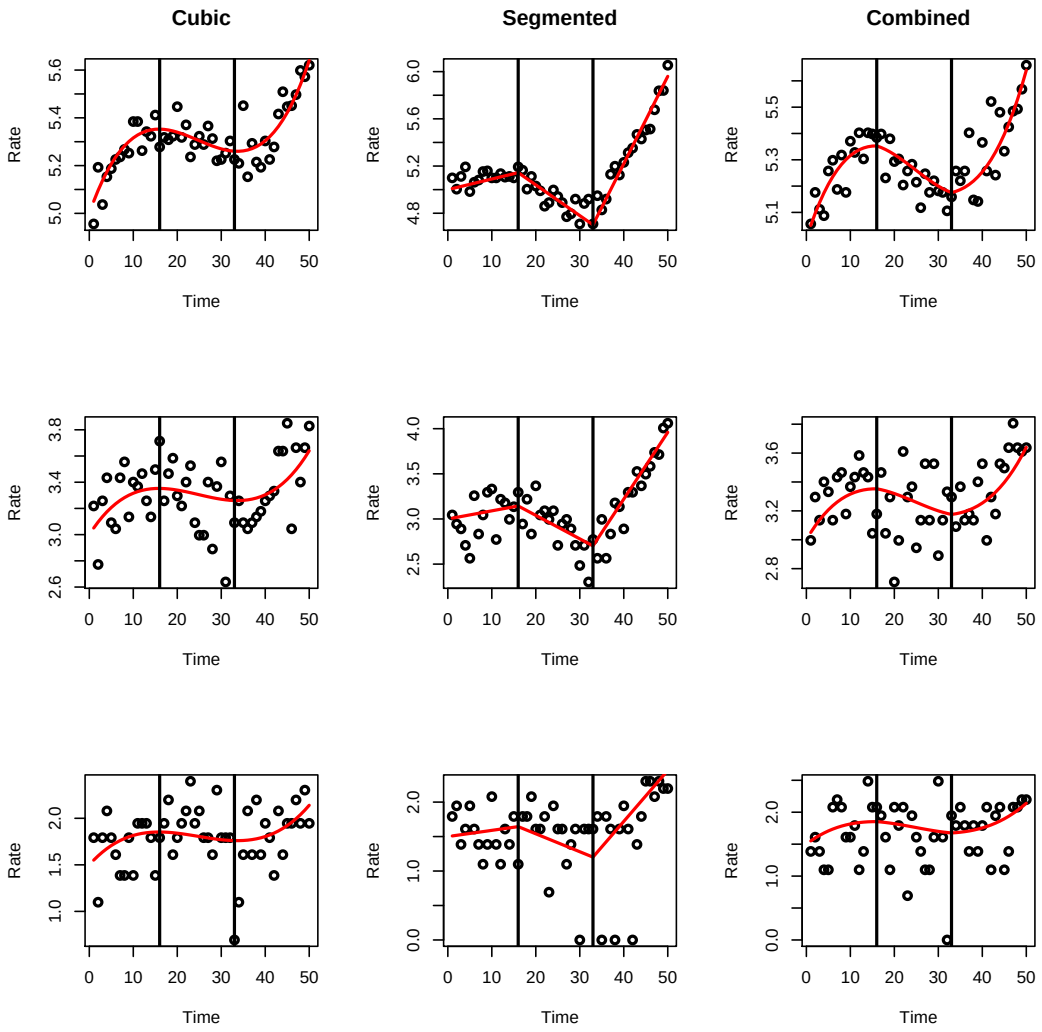



Figure 6.4: Original model (red line) and simulated data (black circles) from cubic, segmented and combined models with two true turning points in the middle (black vertical lines). Plots in the first row associated with $\beta_0 = 5$, the second row with $\beta_0 = 3$ and the third row with $\beta_0 = 1.5$.

Table 6.4: Number of data points=50 and the true turning points at 16 and 33.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 108.4 | 58.0 | <1 | >1000 | -181.3 | 195.2 | 376.5 | 99.5 | 73.5 | 0.0 | 60.5 |
| | C.TP1 | 46.3 | 100.0 | 15.8 | 0.8 | 14.3 | 17.5 | 3.2 | 93.5 | | 0.0 | 100.0 |
| | C.TP2 | | | 33.5 | 0.8 | 31.6 | 34.8 | 3.2 | | 87.0 | | 100.0 |
| | SR1 | 52.4 | 100.0 | | | | | | | | | |
| | SR2 | | 100.0 | | | | | | | | | |
| | SR1 -3 | 55.1 | 99.0 | | | | | | | | | |
| | SR2 -3 | | 100.0 | | | | | | | | | |
| | SR1 +3 | 52.4 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 100.0 | | | | | | | | | |
| | JP1 | 45.0 | 100.0 | 11.7 | 3.7 | 7.6 | 18.7 | 11.1 | 62.0 | | 0.0 | 100.0 |
| | JP2 | | 100.0 | 38.8 | 3.2 | 34.8 | 41.7 | 6.9 | | 33.5 | 0.0 | 100.0 |
| 3 | Q.TP | 56.6 | 13.5 | 35.9 | 785.0 | -133.7 | 172.4 | 306.1 | 98.5 | 87.5 | 0.5 | 62.5 |
| | C.TP1 | 46.9 | 84.0 | 15.8 | 35.2 | 3.5 | 23.1 | 19.6 | 93.5 | | 0.0 | 100.0 |
| | C.TP2 | | | 33.2 | 19.6 | 25.4 | 43.2 | 17.8 | | 92.5 | | 100.0 |
| | SR1 | 45.8 | 42.5 | | | | | | | | | |
| | SR2 | | 68.0 | | | | | | | | | |
| | SR1 -3 | 46.4 | 38.0 | | | | | | | | | |
| | SR2 -3 | | 58.5 | | | | | | | | | |
| | SR1 +3 | 45.7 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 61.0 | | | | | | | | | |
| | JP1 | 42.6 | 88.0 | 19.0 | 12.5 | 4.3 | 38.1 | 33.8 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 95.5 | 37.0 | 8.9 | 18.7 | 46.7 | 28.0 | | 92.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 50.7 | 4.0 | 11.3 | 603.1 | -81.6 | 130.7 | 212.2 | 97.5 | 96.5 | 0.0 | 78.0 |
| | C.TP1 | 47.6 | 35.0 | 15.9 | 237 | -43.1 | 26.7 | 69.8 | 98.5 | | 0.5 | 98.0 |
| | C.TP2 | | | 42.0 | 346.0 | 22.9 | 89.6 | 66.7 | | 92.5 | | 95.5 |
| | SR1 | 45.8 | 13.0 | | | | | | | | | |
| | SR2 | | 16.5 | | | | | | | | | |
| | SR1 -3 | 45.9 | 14.5 | | | | | | | | | |
| | SR2 -3 | | 16.0 | | | | | | | | | |
| | SR1 +3 | 46.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 17.0 | | | | | | | | | |
| | JP1 | 42.8 | 72.5 | 21.6 | 15.0 | 3.0 | 41.4 | 38.4 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 75.5 | 30.9 | 14.8 | 9.8 | 47.7 | 38.0 | | 96.0 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 50.5 | 13.5 | 17.3 | 776.1 | -118.1 | 163.8 | 281.9 | 99.0 | 88.0 | 0.5 | 68.5 |
| | C.TP1 | 46.7 | 43.5 | 16.0 | 137.0 | -25.5 | 25.9 | 51.4 | 94.5 | | 2.0 | 100.0 |
| | C.TP2 | | | 46.0 | 94.7 | 20.9 | 72.9 | 52.0 | | 97.0 | | 97.5 |
| | SR1 | 43.8 | 25.5 | | | | | | | | | |
| | SR2 | | 45.5 | | | | | | | | | |
| | SR1 -3 | 44.2 | 14.5 | | | | | | | | | |
| | SR2 -3 | | 46.0 | | | | | | | | | |
| | SR1 +3 | 44.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 26.5 | | | | | | | | | |
| | JP1 | 41.6 | 80.0 | 22.6 | 12.2 | 3.5 | 40.5 | 37.1 | 98.5 | | 0.0 | 100.0 |
| | JP2 | | 84.5 | 31.2 | 11.2 | 12.6 | 47.1 | 34.5 | | 97.5 | 0.0 | 100.0 |
| 3 | Q.TP | 48.7 | 4.0 | >50 | >1000 | -65.0 | 114.0 | 179.0 | 98.0 | 95.0 | 1.5 | 84.5 |
| | C.TP1 | 47.2 | 11.0 | 15.0 | 429.0 | -62.9 | 28.3 | 91.2 | 99.0 | | 3.5 | 98.0 |
| | C.TP2 | | | >50 | 285.8 | 21.7 | 119.8 | 98.1 | | 97.0 | | 88.5 |

*Continued on next page*

Table 6.4 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR1 | 44.9 | 10.0 | | | | | | | | | |
| | SR2 | | 10.5 | | | | | | | | | |
| | SR1 -3 | 45.1 | 7.5 | | | | | | | | | |
| | SR2 -3 | | 10.5 | | | | | | | | | |
| | SR1 +3 | 45.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 8.5 | | | | | | | | | |
| | JP1 | 42.2 | 69.0 | 20.7 | 14.9 | 2.9 | 43.4 | 40.5 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 75.0 | 28.0 | 15.1 | 8.2 | 48.2 | 40.0 | | 96.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 48.8 | 6.0 | 15.2 | 856.6 | -63.3 | 111.0 | 174.3 | 97.0 | 98.5 | 0.0 | 84.0 |
| | C.TP1 | 49.0 | 9.0 | 15.7 | 617 | -65.6 | 28.0 | 93.6 | 99.0 | | 0.5 | 98.0 |
| | C.TP2 | | | 41.1 | 231.0 | 22.7 | 122.0 | 99.3 | | 95.5 | | 87.0 |
| | SR1 | 45.8 | 3.0 | | | | | | | | | |
| | SR2 | | 8.5 | | | | | | | | | |
| | SR1 -3 | 45.7 | 3.0 | | | | | | | | | |
| | SR2 -3 | | 5.5 | | | | | | | | | |
| | SR1 +3 | 45.8 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 9.0 | | | | | | | | | |
| | JP1 | 42.9 | 57.5 | 22.5 | 15.0 | 2.7 | 43.6 | 40.9 | 98.0 | | 0.0 | 100.0 |
| | JP2 | | 57.5 | 28.1 | 15.4 | 7.7 | 48.6 | 40.9 | | 98.5 | 0.0 | 100.0 |

Original model is combined model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 139.1 | 68.0 | 18.4 | 342.4 | -37.8 | 74.3 | 112.1 | 99.5 | 78.0 | 0.5 | 96.0 |
| | C.TP1 | 47.1 | 100.0 | 13.9 | 0.6 | 12.7 | 15.0 | 2.2 | 55.5 | | 0.0 | 100.0 |
| | C.TP2 | | | 35.3 | 0.5 | 34.4 | 36.2 | 1.9 | | 27.5 | | 100.0 |
| | SR1 | 50.7 | 100.0 | | | | | | | | | |
| | SR2 | | 100.0 | | | | | | | | | |
| | SR1 -3 | 54.3 | 100.0 | | | | | | | | | |
| | SR2 -3 | | 100.0 | | | | | | | | | |
| | SR1 +3 | 52.3 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 100.0 | | | | | | | | | |
| | JP1 | 44.7 | 100.0 | 12.6 | 2.6 | 9.6 | 16.5 | 7.0 | 61.5 | | 0.0 | 100.0 |
| | JP2 | | 100.0 | 37.6 | 2.4 | 34.6 | 40.1 | 5.5 | | 32.0 | 0.0 | 100.0 |
| 3 | Q.TP | 58.7 | 18.0 | 23.8 | 623.5 | -49.6 | 97.5 | 147.1 | 92.5 | 91.5 | 0.5 | 86.5 |
| | C.TP1 | 46.5 | 92.5 | 13.5 | 18.2 | 4.2 | 18.0 | 13.7 | 75.5 | | 0.0 | 100.0 |
| | C.TP2 | | | 35.2 | 57.8 | 30.6 | 41.2 | 10.6 | | 66.0 | | 100.0 |
| | SR1 | 45.0 | 65.0 | | | | | | | | | |
| | SR2 | | 82.5 | | | | | | | | | |
| | SR1 -3 | 45.4 | 52.0 | | | | | | | | | |
| | SR2 -3 | | 82.5 | | | | | | | | | |
| | SR1 +3 | 45.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 72.5 | | | | | | | | | |
| | JP1 | 42.3 | 95.0 | 16.4 | 10.7 | 5.3 | 34.4 | 29.0 | 92.0 | | 0.0 | 100.0 |
| | JP2 | | 98.0 | 35.2 | 8.1 | 20.2 | 45.1 | 24.9 | | 91.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 50.9 | 7.5 | 22.3 | 376.7 | -53.4 | 104.9 | 158.3 | 96.0 | 94.0 | 0.0 | 82.0 |
| | C.TP1 | 47.2 | 38.5 | 14.0 | 140.7 | -30.4 | 24.1 | 54.5 | 95.0 | | 0.5 | 100.0 |
| | C.TP2 | | | 38.8 | 178.8 | 24.7 | 77.8 | 53.1 | | 88.0 | | 95.0 |
| | SR1 | 45.1 | 19.0 | | | | | | | | | |
| | SR2 | | 28.5 | | | | | | | | | |
| | SR1 -3 | 45.1 | 11.0 | | | | | | | | | |
| | SR2 -3 | | 28.0 | | | | | | | | | |
| | SR1 +3 | 45.4 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 28.5 | | | | | | | | | |

*Continued on next page*

Table 6.4 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JP1 | 42.7 | 70.5 | 20.6 | 13.8 | 3.0 | 40.5 | 37.5 | 98.0 | | 0.0 | 100.0 |
| | JP2 | | 77.5 | 30.4 | 13.1 | 10.3 | 47.8 | 37.5 | | 98.5 | 0.0 | 100.0 |

$\beta_0$: The log of number of cases, **S.M:** Simulated model, **Q.TP:** Quadratic simulated model, **C.TP:** Cubic simulated model, **SR:** Segmented simulated model, **JP:** Joinpoint simulated model, **DV:** The mean deviance of simulated models, **SG%:** The percentage of significant parameters of change, **ETP:** The mean of estimated change point, **SD:** The mean of standard error of estimated change point, **LCL:** The mean of lower confidence level of estimated change point, **UCL:** The mean of upper confidence level of estimated change point, **WD:** The mean width of confidence interval of estimated change point, **CI1%:** The percentage of confidence intervals which contains the first true turning point, **CI2%:** The percentage of confidence intervals which contains the second true turning point, **NO%:** The percentage of times that the coefficients of change is zero, **IN%:** The percentage of times where estimated change points are inside the range of data.

### Summary

When the simulation is carried out 200 times with two true turning points in the middle of data, the residual deviance (DV) is calculated for each simulated model. When the original model is cubic or combined (combination model of cubic and segmented regression), the DV of the quadratic model is the largest because the quadratic model fit to the data is simulated from a cubic equation. When the sample size or $\beta_0$ decreases, the difference of DV between quadratic and other models decreases as the simulated data are more random and do not have a cubic pattern. However, the DV of the cubic, segmented regression and joinpoint models are similar. Similar results are obtained when $n = 35$ and 20. On the other hand, when the original model is segmented regression, the DV of all models are similar. Similar results are obtained when $n = 35$ and 20.

SG% measures the significance of the parameter of change in the simulated model. When the original model is cubic or combined, the percentage of significant parameters of change (SG%) in the quadratic model is less than 70% when $n = 50$ and decreases as $\beta_0$ decreases. Moreover, the SG% in the quadratic model is less than 20% when $n = 35$ and 20. However, the cubic model has 100% significant parameters of change when $n = 50$ and $\beta_0 \geq 5$ but SG%

decreases as $\beta_0$ decreases. The joinpoint model has roughly 100% significant parameters of change when $n = 50$ and $\beta_0 \geq 3$ but becomes less than 80% when $\beta_0 \leq 1.5$. Therefore, the SG% in cubic and joinpoint are similar when $n = 50$ and $\beta_0 \geq 5$ otherwise, the SG% in the joinpoint model is larger. Moreover, the segmented regression model has 100% significant parameters of change when $n = 50$ and $\beta_0 \geq 5$ but decreases as $\beta_0$ decreases. Therefore, the SG% in segmented regression and joinpoint are similar when $n = 50$ and $\beta_0 \geq 5$ otherwise, the SG% in joinpoint model is larger. The SG% in the cubic, joinpoint and segmented regression models decreases as $n$ decreases. On the other hand, when the original model is segmented, the SG% in the quadratic model is often less than 10%. The SG% in the joinpoint model is the largest compared with cubic and segmented regression. The parameter of change is equal to zero (NO%) in 1% of time from all simulated models in different sample sizes.

Estimated change point (ETP) from different methods were used to asses the best method of detecting change points. The ETP from the quadratic method did not locate any of the true turning points and has a very large standard error (SD). When the original model is a cubic model, the ETPs from the cubic method often located both true turning points when $n = 50$ and $\beta_0 \geq 3$ with SD increasing as $\beta_0$ decreases. When $n = 35$, ETPs roughly located true turning points when $\beta_0 \geq 5$ with a large SD. But when $\beta_0 < 5$, the first ETP located the first true turning point only with a large SD. However, when $n = 20$, only the first true turning point is located with a large SD. Therefore, the first true turning point is always located but the second true turning point is located only when $n$ and $\beta_0$ are large. This is because small $n$ and $\beta_0$ produce random variation within simulated data. The ETPs from the joinpoint method roughly located true turning points when $n = 20$ otherwise, ETPs did not locate true turning points. On the other hand, when the original model is a segmented

regression, the first ETP from the cubic method usually located the first true turning point but with a large SD. Similar results are obtained when $n = 35$ and 20. However, the ETPs from the joinpoint method roughly located the true turning points when $n = 20$ with SD is approximately 25% of the number of data points otherwise, the ETPs did not occur anywhere close to the true turning points. In contrast, when the original model is combined, the first ETP from the cubic method often located the first true turning point when $n \leq 35$ with a large SD. The second ETP from the joinpoint method roughly located the second true turning point when only $n = 20$ with a large SD.

The confidence interval of the estimated change point is of interest because it explains the range of estimated change points (WD) and the coverage of the true turning point (CI%). The mean width of the confidence interval (WD) for the estimated change point from the quadratic method is very large in all cases. When the original model is cubic or combined, the WD in the cubic model is about 2 to 3 points wide when $n = 50$ and $\beta_0 \geq 5$. The lower confidence level (LCL) and upper confidence level (UCL) are roughly symmetric around the true turning points. However, the WD increases as $\beta_0$ decreases and exceeds $n$ when $\beta_0 \leq 1.5$ where confidence intervals are asymmetric. When $n = 35$, the confidence interval in the cubic model is wide and increases as $\beta_0$ decreases but when $n = 20$ the WD is greater than $n$ and the confidence intervals are asymmetric around the true turning points. The WD in the joinpoint model is greater than the WD in the cubic model when $n = 50$ and $\beta_0 \geq 3$ and it increases as $\beta_0$ decreases but does not exceed $n$ because LCL and UCL have to be within the range of data. However, when $n \leq 35$, the WD in joinpoint is less than the WD in cubic. The confidence intervals are asymmetric in all cases. Moreover, when the data is simulated from a cubic model, the confidence interval of the first estimated change point from the cubic method covers 95% of

the first true turning points when $n = 50$ and $\beta_0 \geq 3$. But when $n = 35$ and $\beta_0 \geq 5$, the confidence intervals of both estimated change points covers 95% of both true turning points. Otherwise, the CI% is out of the 95% coverage or the WD is very large and exceeds $n$. The CI% from the joinpoint model is often under 95% when WD is less than $\frac{n}{2}$ but when WD is greater than $\frac{n}{2}$, the CI% is roughly within 95% coverage. In contrast, when the data is simulated from the combined model, the CI% from the cubic model is under 95% when WD is less than $n$. If WD is greater than $n$, the CI% is either within or out of the 95% coverage. However, the CI% from the joinpoint model is under the 95% coverage when WD is less than $\frac{n}{2}$ but if WD is greater than $\frac{n}{2}$, the CI% is either within or out of 95%. Similar results are obtained when $n = 35$ and 20. On the other hand, when the original model is segmented, the WD of the estimated change points in the joinpoint model are very large and exceeds $n$ in the cubic model. Therefore, the coverage of true turning points (CI%) is mostly above 95% coverage but some confidence intervals cover 95% of true turning points with very large width. Also, the confidence intervals are asymmetric. Similar results are obtained when $n = 35$ and 20.

The estimated change points from the joinpoint method always occur within the range of data (IN% = 100%). However, the estimated change points from the quadratic method occur about 60%- 85% within the range of data. When the data are simulated from cubic or combined models, the first estimated change point from cubic method is often 100% within the range of data but the second estimated change point is roughly 100% within the range of data when $n = 50$ and when $n = 35$ and $\beta_0 \geq 5$. Otherwise, IN% decreases up to 85% as $n$ and $\beta_0$ decrease. However, when the original model is segmented regression, the IN% decreases as $n$ and $\beta_0$ decrease.

### 6.4.2 Two change points occurring close to the beginning and end of data

We assume that the true turning points occur close to the beginning and the end of the dataset. When $n = 50$, the true turning points are assumed to be at 10 and 40, (see Figure 6.5). When $n = 35$ and 20, the true turning points occur at (7 and 27) and (5 and 15), respectively.

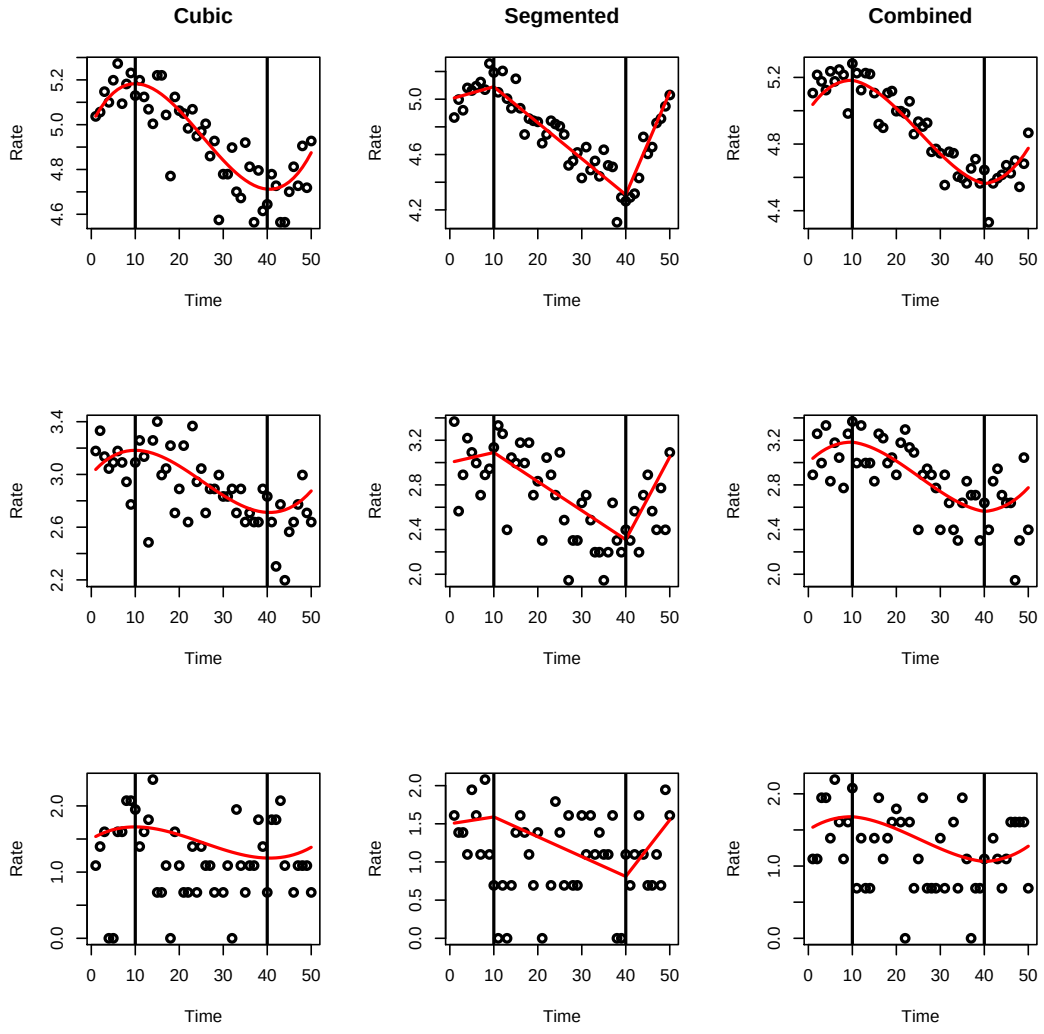



Figure 6.5: Original model (red line) and simulated data (black circles) from cubic, segmented and combined models with two true turning points in the beginning and end (black vertical lines). Plots in the first row associated with $\beta_0 = 5$, the second row with $\beta_0 = 3$ and the third row with $\beta_0 = 1.5$.

Tables 6.5, D.12 and D.13 show the results of simulation studies when $n$ = 50, 35 and 20, respectively.

Table 6.5: Number of data points=50 and the true turning points at 10 and 40.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 89.2 | 4.0 | <1 | >1000 | -810.0 | 817.7 | >1000 | 99.5 | 98.5 | 0.5 | 0.0 |
| | C.TP1 | 46.2 | 100.0 | 9.8 | 0.9 | 7.6 | 11.2 | 3.6 | 90.0 | | 0.0 | 100.0 |
| | C.TP2 | | | 40.5 | 1.0 | 39.1 | 42.8 | 3.7 | | 84.0 | | 100.0 |
| | SR1 | 45.5 | 97.0 | | | | | | | | | |
| | SR2 | | 94.5 | | | | | | | | | |
| | SR1 -3 | 47.8 | 66.0 | | | | | | | | | |
| | SR2 -3 | | 99.5 | | | | | | | | | |
| | SR1 +3 | 48.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 67.0 | | | | | | | | | |
| | JP1 | 43.6 | 100.0 | 12.3 | 4.1 | 8.4 | 18.8 | 10.4 | 76.5 | | 0.0 | 100.0 |
| | JP2 | | 100.0 | 38.5 | 4.4 | 31.1 | 42.8 | 11.6 | | 79.0 | 0.0 | 100.0 |
| 3 | Q.TP | 53.1 | 4.0 | 40.8 | >1000 | -331.7 | 364.2 | 695.9 | 98.5 | 96.0 | 0.0 | 22.0 |
| | C.TP1 | 46.9 | 65.0 | 9.6 | 325.0 | -13.0 | 17.0 | 30.0 | 98.5 | | 0.0 | 98.5 |
| | C.TP2 | | | 41.8 | 600.0 | 31.0 | 66.9 | 35.9 | | 97.0 | | 95.5 |
| | SR1 | 44.8 | 31.0 | | | | | | | | | |
| | SR2 | | 20.5 | | | | | | | | | |
| | SR1 -3 | 45.1 | 19.0 | | | | | | | | | |
| | SR2 -3 | | 34.5 | | | | | | | | | |
| | SR1 +3 | 45.4 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 14.5 | | | | | | | | | |
| | JP1 | 42.3 | 86.5 | 18.2 | 12.8 | 3.5 | 37.3 | 33.8 | 96.0 | | 0.0 | 100.0 |
| | JP2 | | 86.0 | 32.0 | 12.8 | 11.7 | 47.3 | 35.5 | | 95.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 50.9 | 5.5 | 12.2 | >1000 | -143.3 | 179.1 | 322.4 | 97.5 | 93.0 | 0.0 | 59.5 |
| | C.TP1 | 48.5 | 18.0 | 12.6 | 412.0 | -52.4 | 25.0 | 77.4 | 98.0 | | 1.5 | 96.0 |
| | C.TP2 | | | 46.6 | 209.0 | 22.0 | 106.8 | 84.8 | | 97.0 | | 87.0 |
| | SR1 | 46.8 | 8.5 | | | | | | | | | |
| | SR2 | | 11.5 | | | | | | | | | |
| | SR1 -3 | 46.9 | 7.0 | | | | | | | | | |
| | SR2 -3 | | 12.0 | | | | | | | | | |
| | SR1 +3 | 46.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 8.0 | | | | | | | | | |
| | JP1 | 43.8 | 65.5 | 21.7 | 14.5 | 3.0 | 42.6 | 39.7 | 96.0 | | 0.0 | 100.0 |
| | JP2 | | 66.0 | 30.1 | 14.4 | 7.9 | 48.5 | 40.6 | | 97.0 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 49.8 | 18.5 | >50 | >1000 | -107.9 | 141.4 | 249.3 | 95.0 | 80.5 | 1.0 | 72.5 |
| | C.TP1 | 46.8 | 27.0 | 16.4 | 144.0 | -46.0 | 27.9 | 73.9 | 84.5 | | 5.0 | 99.0 |
| | C.TP2 | | | 34.6 | 144.0 | 18.5 | 99.4 | 80.9 | | 78.0 | | 93.0 |
| | SR1 | 44.7 | 6.5 | | | | | | | | | |
| | SR2 | | 27.0 | | | | | | | | | |
| | SR1 -3 | 44.3 | 8.5 | | | | | | | | | |
| | SR2 -3 | | 37.5 | | | | | | | | | |
| | SR1 +3 | 44.8 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 9.0 | | | | | | | | | |

*Continued on next page*

Table 6.5 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JP1 | 41.8 | 79.0 | 24.7 | 15.0 | 3.2 | 43.0 | 39.8 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 81.0 | 33.2 | 12.6 | 12.1 | 48.0 | 35.9 | | 96.5 | 0.0 | 100.0 |
| 3 | Q.TP | 47.8 | 5.0 | 26.1 | >1000 | -63.2 | 116.3 | 179.5 | 93.0 | 94.5 | 0.0 | 75.5 |
| | C.TP1 | 46.6 | 8.0 | 15.9 | 531.0 | -76.3 | 28.4 | 104.7 | 96.5 | | 4.0 | 97.0 |
| | C.TP2 | | | >50 | 203.9 | 21.2 | 125.2 | 104.0 | | 96.5 | | 87.0 |
| | SR1 | 44.5 | 4.0 | | | | | | | | | |
| | SR2 | | 5.0 | | | | | | | | | |
| | SR1 -3 | 44.5 | 2.5 | | | | | | | | | |
| | SR2 -3 | | 11.5 | | | | | | | | | |
| | SR1 +3 | 44.8 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 4.5 | | | | | | | | | |
| | JP1 | 41.7 | 70.0 | 24.0 | 15.4 | 2.8 | 43.7 | 40.9 | 98.0 | | 0.0 | 100.0 |
| | JP2 | | 71.0 | 30.4 | 14.7 | 8.0 | 48.5 | 40.5 | | 99.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 48.8 | 6.5 | 28.8 | 791.3 | -59.5 | 112.1 | 171.5 | 91.5 | 93.5 | 0.0 | 84.0 |
| | C.TP1 | 47.7 | 6.5 | 14.5 | 744.0 | -66.8 | 28.4 | 95.2 | 95.5 | | 0.5 | 98.5 |
| | C.TP2 | | | 44.0 | 304.0 | 22.6 | 117.1 | 94.5 | | 99.0 | | 88.0 |
| | SR1 | 45.9 | 5.0 | | | | | | | | | |
| | SR2 | | 6.0 | | | | | | | | | |
| | SR1 -3 | 45.9 | 3.5 | | | | | | | | | |
| | SR2 -3 | | 7.0 | | | | | | | | | |
| | SR1 +3 | 45.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.5 | | | | | | | | | |
| | JP1 | 42.9 | 55.5 | 22.6 | 15.2 | 2.7 | 43.9 | 41.1 | 97.0 | | 0.0 | 100.0 |
| | JP2 | | 54.5 | 28.0 | 15.6 | 7.6 | 48.7 | 41.1 | | 99.0 | 0.0 | 100.0 |

Original model is combined model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 100.8 | 5.0 | >50 | >1000 | -954.3 | 1001.2 | >1000 | 99.0 | 99.0 | 1.0 | 0.0 |
| | C.TP1 | 45.8 | 100.0 | 9.1 | 0.8 | 7.2 | 10.4 | 3.3 | 81.5 | | 0.0 | 100.0 |
| | C.TP2 | | | 40.8 | 0.8 | 39.5 | 42.7 | 3.3 | | 86.0 | | 100.0 |
| | SR1 | 45.2 | 98.5 | | | | | | | | | |
| | SR2 | | 99.0 | | | | | | | | | |
| | SR1 -3 | 48.0 | 82.5 | | | | | | | | | |
| | SR2 -3 | | 100.0 | | | | | | | | | |
| | SR1 +3 | 47.9 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 82.0 | | | | | | | | | |
| | JP1 | 43.5 | 100.0 | 11.4 | 2.8 | 8.1 | 16.6 | 8.5 | 83.0 | | 0.0 | 100.0 |
| | JP2 | | 100.0 | 39.5 | 3.0 | 34.1 | 43.1 | 9.0 | | 86.0 | 0.0 | 100.0 |
| 3 | Q.TP | 55.5 | 4.5 | <1 | >1000 | -430.0 | 461.2 | 891.2 | 98.5 | 98.0 | 0.5 | 9.0 |
| | C.TP1 | 46.8 | 86.0 | 9.1 | 49.8 | -4.4 | 14.4 | 18.8 | 97.0 | | 0.0 | 99.0 |
| | C.TP2 | | | 41.2 | 109.5 | 35.4 | 56.0 | 20.6 | | 95.0 | | 98.5 |
| | SR1 | 44.8 | 38.5 | | | | | | | | | |
| | SR2 | | 33.5 | | | | | | | | | |
| | SR1 -3 | 45.4 | 17.5 | | | | | | | | | |
| | SR2 -3 | | 51.0 | | | | | | | | | |
| | SR1 +3 | 45.2 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 19.0 | | | | | | | | | |
| | JP1 | 42.6 | 92.0 | 17.1 | 10.7 | 4.5 | 36.4 | 31.9 | 92.0 | | 0.0 | 100.0 |
| | JP2 | | 93.0 | 33.3 | 11.0 | 15.1 | 46.7 | 31.7 | | 92.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 51.4 | 6.5 | >50 | 775.2 | -181.5 | 234.4 | 415.9 | 97.0 | 94.5 | 0.5 | 45.5 |
| | C.TP1 | 48.7 | 24.5 | 10.9 | 143.0 | -47.6 | 25.4 | 73.0 | 99.0 | | 0.5 | 99.0 |
| | C.TP2 | | | 45.1 | 460.0 | 22.9 | 101.4 | 78.5 | | 99.5 | | 91.0 |
| | SR1 | 47.0 | 10.0 | | | | | | | | | |

*Continued on next page*

Table 6.5 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR2 | | 10.5 | | | | | | | | | |
| | SR1 -3 | 47.2 | 7.0 | | | | | | | | | |
| | SR2 -3 | | 13.0 | | | | | | | | | |
| | SR1 +3 | 46.4 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 6.5 | | | | | | | | | |
| | JP1 | 44.0 | 66.0 | 21.2 | 14.5 | 2.8 | 41.4 | 38.6 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 65.5 | 29.7 | 14.2 | 9.0 | 48.4 | 39.4 | | 97.5 | 0.0 | 100.0 |

See Table 6.4 for the definition of $\beta_0$, **S.M, Q.TP, C.TP, SR, JP, DV, SG%, ETP, SD, LCL, UCL, WD, CI1%, CI2%, NO%, IN%**.

### Summary

The residual deviance (DV) of the simulated models has a similar pattern as when two true turning points are in the middle of the dataset. The significance of the parameter of change (SG%) in the quadratic model is about 5% when the original models are cubic or combined (combination model of cubic and segmented regression) at $n = 50$. However, it is about 20% when the data are simulated from segmented regression and the SG% decreases as $\beta_0$ decreases. Similar results are obtained when $n = 35$ and 20. When the original models are cubic or combined, the SG% of the parameter of change in the cubic model has the same pattern when the true turning points occur in the middle of the dataset. Also, the SG% decreases as $n$ decreases and becomes about 5% when $n = 20$. However, the SG% in the joinpoint model is 100% when $n = 50$ and $\beta_0 \geq 5$ and the SG% decreases as $n$ and $\beta_0$ decrease. Therefore, the SG% in the cubic and joinpoint models are similar when $n = 50$ and $\beta_0 \geq 5$ otherwise, the SG% in the joinpoint model is larger and the difference increases as $n$ decreases. In contrast, the results of the SG% in segmented regression has a similar pattern when the true turning points occur in the middle. On the other hand, when the original model is segmented regression and $n = 50$, the SG% of the parameter of change in segmented regression is often very small (about 5%) but the SG% of the parameter of change in the cubic model is often less than 30% and in the joinpoint model is roughly less than 80% where SG% decreases

as $\beta_0$ decreases. The parameter of change of equal to zero (NO%) is rare where in the cubic simulated model when the original data is a segmented data this usually happens 5% of time but about 0% in other original models. Similar results are obtained when $n$ = 35 and 20.

The estimated change point (ETP) from the quadratic model has a similar pattern when the true turning points occur in the middle. When the original model is a cubic model and $n$ = 50 and $\beta_0 \geq 3$, the ETPs from the cubic method often located both true turning points with a standard error (SD) that increases as $\beta_0$ decreases. When $n$ = 35 and $\beta_0 \geq 5$, ETPs often located true turning points with a large SD. However, when $n$ = 20 and $\beta_0 \geq 5$, only the first true turning point is roughly located with a very large SD. Therefore, the first true turning point is always located but the second true turning point is located only when $n$ and $\beta_0$ are large. However, the ETPs from the joinpoint method did not locate true turning points. In contrast, when the original model is combined, the ETPs from the cubic method has similar results when the data are estimated from a cubic model. In addition, when $n$ = 50 and $\beta_0 \leq 1.5$, only the first true turning point is roughly located with a large SD. The ETPs from the joinpoint method roughly located both true turning points with small SD only when $n$ = 50 and $\beta_0 \geq 5$. On the other hand, when the data are simulated from segmented regression, none of the methods locate true turning points.

The mean width of confidence interval (WD) for the estimated change point from the quadratic method has similar results when true turning points occur in the middle. When the original models are cubic or combined, the pattern of results of WD in the cubic model is the same as when the true turning points occur in the middle of the dataset except that the width here is about 3 points and the confidence interval is asymmetric. The WD in the joinpoint model also

has a similar pattern of results when the turning points occur in the middle and the WD increases more than $\frac{n}{2}$ as $\beta_0$ decreases. The percentage of coverage of true turning points (CI%) in the cubic and joinpoint methods is under 95% when the WD is small at $n = 50$ and $\beta_0 \geq 5$. However, when the WD is large, the CI% covers 95% of true turning points when $n = 50$ and $\beta_0 = 3$. Otherwise, the CI% has no clear pattern because the WD is very large and exceeds $n$. Similar confidence interval findings are obtained when the true turning points occur in the middle of the dataset for data simulated from segmented regression models.

The estimated change points from the joinpoint method always occur within the range of data (IN%=100%). The estimated change points occurring within the range of data with two turning points for polynomial methods has similar results to IN% when the true turning points occur in the middle of the dataset (see Table 6.4). Exceptionally, when the data are simulated from a cubic or combined model and $n = 50$ and $\beta_0 \geq 5$ , none of the estimated change points from the quadratic method occur within the range of data and when $\beta_0 < 5$, the IN% is about 10%- 45%.

### 6.4.3 Conclusion

Two change points can be detected roughly in the middle of the data by the cubic polynomial method when the pattern of the data seems cubic and $n \geq 50$ and $\beta_0 \geq 3$. However, two change points can be detected roughly in the beginning and at the end of the data by cubic polynomial and joinpoint methods when the pattern of the data seems mixed between straight lines and cubic models and $n \geq 50$ and $\beta_0 \geq 5$. However, cubic polynomial can detect two change points when $n \geq 50$ and $\beta_0 \geq 5$ where the data is simulated from cubic pattern. If any of the above situations do not meet, different methods need to be investigated to detect the change points.

## 6.5 Discussion and conclusion

In this chapter we have performed simulation studies to compare change point methods and we investigated the best method to detect the change in the trend of count data. Three different patterns of data are assumed; data with no change in trend, data with one change in trend and data with two changes in trend. We identified initial (original) models; polynomial GLM models (linear, quadratic and cubic), segmented models (with one and two changes) and the combinations between polynomial and segmented models. These models are identified according to the method of estimating change points. The polynomial method was discussed in Chapter 4 when change points are estimated from quadratic and cubic models. Chapter 5 explained how to detect change points by segmented regression and joinpoint analysis methods. Since the polynomial models show the quadratic and cubic models as the best to detect change points and segmented models show the joinpoint method as an effective method to detect change points, the combination (polynomial and segmented) model is considered. Also, a linear model is considered as the original model when there is no change in the trend of data. Since the change can occur at any time, the beginning, middle and end of the time period are chosen to locate true turning points. As this research deals with count data which can be rare, different numbers of cases are assumed; $\beta_0$= 5, 3 and 1.5. The limitation in these simulation studies is that the coefficients in original models $\beta_i$s, $i = 1, 2, 3, 4, 5$ are similar but not identical when $\beta_0$ and $n$ change so the comparison between methods is good but not very precise.

The criteria used for comparing different methods are identified. The residual deviance (DV) of the simulated models is used to detect the best fitted model. The percentage of significant parameters of change (SG.CH%) over 200 simulations is calculated in the simulated models to assess the significance

of the change. The location of the estimated change points (TP.ES) detects how close the TP.ES is to the true turning points. The mean width of the estimated confidence intervals of estimated change points (CI.WD) is found to determine the width of the confidence interval of the estimated change points. The percentage of true turning points within the estimated confidence interval (CI.TP%) was used to determine 95% coverage of the true turning point. The use of percentile bootstrap confidence interval led to low coverage of true turning points. The CI.WD and CI.TP% can be improved by using different methods of bootstrap confidence interval such as a bias corrected method or bias corrected and accelerated method. Moreover, choosing $\beta_i$s, $i = 1, 2, 3, 4, 5$ to be slightly different (<0.05) in different sample sizes or different locations of the true turning points may affect the CI.WD and CI.TP%. Thus, using identical values of $\beta_i$s here to produce accurate confidence intervals is recommended (see Section 4.6). Finally the percentage of estimated change points in the range of the dataset (TP.IN%) was computed.

The simulation started off with setting up the initial values to simulate data based on a Poisson distribution. The simulated data was used to fit different models (quadratic, cubic, segmented -at true turning points and at true turning points ±3- and joinpoint models) then, change points were estimated. The simulated models and estimated change points were compared using the criteria explained above to investigate the best method of detecting change points.

Using the methods; polynomial and joinpoint to detect a change where there is actually no change helps explain false positives. The polynomial method detected change points in a linear trend (i.e. with no change points) in approximately 5% of simulations which indicates that the polynomial method is a good method in order to detect the change points. In contrast, the joinpoint

method detected a change point in a linear model in approximately 25% of simulations where the estimated change points from the joinpoint models always occur at the middle with a large standard error and large mean width of confidence interval. This indicates the weakness in the detection of change points. In conclusion, the polynomial methods are good techniques to detect no change when the pattern of data is linear but joinpoint method is not effective. Moreover, a linear pattern may occur with data that has change in the level and therefore in such event different methods should be investigated to detect the change. Taylor (2000) used cumulative sum (CUSUM) with bootstrapping to detect the sudden change in the trend of data over time.

Detecting one change point has different results based on the initial model and the location of change point. Generally, when the true turning point occurs in the middle and the original model is quadratic, all methods; quadratic polynomial and joinpoint analysis are able to detect change points in the middle. However, when $n \geq 50$ and $\beta_0 \geq 5$, the DV of the quadratic model is much smaller than the DV of other models so in such case the quadratic polynomial method is better to detect change points in the middle. However, when $n \leq 20$ and $\beta_0 \leq 1.5$, all methods do not detect change points and therefore different methods are required. In contrast, when the original model is segmented regression, the segmented and joinpoint methods are better to detect change points in the middle. These methods are usually better to detect the change point when the original model is combined and $n \geq 35$ and $\beta_0 \geq 3$. On the other hand, the quadratic polynomial method is better to detect change points when they occur in the beginning or at the end of the data when $n \geq 20$ and $\beta_0 \geq 3$ and the original model is quadratic. Otherwise, better methods need to be derived, (see Figure 6.6).

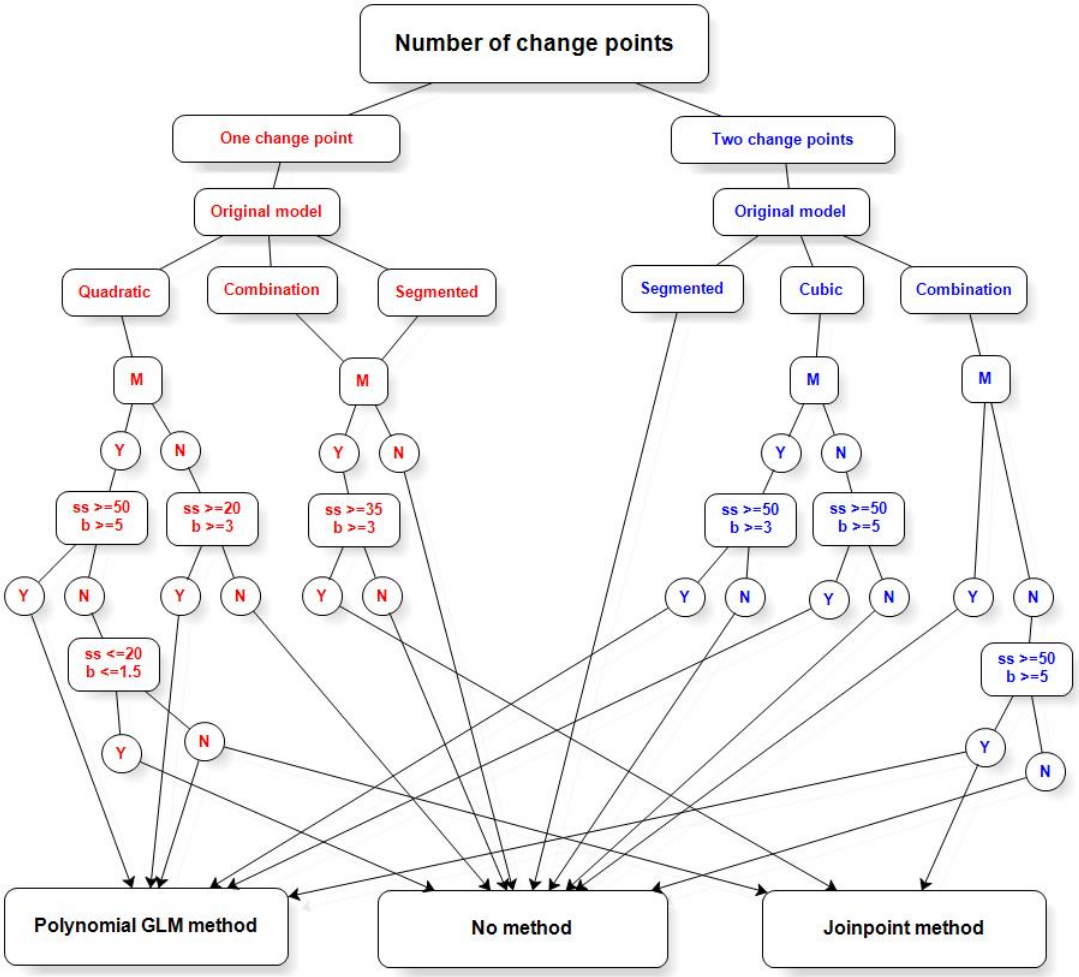



Figure 6.6: The best change point method. The abbreviations mean; M: the true turning point occurs in the middle of dataset, Y: Yes, N: No, b: $\beta_0$ and ss: sample size.

According to different original models, the simulation study gives different conclusions to the best method of detecting two change points. In general, when two true turning points occur roughly in the middle of the data, the cubic polynomial method detects two change points when $n \geq 50$ and $\beta_0 \geq 3$ and the data is simulated from the cubic model. Otherwise, no method can detect two change points in the middle. On the other hand, when two true turning points occur roughly in the beginning and at the end of the data and $n \geq 50$ and $\beta_0 \geq 5$, the cubic polynomial method detects two change points when the original model is cubic model. However, the cubic polynomial and joinpoint

methods detect two change points when the original model is mixed (combination between cubic and segmented) and $n \geq 50$ and $\beta_0 \geq 5$. Otherwise, other methods are required, (see Figure 6.6).

In conclusion, the polynomial regression method depending on one curve over time and the joinpoint method depending on piecewise linear trends over time are useful for detecting change points roughly in large datasets. In the next chapter, we will develop a method for detecting change points of count data using piecewise polynomial. This method uses the generalized additive model (GAM) to fit count data using a smooth spline function of the time variable.

# Chapter 7

# Spline and Generalized Additive Model Analysis

Detecting change points is an important issue in epidemiological studies. Chapters 4 and 5 described methods based on polynomial generalized linear models (GLM) and straight lines models, respectively to detect change points. The nature of HAIs data involves the rate of infection increasing/decreasing steeply and dropping off slowly. A quadratic model cannot fit such data as the feature of a quadratic equation where the rate has an identical pattern before and after the change point but with the structure reversed (increases and decreases), as in a mirror. Cubic and quartic models (polynomial GLMs) can fit the rate of infection well but they may have problems with prediction. They may also estimate the change points outside the range of data. The join-point method is based upon straight lines which may not be appropriate to fit curvature pattern.

In this chapter, a new approach is considered to detect the change points. Generalized additive models (GAM) are more flexible than polynomial GLM and can model a smoother trend to the data. This chapter consists of Section 7.1

which introduces spline and GAM models, Section 7.2 involving the method of estimating change points and their confidence intervals from GAM models, Section 7.3 which presents the results of the HAIs data and Section 7.4 which includes the discussion and conclusion.

## 7.1 Spline function and generalized additive model analysis

### 7.1.1 Introduction

A spline is a function $s(x)$ where mathematically it is piecewise polynomial over an interval $(a, b)$ with continuous derivatives for all points of $(a, b)$. For some given positive integer $r$ and a sequence of knot points; (i.e. points at which the parts of spline function join) $t_1, t_2, ..., t_k$, where $a < t_1 < t_2 < ... < t_k < b$, $s(x)$ is required that on each subinterval $(t_j, t_{j+1})$, the $s(x)$ is a polynomial of order $r$. Also, on the interval $(a, b)$, $s(x)$ is continuous and has $r - 1$ continuous derivatives [Wold (1974) and Bowman and Azzalini (1997)]. When the positive integer $r$ is chosen to be 3, the curve of spline fonction is called a cubic spline and is constructed from parts of a cubic polynomial which is joined together by the knot points. Each part of the cubic polynomial function has different coefficients while at the knot points, the value and first two derivatives of the function are similar in two adjoining parts [Hastie and Tibshirani (1987)].

In statistics, splines are used for data smoothing. The simple (univariate) smooth function is presented in a model with one predictor;

$$y_i = s(x_i) + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{7.1}$$

where $y_i$ is a response variable, $x_i$ is an explanatory variable, $n$ is the number of observations and $s(x_i)$ is given by:

$$s(x) = \sum_{i=1}^{k-1} \beta_i b_i(x), \tag{7.2}$$

which is a smooth function based on some polynomial basis functions $b_i(x)$ with unknown parameters $\beta_i$, $k$ is the number of knot points (i.e. $k-1$ is the number of polynomial basis functions $b_i(x)$) and $\varepsilon_i$ are the errors which have an independent and identical distribution with $N(0, \sigma^2)$ random variables [Wood (2006)]. When the basis functions are cubic splines, the regression function (7.1) is estimated by minimizing the penalized least squares function [Wood (2006)]

$$D = \frac{1}{n} \sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_a^b [s''(x)]^2 dx. \tag{7.3}$$

The first term in expression (7.3) is the residual sum of squares (RSS) which depends on the data, and the second term is called a roughness penalty which depends on a smoothing parameter $\lambda$ and the cubic smoothing spline $s(x)$. $\lambda$ gives a smooth fitted curve but if $\lambda = 0$, minimising $D$ will be associated with the data only (i.e. RSS) which is not estimating the regression spline function (7.1). If $\lambda \to \infty$, the second derivative is restricted to be zero because very large smoothing is equivalent to fitting a straight line to the data (i.e. $s(x)$ is a linear function thus $s''(x) = 0$).

### 7.1.2 Cubic spline in Poisson regression

Spline regression can be fitted in **R** programming language using the package **mgcv** [Wood and Wood (2017)] and smooth curves can be estimated. This approach extends generalized linear regression models (**glm**) to generalized additive models (**gam**). The **gam** function in **R** operates in a similar manner to the linear regression models (**lm**) and **glm** functions [Bowman and Azzalini

(1997)].

In generalized additive models, the predictor variables include some smooth functions (e.g. $s(t)$ is cubic spline function) that estimate the response variable $y$ which can follow an exponential family distribution [Wood (2006)]. In the case of count data, a Poisson distribution is used to fit a spline model within the **gam** function in **R** and the response variable is on a log scale. The model of rate (count per unit) is obtained as:

$$\log(\text{no.Cases}) \sim \text{offset}(\log(\text{Population})) + \beta_0 + s(t, bs = "cr", k), \tag{7.4}$$

where no.Cases is the number of incidences (the response variable), Population is the total number of population at risk and $t$ is the exploratory variable. $s(t, bs = "cr", k)$ is a predictor variable in the form of a spline function and defined in Equation (7.2) where $bs = "cr"$ indicates the penalized smoothing basis function for cubic regression spline and $k$ is the number of knot points used to represent the smooth term (see Equation (7.2)). The coefficient $\beta_0$ estimates the intercept and log(Population) is an offset variable (the denominator of the rate) which has a coefficient of 1.

The criteria of goodness of fit for the GAM model (7.4) depends on the smoothing parameter ($\lambda$) which minimizes the penalized likelihood Poisson regression function,

$$D_{\text{Poisson}} = \frac{-1}{n} \sum_{i=1}^{n} (y_i s(t_i) - e^{s(t_i)}) + \frac{\lambda}{2} \int_a^b [s''(t)]^2 dt, \tag{7.5}$$

where $s(t) \in [a, b], \quad a, b > 0$ [Gu (2013)]. Two methods are reviewed in order to estimate smoothing parameters. When the scale parameter (residual variance $\sigma^2$) is known, minimizing the un-biased risk estimator (UBRE) (see Equation (7.6)) gives good estimates of smoothing parameters leading to a good fitting

model, [Wood and Wood (2017)].

$$\text{UBRE} = \frac{DV}{n} + \frac{2\sigma^2 edf}{n - \sigma^2}, \tag{7.6}$$

However, if the scale parameter is unknown, minimizing the generalized cross validation (GCV), (see Equation (7.7)) results in a good model, [Craven and Wahba (1978)].

$$\text{GCV} = \frac{n \times DV}{(n - edf)^2}. \tag{7.7}$$

Where $DV$ is the residual deviance of the model, $n$ is the number of observations, $\sigma^2$ is the scale parameter and $edf$ is the effective degrees of freedom of the model. The effective degrees of freedom measures the flexibility of the fitted GAM model and it is calculated as the trace of the hat matrix of the GAM model (i.e. $tr(H)$). This is an adjusted method to define the number of degrees of freedom when the number of free parameters is undefined (i.e. depends on smoothing parameters of spline functions) [Wood (2006)].

The residual deviance and percentage of deviance explained (PDE) of the GAM model are also used to detect the best fitted model. PDE is calculated from residual deviances of the GAM and null models (log(rate) = 1) and is obtained as:

$$\text{PDE} = \left(1 - \left(\frac{DV}{NDV}\right)\right) \times 100, \tag{7.8}$$

where $NDV$ is the deviance of the null model.

## 7.2 Estimating change points and their confidence intervals from spline models

In order to investigate change points, a spline model was considered with seasonal effect of time to estimate change points and their confidence intervals. Using **R** (see the algorithm in Appendix E.1) to fit a **gam** model by using Poisson or quasi-Poisson distributions for count data as follows:

$$\log(\text{no.Cases}) \sim \text{offset}(\log(\text{Population})) + \beta_0 + s(t, bs = "cr", k) + \gamma_{Qu(t)}, \quad (7.9)$$

where $t$ is a year, $Qu$ is a seasonal effect (considered as a main factor affecting the rates) and the coefficient $\gamma$ estimates the seasonal effect.

Change points can be estimated using the GAM model (7.9) which involves obtaining predictions of the smooth term (spline function $s(t)$) then, calculating the first and the second derivatives of $s(t)$. The first derivative is obtained by $s'(t_i) = s(t_i) - s(t_{i-1}), i = 2, ..., n$, where $n$ is the number of observations. The second derivative is obtained by $s''(t_j) = s'(t_j) - s'(t_{j-1}), j = 3, ..., n$. The location of change points identifies when the sign of the first derivative changes while the sign of the second derivative determines whether the change points are maximum or minimum values. Therefore, change points can be detected from the first derivative of the spline function $s(t)$ and the magnitude of the second derivative controls how much the gradient changes.

To construct confidence intervals for change points, bootstrapping is used. The following algorithm shows how confidence intervals for change points of the GAM model are constructed, (see Appendix E.1).

1. Calculate the Pearson residuals of the original GAM model (7.9) then, re-sample them.

2. Calculate a new responses ($y_i^*$) (simulated data) which are obtained as $y_i^* = \hat{y}_i + \hat{\varepsilon}_i \times \sqrt{\hat{y}_i}$ where $\hat{y}_i$ are the predictions of the response variable and $\hat{\varepsilon}_i$ are the re-sampled residuals.

3. Fit the GAM model with same number of knot points ($k$) as in the original model to the simulated data of the new response; $y_i^*$.

4. Calculate all possible change points for the bootstrapped spline model by the first and second derivatives as explained above.

5. Find the location (time) of all possible bootstrapped change points.

6. Calculate the number of all possible bootstrapped change points.

7. Save the values of the bootstrapped spline function which occur at the time of all possible bootstrapped change points.

8. Choose the values of the bootstrapped spline function ($s(t)$) based on the actual change points. In the case of one actual change point, the minimum or maximum value of bootstrapped $s(t)$ is chosen. In the case of two actual change points, the minimum and maximum values of the bootstrapped $s(t)$ are chosen.

9. Find the locations of the chosen values from the previous step of bootstrapped $s(t)$ which are the subset of the set including all possible bootstrapped change points. These locations represent the time of the bootstrapped change points.

10. Calculate and save the time of the year of the bootstrapped change points.

11. Repeat steps 1 to 10, B times (say B=10,000).

12. Calculate the 95% percentile bootstrap confidence intervals for each change point using (**quantile**) at 0.025 and 0.975 in **R**.

There is a problem when fitting GAM within bootstrap. When carrying out the procedure, the fitted GAM model sometimes returns a large number of change points with the simulated data. This can occur because the curvature is quite smooth but the variability around the curve is high. This produces over-fitting of data with a large number of changes that do not have a strong evidence of change. In contrast, the fitted GAM model sometimes returns a number of change points less than the number of actual change points from the original GAM model. This happens because the bootstrapped sample (simulated data) has less variability and a fitted GAM model with a very smooth curve does not show any change. Therefore, a pragmatic solution was adopted to discard samples producing change points fewer than the actual number of change points or more than 10% of the actual number of data points used in the analysis. This solution was derived by trial and error method.

## 7.3 Results on HAIs

Fitting GAM models and the method of estimating change points and their confidence intervals was applied to data on HAIs (MRSA bacteraemia, MSSA bacteraemia and CDI) until March 2016. According to the GAM model including a cubic spline as a function of time, the interventions which impact the rates of infection were detected.

### 7.3.1 MRSA bacteraemia

In order to fit the best model which detects the most important interventions impacting the rates of MRSA bacteraemia, the following strategy is conducted. As MRSA bacteraemia is a count data, use a Poisson distribution to fit model

(7.9). The no.Cases is the number of MRSA bacteraemia, Population is the AOBDs which stands for acute occupied bed days in Scottish hospitals, $Qu(t)$ is the seasonal effect and $s(t)$ is the cubic spline function of the time. Fit model (7.9) starting with $k = 10$ for a spline function as a default number in **R** programming language. The deviance (DV) of the model is 57.8056, the percentage of deviance explained (PDE) of the model is 98.2% and the unbiased risk estimator (UBRE) is 0.4583. This model detects one change point during the period of study. In order to get the best model, reduce the number of $k$ for the spline function which gives the following:

1. The GAM model with $k = 9$ gives DV =57.4593, PDE=98.2%, UBRE=0.4530 and detects one change point.

2. The GAM model with $k = 8$ gives DV=58.5215, PDE=98.2%, UBRE=0.4607 and detects one change point.

3. The GAM model with $k = 7$ gives DV=57.4258, PDE=98.2%, UBRE=0.4371 and detects one change point.

4. The GAM model with $k = 6$ gives DV=60.9340, PDE=98.1%, UBRE=0.4689 and detects one change point.

Therefore, the best fitted GAM model to detect one change point is model (7.9) with $k = 7$ for the spline function. This fitted model gives effective degrees of freedom of 5.369 for the spline function (i.e. effective degrees of freedom of the model minus the number of other parameters in the model) and R-sq adjusted is 0.981. The method of detecting change points is used and the location of the change point is estimated at 2005.5 when the maximum rate is predicted. The 95% percentile bootstrap confidence interval for the estimated change point is (2005.25, 2006.25) based on 9,985 iterations out of 10,000 samples, (see Figure 7.1 and Table 7.1). Figure 7.1 shows the predicted line of MRSA bacteraemia on

the spline scale and the estimated change point with 95% confidence interval. It is plotted on the spline scale not on the fitted values because the fitted values have got a seasonal effect on them where the GAM model was adjusted by seasonality which makes the interpretation of the plot more difficult.

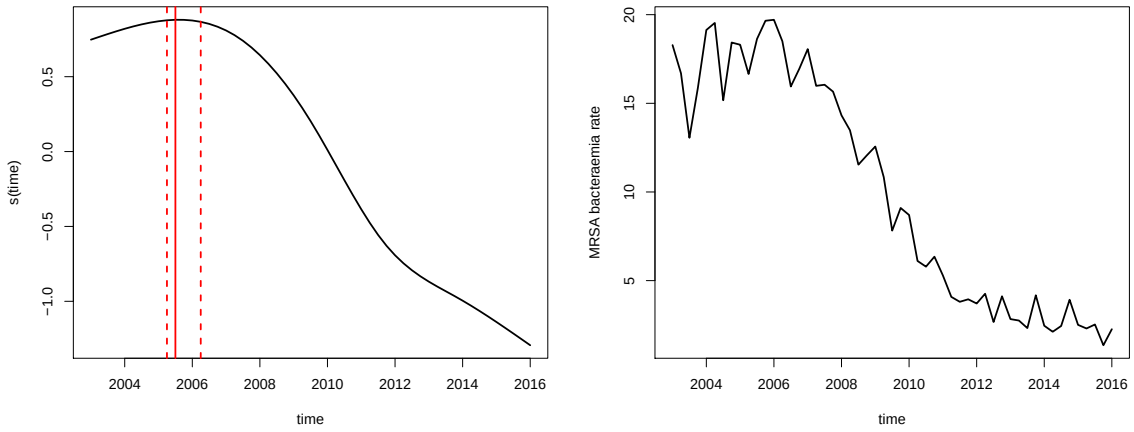



Figure 7.1: Left figure shows GAM model for MRSA bacteraemia (black curve) and estimated change point (red solid line) with 95% confidence interval (red dashed lines). Right figure shows raw data for MRSA bacteraemia.

Table 7.1: Estimated change points from GAM and their confidence intervals.

| Infection | ETP | Bootstrap confidence interval | PDE |
|---|---|---|---|
| MRSA bacteraemia | Qu3, 2005 | (Qu2, 2005 - Qu2, 2006) | 98.2% |
| MSSA bacteraemia | Qu2, 2012 | (Qu3, 2010 - Qu1, 2014) | 33.4% |
| CDI in patients over 65 years | Qu4, 2007 | (Qu3, 2007 - Qu1, 2008) | 98.4% |
| CDI in patients aged 15-64 years | Qu2, 2014 | (Qu1, 2013 - Qu4, 2014) | 93.1% |

**ETP:** Estimated change points, **PDE:** Percentage of deviance explained, **Qu:** Quarter.

In Chapter 4 the method of estimating change points for the rate of MRSA bacteraemia used data up to September 2014. The polynomial GLM model detected three change points while the GAM model detected only one change point in the data up to March 2016. This is partly because the polynomial model may detect changes outside the range of data but GAM does not. Also, GAM fits a smooth curve and as a result it is difficult to detect small changes in the gradient.

### 7.3.2 MSSA bacteraemia

A similar approach was done with MSSA using a Poisson distribution in model (7.9) with MSSA data and $k = 10$ for spline function. This model gives DV=78.9034, PDE=45%, UBRE=1.3206 and detects three change points (2006.5, 2008, 2012.25). There are six points of data only between the first and the second change points. Reducing the value of $k$ to 9 gives DV=78.7725, PDE=45.1% and UBRE=1.3049. This model also, gives three change points (2006.75, 2008, 2012) and five points of data between the first and the second change points. If $k = 8$, DV=85.8009, PDE decreases to 40.2%, the UBRE increases to 1.3888 and this model gives three change points which is not better than the model with $k = 9$.

The occurrence of two change points close to each other (with less than eight data points between them) is not useful which affects the interpretation of change points and cannot illustrate whether the interventions had an impact on the rate of infections, (see Section 7.4). This was addressed by reducing the value of $k$ to have at least eight data points between two change points. Choosing $k = 5$ gives DV=95.5099, PDE is 33.4% and UBRE=1.5071. This model detects one change point. However, this model shows over-dispersion ($p<0.001$) so a quasi-Poisson distribution should be used. The model with quasi-Poisson and five knot points shows DV=98.9909, PDE=31.0%, GCV=3.0828 and one change point. In order to find the best model under a quasi-Poisson distribution, the number of $k$ is increased and the results are recorded.

1. Fitting GAM model with $k = 6$ gives smaller GCV (3.0693), smaller DV (97.3657), larger PDE (32.1%) and detects one change point.

2. Fitting GAM model with $k = 7$ gives GCV=3.0613, DV=96.2481, PDE=32.9% and detects one change point.

3. Fitting GAM model with $k = 8$ decreases GCV to 3.0596, DV to 95.7597, increases PDE to 33.2% and detects one change point.

4. Fitting GAM model with $k = 9$ gives GCV=3.0595, DV=95.5458, PDE=33.4% and detects one change point.

5. Fitting GAM model with $k = 10$ increases GCV to 3.0633, DV to 96.1233, reduces PDE to 33.0% and detects one change point.

Therefore, the GAM model with a quasi-Poisson distribution and $k = 9$ for the spline function is the best model in detecting one change point. It gives effective degrees of freedom of 2.931 for the spline function and R-sq adjusted is 0.338. The method of estimating change points locates the change point at 2012.25 with minimum rate. The 95% percentile bootstrap confidence interval of the estimated change point is (2010.50, 2014) based on 8,403 simulations out of 10,000 samples, as shown in Figure 7.2 and Table 7.1.

In Chapter 4, data up to September 2014 was used and there was no evidence of a change point when polynomial GLM was fitted to MSSA rates. Figure 3.9 showed the general trend of MSSA up to September 2014 which was almost a straight line, thus it is only with extra data the change point can be detected.

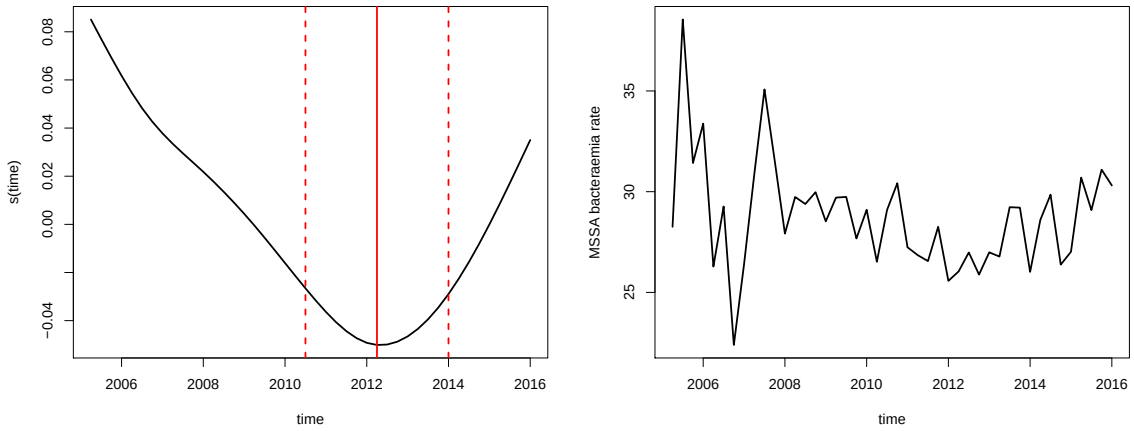



Figure 7.2: Left figure shows GAM model for MSSA bacteraemia (black curve) and estimated change point (red solid line) with 95% confidence interval (red dashed lines). Right figure shows raw data for MSSA bacteraemia.

### 7.3.3 CDI in patients over 65 years

The same procedure was adopted to fit a GAM model for CDI in patients over 65 years. Using the Poisson distribution and a spline function with $k = 7$ showed over-dispersion ($p<0.001$). A quasi-Poisson distribution is better to fit the data and the spline function has effective degrees of freedom of 5.394, the model deviance is 182.2533, percentage of deviance explained is 98.4%, R-sq adjusted is 0.979 and generalized cross validation (GCV) is 8.4633. The location of the change point is at 2007.75 when a maximum rate is predicted with 95% percentile bootstrap confidence interval (2007.50, 2008) based on 10,000 out of 10,000 samples as shown in Figure 7.3 and Table 7.1.

In Chapter 4, the data of CDI in patients over 65 years up to September 2014 was used and the polynomial GLM detected three change points. GAM detected only one change on the data up to March 2016 because GAM fits a smooth curve which does not easily detect small changes.

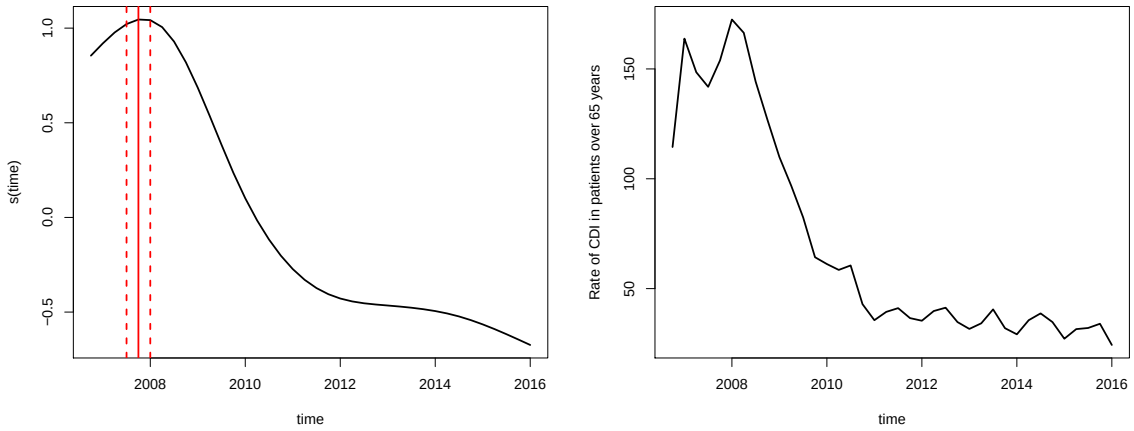



Figure 7.3: Left figure shows GAM model for CDI in patients over 65 years (black curve) and estimated change point (red solid line) with 95% confidence interval (red dashed lines). Right figure shows raw data for CDI in patients over 65 years.

### 7.3.4 CDI in patients aged 15-64 years

Fitting a GAM model to CDI in patients aged 15-64 years with a Poisson distribution and spline function with $k = 10$ identified three change points. However, the first and second change points have poor locations (i.e. there are only two time points between the first and the second change points). Re-fitting the GAM model with a spline function with $k = 7$ gave 5.226 effective degrees of freedom, the model deviance was 29.4231, percentage of deviance explained was 93.1%, UBRE was 0.7098 and R-sq adjusted was 0.93. The change point is located at 2014.25 with a minimum rate and 95% percentile bootstrap confidence interval (2013, 2014.75) based on 9,679 iterations out of 10,000 samples, see Figure 7.4 and Table 7.1.

GAM detected one change point in the data of CDI in patients aged 15-64 years up to March 2016 whereas in Chapter 4 the polynomial GLM did not detect changes in the data up to September 2014 due to complex numbers.

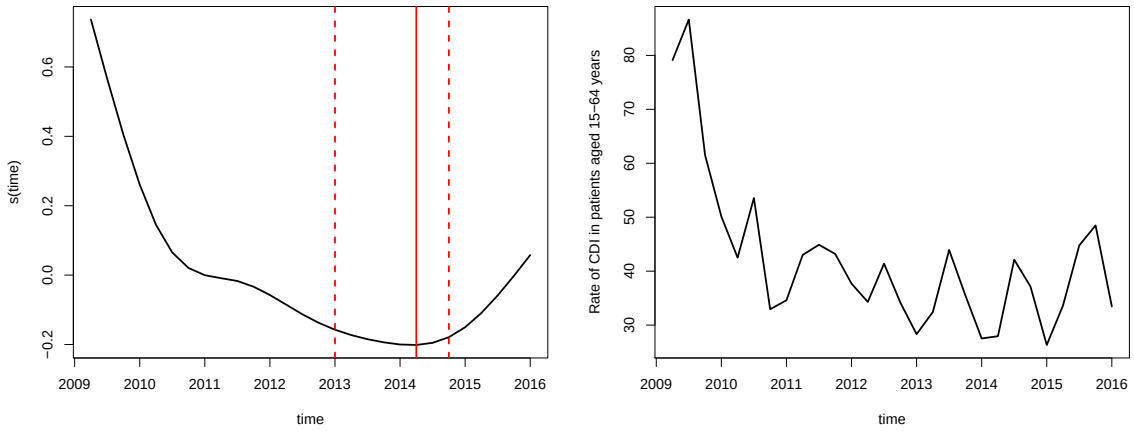



Figure 7.4: Left figure shows GAM model for CDI in patients aged 15-64 years (black curve) and estimated change point (red solid line) with 95% confidence interval (red dashed lines). Right figure shows raw data for CDI in patients aged 15-64 years.

In conclusion, using GAM models with spline functions give smooth fitted curves that detect one change point. This change point gives an indication of the time when the intervention had an impact on the rate of infections.

## 7.4 Discussion and conclusion

Estimating change points from regression models is an important approach in identifying healthcare interventions which impact the rate of HAIs. Spline regression models can be used to describe the change in the trend of HAIs. These detect the change at the time when the data was monitored (quarterly time during the period of study). This gives the most significant change points through the smooth spline function.

The methods of modelling GAM and obtaining the first derivatives from smoothing functions within GAM to detect the change are used by few authors. The GAM model with two spline functions; a cubic regression spline for the year and cyclic cubic regression spline for seasonal effect was fitted by Curtis and

Simpson (2014). However, our approach assumes one cubic regression spline for the year and the response variable (count observations) is adjusted by the categorical variable; seasonal effect with four factors. The finite differences method was used by Curtis and Simpson (2014) to obtain the first derivatives of the cubic regression spline for the year. This method detected the period of significant change in the trend over time. In contrast, our approach applied the method to detect significant change points where the trend changed. Following this, the second derivatives are calculated to determine whether the trend changes to decrease or to increase.

### 7.4.1 GAM and polynomial GLM regression

The GAM model is more flexible than the polynomial GLM regression, especially when constructing confidence intervals for change points. In polynomial GLM regression, when the change is detected from a quadratic model, there is one change point. When using bootstrapping to construct a 95% confidence interval for that change point, one estimated change point is detected each time and the confidence interval is then calculated from these estimated points. However, in a GAM model when the model detects one change point, bootstrapping cannot easily control the number of estimated change points each time. If this happens (i.e. fixing the number of bootstrapped change points each time to be same as the number of actual change points), many bootstrap samples may be discarded. In bootstrapping, $k$ is fixed as in the original GAM model and the location of knot points are similar in all bootstrap models because cubic regression spline ($bs = "cr"$) places knot points by quantile. The quality of a fitted GAM model is then only based on the bootstrap samples. The bootstrap sample produces change points either less, equal or more than the number of actual change points which indicates that changes in the gradient occur as a result of random variation. Less change points occur because the

bootstrap sample has much less variation and therefore the fitted GAM model with a very smooth curve does not change in gradient. In such cases, the bootstrap samples were discarded. The same number of change points occur because the bootstrap sample has less variation and the GAM model is fitted well. More change points occur because the curvature is quite smooth with large variability around it. Change points may occur more than 10% of the actual number of data points because high smoothness of the curvature with large variation results in over-fitting of data with a large number of changes which may not have strong evidence of change (i.e. small gradient). In such cases, the bootstrap samples were discarded. The algorithm of calculating confidence intervals for the actual change points from the original GAM model picks each time the closet estimated change point to the actual change point and then calculates the confidence interval from estimated change points using the percentile bootstrap method.

With polynomial GLM, at times the change points are estimated outside the range of data. The advantage of the GAM approach is that all change points would be within the range of data because the GAM method does not estimate any change outside the data. This method only looks at the first and second derivatives within the range of data. The GAM model also estimates the change points within the range of data which can then be matched with one or more interventions during the period of study. Moreover, in both polynomial GLM and GAM methods, when the estimated change points occur at the beginning or at the end of the dataset and the fitted line is smoother around the estimated change point, the confidence interval of estimated change point is asymmetric. In polynomial GLM when estimated change points occur in the beginning, they are close to the upper level and this is the opposite if they occur at the end, they are close to the lower level. In GAM, if the estimated change point occurs

in the beginning, it is close to the lower level and if it occurs at the end, it is close to the upper level. Such difference between the two methods is due to GAM estimating the change point and its confidence interval to occur within the range of data so when the estimated change point occurs in the beginning there are few data points before it and more data points after it. The estimated change point and its confidence interval from the polynomial GLM can occur at any time, even outside the range of data.

The number of estimated change points from GAM models is different from polynomial GLM models of HAIs data and this is due to various reasons. Different datasets were used to fit the models. GAM models fit data up to March 2016 but polynomial GLM models only fit data up to September 2014 due to data availability at the time of analysis. Also, GAM models fit quite smooth curves to data which detects less change points whilst polynomial GLM estimates change points based upon the order of the polynomial function. In addition, polynomial GLM detects change points either inside or outside the range of data but GAM estimates the change points only within the range of data. Finally, the polynomial GLM may not detect any change points if there is no real solution to the polynomial equation (i.e. complex roots).

### 7.4.2 Challenges of GAM models

Using GAM spline models to estimate the number of change points and their locations has some difficulties and challenges. In MSSA and CDI in patients aged 15-64 years, the fitted models with $k = 10$ detected more than one change point during the period of study. The challenge with this type of detection is that some of the estimated change points occur close to each other (less than eight data points between them) which affects the interpretation of change points and cannot illustrate whether the interventions had an impact on the

rate of infections. Also, confidence intervals for change points close to each other overlap indicating that there should be only one change point during this period of time, especially with a smooth trend. We then searched for the model with change points that are not close to each other with at least eight data points (two years of data) between them. The second derivative controls how quickly the gradient changes. When the second derivatives are large, the gradient had a large change but when the second derivatives are small, small changes in gradient occur which indicates that there may be no change points. The small $k$ is related to a large gradient. Therefore, reducing the value of $k$ decreases the oscillations in the trend and reduces the number of change points. The change point with the smallest GCV in MSSA occurs when $k = 9$, and the smallest UBRE in CDI in patients aged 15-64 years occurs when $k = 7$. In other infections; MRSA and CDI in patients over 65 years we found one change point when the number of $k = 10$ but the smallest UBRE and GCV were observed when $k = 7$ which indicates a better fit of the models.

The other challenge is associated with constructing bootstrap confidence intervals using a fixed number of iterations. When the bootstrap GAM model gives too many or too few bootstrapped change points, the bootstrap samples are discarded. The large number of bootstrapped change points more than 10% of the number of data points is unexpected and lots of change points will occur with less than eight data points between them. The number of bootstrapped change points less than the number of actual change points from the original model should not be allowed because the bootstrapping aims to simulate change points for each actual change point. Each actual change point should have a unique sample of bootstrapped change points. Thus, one bootstrapped change point cannot be repeated in samples of two actual change points. Also, if the bootstrap model does not find a change point when there is

one actual change point, this bootstrap sample is discarded.

### 7.4.3 Confidence intervals for estimated change points from GAM models

Confidence intervals for estimated change points from spline models depend on the pattern and goodness of fit of the smooth spline fitted line. The confidence intervals of estimated change points from spline models become narrower when the model is fitted well (i.e. percentage of deviance explained $\geq$ 95%). This is the case for MRSA bacteraemia and CDI in patients over 65 years, (see Table 7.1 and Figures 7.1 and 7.3). However, if the percentage of deviance explained is less than 95% then, the confidence interval of the estimated change point is wide. This is the case for MSSA bacteraemia and CDI in patients aged 15-64 years, (see Table 7.1 and Figures 7.2 and 7.4). The best fitted model for MSSA bacteraemia and CDI in patients aged 15-64 years detected three change points as explained in the results (see Section 7.3). In order to estimate fewer change points, different models were fitted which had less percentage of deviance explained. This can lead to wide confidence intervals for the estimated change point.

Confidence intervals of estimated change points from the spline model are almost symmetric if the trend of the model changes with large curvature (i.e. large gradient) before and after the estimated change points, such as the confidence intervals of estimated change points from MSSA bacteraemia and CDI in patients over 65 years. On the other hand, the confidence interval is asymmetric in MRSA bacteraemia and CDI in patients aged 15-64 years due to a smooth change (flatter) of trend before and after the estimated change points. The estimated change point and its confidence interval occur within the range of the data. The estimated change point from MRSA bacteraemia is closer to

the lower level than to the upper because the estimated change point occurs in the beginning of the data. Similarly, the estimated change point from CDI in patients aged 15-64 years is closer to the upper level than to the lower because the estimated change point occurs at the end of the data.

### 7.4.4 Effective healthcare interventions

The location of maximum estimated change points in the spline models gives an idea of when the rates start to decrease and may help identify clinical interventions associated with this. Some of the healthcare interventions listed in Table 1.2 are associated with a reduction in infection rates. Spline models can also estimate minimum change points within the range of data which may indicate healthcare interventions having no impact or may indicate an outbreak of the infection. The estimated change points from spline regression models around 2012- 2013 for MSSA bacteraemia and around 2014- 2015 for CDI in patient aged 15-64 years may indicate the start of an upward trend. In conclusion, Table 7.2 summarizes the interventions which may have impacted the rates of HAIs in Scotland.

Table 7.2: Summary of the interventions that have been detected by spline GAM models which may have impacted the rate of HAIs in Scotland.

| Point of change | Time of intervention | Intervention | MRSA | CDI |
|---|---|---|---|---|
| Qu3, 2005 | February 2005 | CNO letter on alcohol based hand rubs and infection control. | Yes | |
| | March 2005 | CNO requested that all G Grade Sisters/ Charge Nurses (Senior Charge Nurses) undertake the Cleanliness Champions Course commenced. | Yes | |
| | July 2005 | New IC structure in Boards, including ICM funding. | Yes | |
| | August 2005 | Antimicrobial Prescribing Policy and Practice in Scotland- Recommendations for good antimicrobial practice in acute hospitals. | Yes | |

*Continued on next page*

Table 7.2 – *Continued from previous page*

| Point of change | Time of intervention | Intervention | MRSA | CDI |
|---|---|---|---|---|
| Qu4, 2007 | March 2007 | Scottish Patient Safety Programme (SPSP) announced. | | Yes |
| | December 2007 | First national hand hygiene compliance report issued. | | Yes |

**MRSA:** Methicillin-resistant staphylococcus aureus, **CDI:** Clostridium difficile infection, **Qu:** Quarter, **CNO:** Chief nursing officer, **IC:** Infection control, **ICM:** Intensive care medicine.

In conclusion, the change points problem is an important analysis to detect the presence of changes in infection rates. Spline regression models detect changes at any time within the range of data where they are powerful in the detection of change, especially when the model is a good fit. The idea in this research is to detect the time at which an intervention had taken place to reduce the rate of healthcare associated infections. The research recommends healthcare interventions such as giving training courses for hospital staff to deal with infection, applying antibiotic policy to reduce and prevent the infection in hospitals and healthcare systems and improved hand hygiene.

# Chapter 8

# Conclusions and Further Work

Healthcare associated infections (HAIs) are a major factor of patient morbidity and mortality, especially methicillin-resistant staphylococcus aureus (MRSA) bacteraemia, methicillin-sensitive staphylococcus aureus (MSSA) bacteraemia and clostridium difficile infection (CDI). The infections can be transmitted either outside or inside hospital settings, causing serious diseases. The National Health Services (NHS) and Health Protection Scotland (HPS) established, developed and improved healthcare interventions to control infection as means to avoid HAIs. Some of these interventions took place in Scotland between 2004-2011 (see Table 1.2) to tackle the rates of HAIs and the infection rates have subsequently decreased [HPS (2015b)].

It is of interest to identify time points when changes occur in order to identify interventions associated with these changes. Usually change point detection analysis has two aspects in statistical inference. The first is the hypothesis test to detect if there is any change in the observed data. The second is to estimate the number of changes and their locations. Statisticians implement many different change point detection methods to estimate the number and location of significant change points.

This thesis aimed to discuss and develop some statistical methods to detect change points which reflect impacted interventions on HAIs and to observe the impact of seasonal effect on these changes. The first approach was polynomial generalized linear models (GLM) which is a new method for this data. Segmented regression and joinpoint analysis were then applied and these use linear trends. The joinpoint analysis was developed by considering a seasonal effect in the model. The third method was a generalized additive model (GAM) which was used to estimate change points where a spline function of time was fitted by GAM and adjusted with a seasonal effect. All these methods are considered within the scope of HAIs. This research also considered methods of constructing confidence intervals for the change points. Finally, simulation studies were established to investigate the best method of detecting change points and the best method of constructing confidence intervals for the change points.

A background on HAIs including MRSA, MSSA and CDI and some related risk factors were introduced in Chapter 1. A literature review of modelling and change point detection methods was presented in Chapter 2. Modelling the rate of HAIs was investigated in Chapter 3. The strength of this research was presented in the rest of the PhD thesis where change point methods were investigated, developed and applied to HAIs. These methods were addressed to detect the change points in count data and they have a flexibility to investigate the impact of more than one covariates. Joinpoint software was developed using R programming language and polynomial and joinpoint methods were combined to develop spline GAM to detect change points. In addition, the change point methods were compared by simulation algorithm in order to investigate the best way of detection change points. A brief summary and

conclusion from these methods will be discussed and some future research suggested in the following sections.

## 8.1 Modelling rare events of count data

In this research, our methodology was built on the Poisson distribution and quasi-Poisson distribution in the case of over-dispersion. These fit count data of rare events accurately. However, if the data had a lot of zeros (like in small health boards in our data), the Poisson and quasi-Poisson are not appropriate because of sample bias. The use of a zero-inflated Poisson distribution and zero-inflated negative binomial distribution to model this type of data is recommended.

Modelling the data describes the change over Scotland and also in individual health boards. The data for Scotland (large number of cases of infection) can be fitted by polynomial GLM and the inclusion of a seasonal effect describes the change in the trend well. However, this polynomial Poisson model is not appropriate to investigate the change in small datasets (e.g. when there is a small number of cases in some health boards) because there is not enough power to fit a complicated model to small data.

Adding health board as a factor to the polynomial GLM model allows us to describe the change in each health board however the change in Scotland overall cannot be described. The model with the health board factor explains that most health boards have similar patterns but some of small health boards have a different pattern because the Poisson distribution was used with data containing lots of zeros.

Change in the trend of the rate of HAIs is discussed in literature by many authors worldwide. Most of studies agreed with our conclusion of general change (reduction in HAI rates) and seasonal change. Daneman et al. (2012) reported a reduction in CDI in Canada between 2002- 2010. They used Poisson regression and account for correlation using autoregressive models. Modelling time trend was adjusted with other factors such as age and hospital type. In addition, Worth et al. (2016) evaluated the trend for CDI incidence in Australia between 2010- 2014 using Poisson regression which demonstrated general reduction started after the peak in fifth quarter. There was evidence of seasonal trend with higher rates in summer and lower rates in winter which agreed with our analysis. Moreover, a significant reduction in MRSA rate was observed in 2012 (40%) compared to 2010 (53%) [Perovic et al. (2015)]. They attributed this decline to the implementation of the infection control. Kinoshita et al. (2017) studied the change in trend of MRSA in seven European countries between 1999- 2015. They reported that countries with more interventions (policies) to control MRSA (the UK, France, Belgium, Germany, and the Netherlands) had a greater reduction in MRSA rates than those with fewer interventions (Spain and Italy). The reduction in MRSA rates after implementation of mandatory surveillance was observed in the UK, France, Belgium and Germany. The change in trends may be due to budgets are different in each country to implement control infection policies.

In contrast, some few studies reported an increasing trend of HAIs. Moxnes et al. (2015) estimated and predicted the trend of MRSA in Norway between 1997- 2010. They used Poisson and gamma Poisson distributions to model the trend of MRSA. They found that the incidence of MRSA was increasing and will continue to increase until 2017. They attributed this increasing to rise of importing from abroad due to population mobility which was investigated and

discussed in the recent study by Di Ruscio et al. (2017).

Funnel plots were used to investigate if any health board is significantly different from Scotland overall in terms of infection rates. This method investigates whether the risk factors (surgical procedure and teaching hospital) affect the rates of infection and show high rates for Scotland overall. It would be of interest to use the intervention information in funnel plots to see if rates of HAIs differ. This suggestion investigates whether funnel plots would be useful to compare rates after a specific intervention to identify the impact of the intervention on individual health boards.

## 8.2 Change points of polynomial GLM and GAM

The polynomial generalized linear model (GLM) estimates the number and location of change points based on the order of a polynomial function and the coefficients of polynomial terms. HPS (2016a) reported general reduction of HAI rates but it is not clear when the change took place. However, fitting polynomial regression as in Chapter 3 and using the change point detection method as explained in Chapter 4 gave an illustration when the change took place and reflect on the associated interventions. We concluded that if the model fits the data well, with small deviance, this gives a good estimate of the change points. If the model is a poor fit, the estimated change points are inaccurate. Parts of these works has been drafted for journal publication in the Journal of Hospital Infection and is titled as "The Impact of NHS Infection Control Interventions on Rates of Healthcare Associated Infections".

A spline function within a generalized additive model (GAM) can be used to estimate the change point. This approach fits a smoother model to the data. If a change point is detected, it is likely to be a significant change since it was

estimated from a smoother trend. Curtis and Simpson (2014) used the finite differences method to detect the period of change when the trend of spline function is significantly increased or decreased. However, our method used spline function and finite differences method and focused on detecting change point when significant change in the trend took place (i.e. when the trend change from increase to decrease or vice versa) and then reflected the associated intervention with these changes. In addition, we constructed confidence interval for the change points using bootstrapping. Parts of this work has been drafted for journal publication in Communications in Statistics journal and is titled as "Estimation of Change Points from Regression models of Count Data".

A spline GAM model is more flexible than a polynomial GLM model. A bootstrap method was used in both models to construct 95% confidence intervals for estimated change points. If the original data gives one estimated change point, a polynomial GLM model was fitted to each bootstrap sample to give one change point. In a GAM model, each bootstrap sample was fitted to give the best model which may estimate one or more change points. The algorithm then chooses the closest point to the actual change point (from the original data). Bootstrap confidence intervals of estimated change points from the polynomial GLM model are usually narrower than those of estimated change points from the GAM model because GAM gives a flatter (smoother) trend than polynomial GLM. Also, the bootstrap confidence interval in the polynomial GLM model discards the bootstrap samples which gives complex numbers as a result of change points.

Polynomial GLMs can estimate change points outside the range of data which are uninterpretable in relation to interventions, especially if the change was estimated before any intervention took place. In contrast, GAM models

estimate the change points within the range of data so this is easier to attribute to one or more interventions. However, estimated change points from a spline GAM model is approximately similar to one of the estimated change points from polynomial GLM. This estimated change point is similar and in the same position (minimum or maximum) in both methods. This similarity illustrates the strength of both methods to detect change points since polynomial GLM models and spline GAM models fit a curvature trend to the data. Table 8.1 shows the results of change points from different HAIs up to June 2016. For example, an estimated change point from the polynomial GLM when the rate of MRSA bacteraemia decreased at 2005.31 and an estimated change point from GAM at 2005.50, also when the rate decreased.

The number of estimated change points of HAI data from polynomial and GAM models are different. The spline GAM model detects one change point exactly on one of the data points. However, the estimated change points from the polynomial GLM can be at any time during the period of study. Polynomial GLM estimates change points based on the order of polynomial function but a GAM model fits a smooth curve to the data which detects the most important change point (with large gradient). In addition, polynomial GLM may detect change points outside or very close to sides (i.e. at the very beginning or at the very end) of the data which estimates more change points. GAM estimates the change points only within the range of data. Finally, complex numbers may occur when solving polynomial equation so polynomial GLM cannot detect any change points.

In polynomial GLM models, the confidence interval of the estimated change point can be constructed easily using the delta method which calculates the expected value and variance for the ratio of two random variables based on a

Table 8.1: Change points using different methods for HAIs data up to June 2016.

| Infection | Change point | | | |
|---|---|---|---|---|
| | Polynomial GLM with 95% BCI | GAM with 95% BCI | Segmented regression | Joinpoint analysis with 95% BCI |
| MRSA bacteraemia | Min@ 1998.12 (1975.07, 2002.47) Max@ 2005.31 (2004.99, 2005.61) Min@ 2015.05 (2014.60, 2015.89) | Max@ 2005.50 (2005.25, 2006.25) | 2006.25 2008.50 2011 | 2007.50 (2007.25, 2007.75) 2012 (2010.25, 2013.25) |
| MSSA bacteraemia | Min@ 2011.7 (2010.67, 2014.56) | Min@ 2012.25 (2010.50, 2014) | 2011 | - |
| CDI in patients over 65 years | Max@ 2007.62 (2007.50, 2007.72) Min@ 2012.75 (2012.47, 2013.14) Max@ 2014.60 (2014.28, 2014.92) | Max@ 2007.75 (2007.50, 2008) | 2007.75 | 2008.25 (2008, 2008.50) 2011 (2010.50, 2011.50) |
| CDI in patients aged 15-64 years | Min@ 2013.03 (2012.42, 2013.69) Max@ 2016.48 (2014.75, 2028.77) | Min@ 2013.75 (2013, 2014.75) | 2009.75 | 2010.75 (2010.50, 2011.25) |

**BCI:** Bootstrap confidence interval, **Min:** Minimum rate, **Max:** Maximum rate.

Taylor expansion. This approach is difficult to apply for more than one change point because the change points formulas are complicated nonlinear functions with more than two variables. Therefore, a bootstrap method is used to construct confidence intervals for more than one change point. Some bootstrap samples return complex numbers and in such cases these samples are discarded. Limitations relating to the bootstrap method require further research. Some bootstrap samples return one or more change points outside the range of data which can affect the calculation of the mean and standard error and give wide confidence intervals. Bootstrap samples which return at least one change point outside the range of data should be discarded. Also, the change points estimated in each bootstrap sample may not have the same position as the estimated change points from the original data where they may occur at minimum or maximum values. Future research may improve this issue to ensure the position of the bootstrap change point is in the same position as the

estimated change point from the original data.

In spline GAM models, the bootstrap confidence interval is based on the estimated change points from the bootstrap samples. When the original data has one change point, the bootstrap sample may estimate no change point or large number of change points. The large number of change points is defined to be greater than 10% of the number of data points (e.g. when sample size is 50, the large number of change points is greater than 5). In such cases, these samples are discarded. However, when the bootstrap sample estimates the number of change points to be between one and 10% of the number of data points, the closest change point to the actual change point with same position (minimum or maximum) was selected. This algorithm gives a more accurate confidence interval, especially with a smoother trend since all change points close to the actual change point are considered.

## 8.3 Segmented linear regression and joinpoint analysis

Chapter 5 introduced segmented linear regression and the method to detect change points. This technique fits a linear trend before and after an intervention took place. This detects the most significant intervention if the period of implementation is short which allows us to investigate if the intervention had an impact on the rate of infection before any other interventions took place. For a long time of implementation, segmented polynomial regression may be better to describe the data before and after the intervention. It is of interest to develop segmented linear to segmented polynomial regression to study the curvature fits before and after the specific intervention. This approach would be useful when there is only one intervention investigated. If there is one particular in-

tervention and polynomial segmented regression fits the data well, polynomial segmented regression gives an indication of whether there are previous or later turning points. If there are previous turning points, other interventions may have impacted the rates. If there are later turning points, there is an indication that a particular intervention had an impact later in time (i.e. not at the same time when a particular intervention took place).

If there are many different interventions during a period of time, our method of segmented linear regression detects change points at the time $i$ where there could be more than one intervention before time $i$. Thus, the change in the rate may occur due to these interventions working together impacting the infection rate. Therefore, the intervention at time $i$ may or may not have impacted the rate.

Joinpoint analysis was developed to improve segmented linear regression analysis. The joinpoint method looks for the change point at each data point and can detect the change point either at the time of intervention or after the intervention took place. Joinpoint software ([NCISR (2017)]) considers model with one independent variable and detects the number and location of change points on that variable. However, we developed the joinpoint model to be able to add more covariates to the model and then detects the change points in the main independent variable. This development allows to understand the impact of the covariates on the trend and then may affect the detection of joinpoints. This work has been drafted for publication in Communications in Statistics journal and is titled as "Change Points Analysis for the Trend of Count Data". It would be of interest to use Hudson's method [Hudson (1966)] with our modified joinpoint model to detect change points not only at each point of data but also between them.

Joinpoint analysis allows the interpretation of the change in the rate. If the change is detected after the intervention took place, it suggests that this intervention had an impact on the rate and the effect occurs at a later point. However, if the change is detected after several interventions took place, it is difficult to identify which intervention had a major impact on the rate. In conclusion, the nature of the data and interventions explain that the change point should occur after the time of the intervention. This can be clearly shown in Table 8.1 where the first detected change point in MRSA bacteraemia using joinpoint analysis occurred at 2007.50 and after 5 quarters from the detected change point by using segmented regression where the intervention took place at 2006.25. Similarly, the second detected change point in MRSA bacteraemia using joinpoint analysis was shown at 2012, after one year from the detected change point using segmented regression. Similar results are concluded from the CDI data. Therefore, joinpoint analysis is better than segmented linear regression for such type of data where some interventions may take time to have an impact because the intervention takes time to spread over all Scotland and also intervention requires behavioural change from the medical professionals which may take a while to become ingrained.

On the other hand, if the nature of the intervention allows the change to occur exactly when the intervention took place, segmented regression is better. For example, if a sudden and great earthquake (8 Richter or more) occurred in a city, this may increase mortality at this time point. We can say that this intervention (a sudden event) directly caused high mortality. In analysing such cases, the segmented regression is better than joinpoint analysis. This instant impact of an intervention happens rarely in medical studies because most medical studies aim to control disease and reduce the rate over a period of time.

Moreover, the impact of the interventions takes time to spread across all health boards in Scotland and therefore it is not a straightforward problem to detect when there is an impact on the rate of infection. A lot of interventions took place, some of them give evidence of impacting the rates of HAIs and others do not. Change points may occur after two or more interventions and if one intervention took place and had an effect over one year, for example, and then an other intervention happened, it is difficult to determine exactly which intervention had an impact on the rate of infection. Therefore, effective interventions are quite difficult to identify in this research because of the nature of the type of interventions implemented.

Constructing confidence intervals for joinpoints using the profile likelihood method is a good approach for one joinpoint where the models are associated with residual deviance. Lerman (1980) approach to construct confidence intervals for the change point from joinpoint model is similar to our approach when profile likelihood confidence interval was used. However, he used a function depending on residual sum of squares which constructs confidence intervals for more than one change point while our method uses the residual deviance of the models. In contrast, when there is more than one joinpoint, profile likelihood method is not able to find confidence intervals for joinpoints thus we have developed bootstrap methods to construct confidence intervals for one or more joinpoints. This approach gives similar result to profile likelihood method but it requires a lot of computation.

## 8.4 Limitations and further work on simulation

In this thesis, several simulation studies have been conducted to compare different methods which were used in this research.

In Chapter 4, a simulation was carried out to compare bootstrap and delta methods to construct confidence intervals for one change point where the change point is estimated from a polynomial GLM model. Both methods are not recommended for use with small sample sizes ($n \leq 20$). Some technical issues were associated with failure of using the bootstrap method especially with small sample sizes. In order to find two change points, the cubic model is fitted but for small sample sizes ($n \leq 20$) the more complicated polynomial GLM model does not fit well and change points are not accurately estimated. The recommendation here is to use the bootstrap method with a sample size greater than 20. Our approach uses percentile bootstrap confidence intervals which were inaccurate to estimate confidence intervals and cover 95% of the true turning points. A bootstrap bias corrected (BC) or bias corrected and accelerated (BCa) methods may improve this coverage. Sample size may also affect the simulated data and produce samples with very few cases and lots of zeros. Our approach always uses a Poisson distribution to fit a model to the data and because the Poisson distribution is not an appropriate distribution to fit to data with lots of zeros, poorly fitting models are observed. Change points and related confidence intervals from these models are therefore not accurate. Zero-inflated distributions should be used in the case of many zeros observed in simulated data.

In Chapter 5, a simulation study compared bootstrap and profile likelihood methods of constructing confidence intervals for one joinpoint. Profile likelihood confidence interval is a good method when the sample size is $\geq 50$

otherwise, the bootstrap method is better. If the data are simulated originally from a quadratic model, the bootstrap method is good but not the best. It is of interest to investigate other approaches here. The profile likelihood confidence interval method cannot be used with more than one joinpoint so the bootstrap method was used. However, bootstrap confidence intervals need to be improved to cover 95% of actual estimated joinpoints and a way to do this is to use different methods of bootstrap confidence intervals such as BC or BCa confidence intervals.

In Chapter 6, polynomial GLM, segmented and joinpoint approaches to detect change points were compared in simulation studies. There are three main results from this simulation study. Firstly, if the trend of data seems to be linear, polynomial methods usually report no change points, however, the joinpoint approach may sometimes detect changes even in the absence of a change where this illustrate inaccuracy of joinpoint method. Secondly, if the trend of data seems to change at one point in the middle with quadratic pattern, quadratic polynomial and joinpoint analysis are able to detect change points in the middle. However, when the sample size is large, quadratic is better but when the sample size is small, different methods should be investigated. In contrast, when the trend of data before and after a change seems linear, the segmented and joinpoint methods are better to detect change points in the middle when $n \geq 35$ and $\beta_0 \geq 3$. On the other hand, when the change point occurs in the beginning or at the end of the data and $n \geq 20$ and $\beta_0 \geq 3$, the quadratic polynomial method is better to detect change points when the pattern of the data is curved. Otherwise, better methods need to be investigated. Finally, when two true turning points occur roughly in the middle of the data with curved pattern, the cubic polynomial method detects two change points when $n \geq 50$ and $\beta_0 \geq 3$. On the other hand, when two true turning points occur

roughly in the beginning and at the end of the data and $n \geq 50$ and $\beta_0 \geq 5$, the cubic polynomial method detects two change points in a curved pattern of data. However, the cubic polynomial and joinpoint methods detect two change points when the original model is a combination between cubic and segmented. Otherwise, other methods are required.

There are many possible assumptions in which the results in this simulation could be improved in future work. One of these recommendations is to assume serial correlation on the data. It may also be worth considering occurrence of outliers in the data (i.e. 10% or less of data are outliers) to investigate the impact of extreme values on the results of the simulation. It would also be of interest to compare spline GAM methods of estimating change points with other methods through a simulation study. This includes also the simulation studies to estimate change points from GAM and construct bootstrap confidence intervals. Choosing $\beta_i$s to be different in different sample sizes gives different curvature for the trend of data in each sample size. Here the results of confidence intervals between different sample sizes are not accurately comparable. To improve the procedure and have more accurate results for constructing confidence intervals using the bootstrap method, identical values of $\beta_i$s in different sample sizes should be chosen. Furthermore, various values of $\beta_i$s should be investigated.

## 8.5 Conclusion

In conclusion, the contributions to knowledge of this research are to improve and develop statistical models which detect changes in trend of count data. Adapting joinpoint software to include categorical explanatory variables that may influence rates. Also, detecting change point from spline regression on GAM using finite differences method. In addition, constructing confidence intervals for change points using bootstrapping and coding algorithms of simulation

study to compare change point methods. Finally, applying these methods to detect change points in HAI data which gives some recommendations of some impacted interventions which need to be improved to control infection to reduce the coast and mortality. Spline GAM method is suggested to practitioners who try to define the point of change on their systems especially those who are dealing with count data. For example, people who work in medical and epidemiological studies (e.g. policy makers and NHS infection control people) and people who want to estimate the impact of interventions or process control on manufactory process to make their systems more professional or to have costless products. The spline GAM method combined polynomial and joinpoint methods which detects the most important change that have large gradient.

In HPS data (HAIs and interventions), the impact of the interventions takes time to spread across all health boards in Scotland and therefore it is not a straightforward problem to detect when there is an impact on the rate of infection. A lot of interventions took place, some of them give evidence of impacting the rates of HAIs and others do not. Change points may occur after two or more interventions and if one intervention took place and had an effect over one year, for example, and then another intervention happened, it is difficult to determine exactly which intervention had an impact on the rate of infection. Therefore, effective interventions are quite difficult to identify in this research because of the nature of the type of interventions implemented.

# Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Abramson, M. A. and Sexton, D. J. (1999). Nosocomial methicillin-resistant and methicillin-susceptible staphylococcus aureus primary bacteremia: at what costs? *Infection Control and Hospital Epidemiology*, 20(6):408–411.

Agay-Shay, K., Friger, M., Linn, S., Peled, A., Amitai, Y., and Peretz, C. (2012). Periodicity and time trends in the prevalence of total births and conceptions with congenital malformations among Jews and Muslims in Israel, 1999-2006: A time series study of 823,966 births. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 94(6):438–448.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Information Theory, Proceedings of the 2nd International Symposium, 1973, eds. Petrov, B.N. and Caski, F.*, pages 267–281.

Ansari, F., Gray, K., Nathwani, D., Phillips, G., Ogston, S., Ramsay, C., and Davey, P. (2003). Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis. *Journal of Antimicrobial Chemotherapy*, 52(5):842–848.

Austin, P. C. (2008). Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in medicine*, 27(17):3286–3300.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Bellman, R. and Roth, R. (1969). Curve fitting by segmented straight lines. *Journal of the American Statistical Association*, 64(327):1079–1084.

Bellman, R. E. and Dreyfus, S. E. (1962). *Applied dynamic programming*. Princeton university press.

Bielakova, K., Fernandova, E., Matejovska-Kubesova, H., Weber, P., Prudius, D., and Bednar, J. (2016). Can we improve the therapy of clostridium difficile infection in elderly patients? *Wiener klinische Wochenschrift*, 128(15):592–598.

Blischke, W. R. (1961). Least squares estimators of two intersecting lines. Technical report, New York State coll of agriculture and life sciences ithaca.

Borg, M., Hulscher, M., Scicluna, E., Richards, J., Azanowsky, J.-M., Xuereb, D., Huis, A. v., Moro, M., Maltezou, H., and Frank, U. (2014). Prevention of meticillin-resistant staphylococcus aureus bloodstream infections in European hospitals: moving beyond policies. *Journal of Hospital Infection*, 87(4):203–211.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, 33(1):38–44.

Breslow, N. E., Day, N. E., et al. (1987). *Statistical methods in cancer research*, volume 2. International Agency for Research on Cancer Lyon.

Bugna, B. A. L. (2015). *Poisson versus negative binomial regression in the analysis of count data*. PhD thesis, Western Michigan University.

Bühlmann, P. et al. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Cardano, G., Witmer, T. R., and Ore, O. (2007). *The rules of algebra: Ars Magna*, volume 685. Courier Corporation.

Carignan, A., Poulin, S., Martin, P., Labbé, A.-C., Valiquette, L., Al-Bachari, H., Montpetit, L.-P., and Pépin, J. (2016). Efficacy of secondary prophylaxis with vancomycin for preventing recurrent clostridium difficile infections. *The American Journal of Gastroenterology*, 111:1834–1840.

Carpenter, B. P., Hennessey, E. K., Bryant, A. M., Khoury, J. A., and Crannage, A. J. (2016). Identification of factors impacting recurrent clostridium difficile infection and development of a risk evaluation tool. *Journal of Pharmacy and Pharmaceutical Sciences*, 19(3):349–356.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164.

Carter, R., Quesnel-Vallée, A., Plante, C., Gamache, P., and Lévesque, J.-F. (2016). Effect of family medicine groups on visits to the emergency

department among diabetic patients in Quebec between 2000 and 2011: a population-based segmented regression analysis. *BMC family practice*, 17(1):23.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

CFSPH (2011). Methicillin resistant staphylococcus aureus. Available from: The center for food security and public health. `http://www.cfsph.iastate.edu/Factsheets/pdfs/mrsa-citations.pdf`. (Accessed: 20-6-2014).

Chang, S. Y. and Perron, P. (2016). A comparison of alternative methods to construct confidence intervals for the estimate of a break date in linear regression models. *Econometric Reviews*, pages 1–25.

Chen, J. (1998). Testing for a change point in linear regression models. *Communications in Statistics-Theory and Methods*, 27(10):2481–2493.

Chen, J. and Gupta, A. K. (2001). On change point detection and estimation. *Communications in statistics-simulation and computation*, 30(3):665–697.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science and Business Media.

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine*, 21(10):1461–1469.

Chou, Y.-C., Chuang, H. H.-C., and Shao, B. (2015). Information initiatives of mobile retailers: a regression analysis of zero-truncated count data with underdispersion. *Applied Stochastic Models in Business and Industry*, 31(4):457–463.

Christensen, R. (1996). *Analysis of variance, design, and regression: applied statistical methods*. CRC Press.

Chu, F.-L. (2004). Forecasting tourism demand: a cubic polynomial approach. *Tourism Management*, 25(2):209–218.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, 31(6):1852–1884.

Clements, M. S., Armstrong, B. K., and Moolgavkar, S. H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, 6(4):576–589.

Coia, J. E., Leanord, A. T., and Reilly, J. (2014). Screening for meticillin resistant staphylococcus aureus (MRSA): who, when, and how. *BMJ*, 348(1):g1626.

Conterno, L. O., Wey, S. B., and Castelo, A. (1998). Risk factors for mortality in staphylococcus aureus bacteremia. *Infection Control and Hospital Epidemiology*, 19(1):32–37.

Cosgrove, S. E., Qi, Y., Kaye, K. S., Harbarth, S., Karchmer, A. W., and Carmeli, Y. (2005). The impact of methicillin resistance in staphylococcus aureus bacteremia on patient outcomes: mortality, length of stay, and hospital charges. *Infection Control and Hospital Epidemiology*, 26(2):166–174.

Cosgrove, S. E., Sakoulas, G., Perencevich, E. N., Schwaber, M. J., Karchmer, A. W., and Carmeli, Y. (2003). Comparison of mortality associated with methicillin-resistant and methicillin-susceptible staphylococcus aureus bacteremia: a meta-analysis. *Clinical infectious diseases*, 36(1):53–59.

Cox, R., Conquest, C., Mallaghan, C., and Marples, R. (1995). A major outbreak of methicillin-resistant staphylococcus aureus caused by a new phage-type (EMRSA-16). *Journal of Hospital Infection*, 29(2):87–106.

Coxe, S., West, S. G., and Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment*, 91(2):121–136.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Currie, K., Knussen, C., Price, L., and Reilly, J. (2014). Methicillin-resistant staphylococcus aureus screening as a patient safety initiative: using patients' experiences to improve the quality of screening practices. *Journal of clinical nursing*, 23(1-2):221–231.

Curtis, C. J. and Simpson, G. L. (2014). Trends in bulk deposition of acidity in the UK, 1988–2007, assessed using additive models. *Ecological Indicators*, 37:274–286.

Czajkowski, M., Gill, R., and Rempala, G. (2008). Model selection in logistic joinpoint regression with applications to analyzing cohort mortality patterns. *Statistics in medicine*, 27(9):1508–1526.

Daneman, N., Stukel, T. A., Ma, X., Vermeulen, M., and Guttmann, A. (2012). Reduction in clostridium difficile infection rates after mandatory hospital public reporting: findings from a longitudinal cohort study in Canada. *PLoS medicine*, 9(7):e1001268.

Daniel, M., Booth, M., Ellis, K., Maher, S., and Longmate, A. (2015). Details behind the dots: How different intensive care units used common and contrasting methods to prevent ventilator associated pneumonia. *BMJ quality improvement reports*, 4(1):u207660–w3069.

Dathe, E. and Müller, P. (1980). A contribution to spline-regression. *Biometrical Journal*, 22(3):259–269.

Davison, A. and Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review/Revue Internationale de Statistique*, 60(3):337–353.

Davison, A. C. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472.

Di Ruscio, F., Bjørnholt, J. V., Leegaard, T. M., Moen, A. E. F., and De Blasio, B. F. (2017). MRSA infections in Norway: A study of the temporal evolution, 2006-2015. *PloS one*, 12(6):e0179771.

Dianat, R. and Kasaei, S. (2010). Change detection in optical remote sensing images using difference-based methods and spatial information. *IEEE Geoscience and Remote Sensing Letters*, 7(1):215–219.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, 11(3):189–212.

Diniz, C. A. R., Milan, L. A., and Mazucheli, J. (2003). Bayesian inference of a linear segmented regression model. *Brazilian Journal of Probability and Statistics*, 17(1):1–16.

Dobson, A. J. and Barnett, A. (2011). *An introduction to generalized linear models*. CRC press.

Dover, D. C. and Schopflocher, D. P. (2011). Using funnel plots in public health surveillance. *Population Health Metrics*, 9(1):58–69.

Draper, N. R., Smith, H., and Pownell, E. (1966). *Applied regression analysis*, volume 3. Wiley New York.

Dukić, V., Hayden, M., Forgor, A. A., Hopson, T., Akweongo, P., Hodgson, A., Monaghan, A., Wiedinmyer, C., Yoksas, T., Thomson, M. C., et al. (2012). The role of weather in meningitis outbreaks in Navrongo, Ghana: a generalized additive modeling approach. *Journal of agricultural, biological, and environmental statistics*, 17(3):442–460.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, 37:409–428.

Eayres, D. (2008). Technical briefing 3: Commonly used public health statistics and their confidence intervals. `http://www.sld.cu/galerias/pdf/sitios/revsalud/usos_comunes_de_la_estad_ultimo.pdf`. (Accessed: 30-1-2014).

ECDPC (2012). Antimicrobial resistance surveillance in Europe. Available from: European Centre for Disease Prevention and Control. `https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/antimicrobial-resistance-surveillance-europe-2012.pdf`. (Accessed: 30-6-2017).

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics*, 9(2):139–158.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ekkelenkamp, M. B. (2011). *Staphylococcus aureus and healthcare-associated infections*. PhD thesis, Utrecht University.

Elliott, G. and Müller, U. K. (2007). Confidence sets for the date of a single break in linear time series regressions. *Journal of Econometrics*, 141(2):1196–1218.

Engemann, J. J., Carmeli, Y., Cosgrove, S. E., Fowler, V. G., Bronstein, M. Z., Trivette, S. L., Briggs, J. P., Sexton, D. J., and Kaye, K. S. (2003). Adverse clinical and economic outcomes attributable to methicillin resistance among patients with staphylococcus aureus surgical site infection. *Clinical infectious diseases*, 36(5):592–598.

Engemann, J. J., Friedman, J. Y., Reed, S. D., Griffiths, R. I., Szczech, L. A., Kaye, K. S., Stryjewski, M. E., Reller, L. B., Schulman, K. A., Corey, G. R., et al. (2005). Clinical outcomes and costs due to staphylococcus aureus bacteremia among patients receiving long-term hemodialysis. *Infection Control and Hospital Epidemiology*, 26(6):534–539.

Eo, Y. and Morley, J. (2015). Likelihood-ratio-based confidence sets for the timing of structural breaks. *Quantitative Economics*, 6(2):463–497.

Fay, M. P., Kim, H.-J., and Hachey, M. (2007). On using truncated sequential probability ratio test boundaries for monte carlo implementation of hypothesis tests. *Journal of Computational and Graphical Statistics*, 16(4):946–967.

Feder, P. I. (1975). On asymptotic distribution theory in segmented regression problems- identified case. *The Annals of Statistics*, 3(1):49–83.

Fernández, E., González, J., Borras, J., Moreno, V., Sánchez, V., and Peris, M. (2001). Recent decline in cancer mortality in Catalonia (Spain). A joinpoint regression analysis. *European Journal of Cancer*, 37(17):2222–2228.

Ferraro, M. B., Coppi, R., and González-Rodríguez, G. (2013). Bootstrap confidence intervals for the parameters of a linear regression model with fuzzy random variables. In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, pages 33–42. Springer.

Ferreira, P. E. (1975). A bayesian analysis of a switching regression model: known number of regimes. *Journal of the American Statistical Association*, 70(350):370–374.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 175–185.

Filice, G. A., Nyman, J. A., Lexau, C., Lees, C. H., Bockstedt, L. A., Como-Sabetti, K., Lesher, L. J., and Lynfield, R. (2010). Excess costs and utilization associated with methicillin resistance for patients with staphylococcus aureus infection. *Infection Control and Hospital Epidemiology*, 31(4):365–373.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Private Limited.

Fletcher, D. (2008). Confidence intervals for the mean of the delta-lognormal distribution. *Environmental and Ecological Statistics*, 15(2):175–189.

Fong, Y., Di, C., and Permar, S. (2015). Change point testing in logistic regression models with interaction term. *Statistics in medicine*, 34(9):1483–1494.

Freeman, J. M. (2010). Inference for binomial change point data. In *Advances in Data Analysis*, pages 345–352. Springer.

Frome, E. L. (1983). The analysis of rates using Poisson regression models. *Biometrics*, 39(3):665–674.

Furuya-Kanamori, L., Clements, A., Foster, N., Huber, C., Hong, S., Harris-Brown, T., Yakob, L., Paterson, D., and Riley, T. (2017). Asymptomatic clostridium difficile colonization in two Australian tertiary hospitals, 2012–2014: prospective, repeated cross-sectional study. *Clinical microbiology and infection*, 23(1):48.e1–48.e7.

Gail, M. and Benichou, J. (2000). *Encyclopedia of Epidemiologic Methods*. Wiley Reference Series in Biostatistics. Wiley.

Gardner, W., Mulvey, E. P., and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin*, 118(3):392–404.

Gerding, D. N., Meyer, T., Lee, C., Cohen, S. H., Murthy, U. K., Poirier, A., Van Schooneveld, T. C., Pardi, D. S., Ramos, A., Barron, M. A., et al. (2015). Administration of spores of nontoxigenic clostridium difficile strain m3 for prevention of recurrent c difficile infection: a randomized clinical trial. *Jama*, 313(17):1719–1727.

Ghorbanzadeh, D., Durand, P., and Jaupi, L. (2016). Problem of change-point detection for the Poisson observations. In *Proceedings of the World Congress on Engineering*, volume 1.

Ghosh, P., Basu, S., and Tiwari, R. C. (2009a). Bayesian analysis of cancer rates from seer program using parametric and semiparametric joinpoint regression models. *Journal of the American Statistical Association*, 104(486):439–452.

Ghosh, P., Huang, L., Yu, B., and Tiwari, R. C. (2009b). Semiparametric bayesian approaches to joinpoint regression for population-based cancer survival data. *Computational statistics and data analysis*, 53(12):4073–4082.

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science and Business Media.

Gu, K., Ng, H. K. T., Man, L. T., and Schucany, W. R. (2008). Testing the ratio of two Poisson rates. *Biometrical Journal*, 50:283–298.

Guisan, A., Edwards, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2):89–100.

Gurevich, G. and Vexler, A. (2005). Change point problems in the model of logistic regression. *Journal of Statistical Planning and Inference*, 131(2):313–331.

Gurmu, S. (1998). Generalized hurdle count data regression models. *Economics Letters*, 58(3):263–268.

Habibi, R., Sadooghi-Alvandi, S., and Nematollahi, A. (2005). Change point detection in a general class of distributions. *Communications in Statistics-Theory and Methods*, 34(9-10):1935–1938.

Halinski, R. S. and Feldt, L. S. (1970). The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 7(3):151–157.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, 20(2):695–711.

Han, S. W., Zhong, H., and Putt, M. (2015). An efficient operator for the change point estimation in partial spline model. *Communications in Statistics-Simulation and Computation*, 44(5):1171–1186.

Härdle, W., Huet, S., and Jolivet, E. (1995). Better bootstrap confidence intervals for regression curve estimation. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):287–306.

Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20(2):265–300.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Heinze, G., Ploner, M., and Beyea, J. (2013). Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. *Statistics in medicine*, 32(29):5062–5076.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56(3):635–639.

Hinkley, D. V. (1971). Inference in two-phase regression. *Journal of the American Statistical Association*, 66(336):736–743.

HIS (2008). Healthcare associated infection (HAI) standards: 2008. Available from: Healthcare improvment Scotland. `http://www.healthcareimprovementscotland.org/previous_resources/standards/healthcare_associated_infectio.aspx`. (Accessed: 5-1-2015).

HIS (2009). Healthcare environment inspectorate. Available from: Healthcare improvment Scotland. `http://www.healthcareimprovementscotland.org/programmes/inspecting_and_regulating_care/environment_inspectorate_hei.aspx`. (Accessed: 5-1-2015).

Hole, A. R. (2007). A comparison of approaches to estimating confidence intervals for willingness to pay measures. *Health economics*, 16(8):827–840.

HPS (2006). HEAT targets due for delivery. Available from: Scottish government. `http://www.gov.scot/Resource/0046/00467027.pdf`. (Accessed: 5-1-2015).

HPS (2008a). Model infection control policies on transmission based precaution. Available from: Health Protection Scotland. `http://www.hps.scot.nhs.uk/news/newsdetail.aspx?id=20103`. (Accessed: 30-6-2017).

HPS (2008b). Quarterly report on staphylococcus aureus bacteraemias in Scotland from January 2003 - December 2007. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/haiic/sshaip/documents/mrsa-quarterly-reports/jan-2003-to-dec-2007.pdf. (Accessed: 27-2-2014).

HPS (2009). Hospital infection incident assessment tool. Available from: Health protection Scotland. http://www.hps.scot.nhs.uk/resourcedocument.aspx?id=1476. (Accessed: 30-6-2017).

HPS (2012). Protocol for the Scottish mandatory surveillance programme for staphylococcus aureus bacteraemia. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/resourcedocument.aspx?id=577. (Accessed: 30-9-2014).

HPS (2013). Health Protection Scotland. http://www.hps.scot.nhs.uk/. (Accessed: 5-1-2014).

HPS (2014). Quarterly report on the surveillance of staphylococcus aureus bacteraemias in Scotland, October- December 2013. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/ewr/article.aspx. (Accessed: 10-7-2017).

HPS (2015a). Commentary on quarterly epidemiological data on clostridium difficile infection (CDI) in Scotland, July to September (Q3) 2014. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/ewr/article.aspx. (Accessed: 10-7-2017).

HPS (2015b). National hand hygiene campaign. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/haiic/ic/nationalhandhygienecampaign.aspx. (Accessed: 7-4-2014).

HPS (2015c). Quarterly report methods and caveats for the surveillance of staphylococcus aureus (s. aureus) bacteraemia infection in Scotland. Available from: Health protection Scotland. http://www.hps.scot.nhs.uk/resourcedocument.aspx?resourceid=2064. (Accessed: 05-10-2017).

HPS (2016a). Commentary on quarterly epidemiological data on clostridium difficile infection, staphylococcus aureus bacteraemias and surgical site infection in Scotland. April to June (Q2) 2016. Available from: Health Protection Scotland. http://www.hps.scot.nhs.uk/pubs/detail.aspx?id=3084. (Accessed: 27-2-2017).

HPS (2016b). HIIAT green incidents and outbreak mandatory reporting. Available from: Health protection Scotland. http://www.hps.scot.nhs.uk/ewr/article.aspx. (Accessed: 10-7-2017).

HPS (2017). Healthcare associated infection annual report 2016. Available from: Health protection Scotland. http://www.hps.scot.nhs.uk/resourcedocument.aspx?id=5934. (Accessed: 05-10-2017).

Hu, M.-C., Pavlicova, M., and Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. *The American journal of drug and alcohol abuse*, 37(5):367–375.

Huang, C.-Y. and Lyu, M. R. (2011). Estimation and analysis of some generalized multiple change-point software reliability models. *IEEE Transactions on reliability*, 60(2):498–514.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *Journal of the american statistical association*, 61(316):1097–1129.

Huet, S., Denis, J.-B., and Adamczyk, K. (1999). Bootstrap confidence intervals in nonlinear regression models when the number of observations is fixed

and the variance tends to 0. application to biadditive models. *Statistics: A Journal of Theoretical and Applied Statistics*, 32(3):203–227.

Huh, J. (2012). Nonparametric estimation of the regression function having a change point in generalized linear models. *Statistics and Probability Letters*, 82(4):843–851.

Hušková, M. and Kirch, C. (2008). Bootstrapping confidence intervals for the change-point of time series. *Journal of Time Series Analysis*, 29(6):947–972.

Hušková, M. and Kirch, C. (2010). A note on studentized confidence intervals for the change-point. *Computational Statistics*, 25(2):269–289.

Hutchinson, M. K. and Holtman, M. C. (2005). Analysis of count data using Poisson regression. *Research in nursing and health*, 28(5):408–418.

ICT, I. C. T. (2015). Transmission based precautions literature review: Definitions of transmission based precautions. Available from: Health Protection Scotland. `http://www.nipcm.hps.scot.nhs.uk/documents/tbp-definitions-of-transmission-based-precautions/`. (Accessed: 30-6-2017).

Ilic, M. and Ilic, I. (2016). Prostate cancer mortality in Serbia, 1991–2010: A joinpoint regression analysis. *Journal of Public Health*, 38(2):e63–e67.

Ilic, M. and Ilic, I. (2017). Diabetes mortality in Serbia, 1991–2015 (a nationwide study): A joinpoint regression analysis. *Primary Care Diabetes*, 11(1):78–85.

ISD (2005). Scottish intensive care society. Available from: Information Services Division. `http://www.sicsag.scot.nhs.uk/docs/SICSAG_Report2005_06.pdf`. (Accessed: 5-1-2015).

ISD (2014). Information services division. `http://www.isdscotland.org/`. (Accessed: 30-2-2014).

ISMR (2006). Burden of methicillin-resistant staphylococcus aureus on healthcare cost and resource utilization. Available from: Inernational society of microbial resestant. http://www.microresistance.org/docs/ISMR-Burden-of-MRSA-on-Healthcare-Cost.pdf. (Accessed: 20-9-2016).

Jahren, T., Storaas, T., Willebrand, T., Moa, P. F., and Hagen, B.-R. (2016). Declining reproductive output in capercaillie and black grouse–16 countries and 80 years. *Animal Biology*, 66(3-4):363–400.

Jandhyala, V. and Minogue, C. (1993). Distributions of bayes-type change-point statistics under polynomial regression. *Journal of statistical planning and inference*, 37(3):271–290.

Jiang, R. (2012). Determination of degradation change point using spline function. In *Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2012 International Conference*, pages 900–904. IEEE.

Jones, R. H. and Dey, I. (1995). Determining one or more change points. *Chemistry and Physics of LIPIDS*, 76(1):1–6.

Julious, S. A. (2001). Inference and estimation in a changepoint regression problem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1):51–61.

Karlsson, A. (2009). Bootstrap methods for bias correction and confidence interval estimation for nonlinear quantile regression of longitudinal data. *Journal of Statistical Computation and Simulation*, 79(10):1205–1218.

Kastenbaum, M. A. (1959). 137. note: A confidence interval on the abscissa of the point of intersection of two fitted linear regressions. *Biometrics*, 15(2):323–324.

Katikireddi, S. V., Der, G., Roberts, C., and Haw, S. (2016). Has childhood smoking reduced following smoke-free public places legislation? a segmen-

ted regression analysis of cross-sectional UK school-based surveys. *Nicotine and Tobacco Research*, 18(7):1670–1674.

Kavanagh, K., Pan, J., Marwick, C., Davey, P., Wiuff, C., Bryson, S., Robertson, C., and Bennie, M. (2017). Cumulative and temporal associations between antimicrobial prescribing and community-associated clostridium difficile infection: population-based case–control study using administrative data. *Journal of Antimicrobial Chemotherapy*, 72(4):1193–1201.

Keeping, E. S. (1962). *Introduction to statistical inference*, volume 26. Courier Corporation.

Kihoro, J., Mundia, S., and Gichuhi, A. (2017). The power of likelihood ratio test for a change-point in binomial distribution. *Journal of Agricultural Science and Technology*, pages 105–132.

Kim, H.-J. (1994). Tests for a change-point in linear regression. *Lecture Notes-Monograph Series*, 23:170–176.

Kim, H.-J. (1996). Change-point detection for correlated observations. *Statistica Sinica*, 6(1):275–287.

Kim, H.-J. and Cai, L. (1993). Robustness of the likelihood ratio test for a change in simple linear regression. *Journal of the American Statistical Association*, 88(423):864–871.

Kim, H.-J., Fay, M. P., Feuer, E. J., Midthune, D. N., et al. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in medicine*, 19(3):335–351.

Kim, H.-J., Fay, M. P., Yu, B., Barrett, M. J., and Feuer, E. J. (2004). Comparability of segmented line regression models. *Biometrics*, 60(4):1005–1014.

Kim, H.-J., Luo, J., Kim, J., Chen, H.-S., and Feuer, E. J. (2014). Clustering of trend data using joinpoint regression models. *Statistics in medicine*, 33(23):4087–4103.

Kim, H.-J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, 76(3):409–423.

Kim, H.-J., Yu, B., and Feuer, E. J. (2008). Inference in segmented line regression: a simulation study. *Journal of Statistical Computation and Simulation*, 78(11):1087–1103.

Kim, H.-J., Yu, B., and Feuer, E. J. (2009). Selecting the number of change-points in segmented line regression. *Statistica Sinica*, 19(2):597–609.

Kim, J. and Kim, H.-J. (2008). Asymptotic results in segmented multiple regression. *Journal of Multivariate Analysis*, 99(9):2016–2038.

Kim, J. and Kim, H.-J. (2016). Consistent model selection in segmented line regression. *Journal of statistical planning and inference*, 170:106–116.

Kim, W., Linton, O. B., and Hengartner, N. W. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, 8(2):278–297.

Kinoshita, T., Tokumasu, H., Tanaka, S., Kramer, A., and Kawakami, K. (2017). Policy implementation for methicillin-resistant staphylococcus aureus in seven European countries: a comparative analysis from 1999 to 2015. *Journal of Market Access and Health Policy*, 5(1):1351293.

Kligienė, N. (1977). On the estimation of the change point in the autoregressive sequence. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*, pages 325–334. Springer.

Kociolek, L. K., Bovee, M., Carter, D., Ciolino, J. D., Patel, R., O'Donnell, A., Rupp, A. H., Zheng, X., Shulman, S. T., and Patel, S. J. (2017). Impact of a healthcare provider educational intervention on frequency of clostridium difficile polymerase chain reaction testing in children: a segmented regression analysis. *Journal of the Pediatric Infectious Diseases Society*, 6(2):142–148.

Lalani, T., Chu, V. H., Grussemeyer, C. A., Reed, S. D., Bolognesi, M. P., Friedman, J. Y., Griffiths, R. I., Crosslin, D. R., Kanafani, Z. A., Kaye, K. S., et al. (2008). Clinical outcomes and costs among patients with staphylococcus aureus bacteremia and orthopedic device infections. *Scandinavian journal of infectious diseases*, 40(11-12):973–977.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Land, K. C., McCall, P. L., and Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models with empirical applications to criminal careers data. *Sociological Methods and Research*, 24(4):387–442.

Lawes, T., Edwards, B., López-Lozano, J.-M., and Gould, I. (2012). Trends in staphylococcus aureus bacteraemia and impacts of infection control practices including universal MRSA admission screening in a hospital in Scotland, 2006–2010: retrospective cohort study and time series intervention analysis. *BMJ open*, 2(3):e000797.

Lawes, T., López-Lozano, J.-M., Nebot, C., Macartney, G., Subbarao-Sharma, R., Dare, C. R., Edwards, G. F., and Gould, I. M. (2015). Turning the tide or riding the waves? impacts of antibiotic stewardship and infection control on MRSA strain dynamics in a Scottish region over 16 years: non-linear time series analysis. *BMJ open*, 5(3):e006596.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lee, P. H. (2016). Examining non-linear associations between accelerometer-measured physical activity, sedentary behavior, and all-cause mortality using segmented cox regression. *Frontiers in physiology*, 7:272.

Lerman, P. (1980). Fitting segmented regression models by grid search. *Applied Statistics*, 29(1):77–84.

Lessa, F. C., Mu, Y., Bamberg, W. M., Beldavs, Z. G., Dumyati, G. K., Dunn, J. R., Farley, M. M., Holzbauer, S. M., Meek, J. I., Phipps, E. C., et al. (2015). Burden of clostridium difficile infection in the United States. *New England Journal of Medicine*, 372(9):825–834.

Lindberg, M. (2012). *Methicillin-resistant Staphylococcus aureus (MRSA) an Unclear and Untoward Issue: Patient-Professional Interactions, Experiences, Attitudes and Responsibility*. PhD thesis, Acta Universitatis Upsaliensis.

Liu, C., Zhang, X., and Wan, J. (2015). Public reporting influences antibiotic and injection prescription in primary care: a segmented regression analysis. *Journal of evaluation in clinical practice*, 21(4):597–603.

Liu, H. (2008). Generalized additive model. *Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN*, 55812.

Liu, J., Wu, S., and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, 7(2):497–525.

Liu, R. Y. et al. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.

Liu, Z. and Qian, L. (2009). Changepoint estimation in a segmented linear regression via empirical likelihood. *Communications in Statistics-Simulation and Computation*, 39(1):85–100.

Loader, C. R. (1992). A log-linear model for a Poisson process change point. *The Annals of Statistics*, 20(3):1391–1411.

Ludden, T. M., Beal, S. L., and Sheiner, L. B. (1994). Comparison of the Akaike information criterion, the Schwarz criterion and the F-test as guides to model selection. *Journal of pharmacokinetics and biopharmaceutics*, 22(5):431–445.

Luwel, K., Beem, A. L., Onghena, P., and Verschaffel, L. (2001). Using segmented linear regression models with unknown change points to analyze strategy shifts in cognitive tasks. *Behavior Research Methods, Instruments and Computers*, 33(4):470–478.

Lyazrhi, F. (1997). Bayesian criteria for discriminating among regression models with one possible change point. *Journal of statistical planning and inference*, 59(2):337–353.

MacNeill, I. B. (1978). Properties of sequences of partial sums of polynomial regression residuals with applications to tests for change of regression at unknown times. *The Annals of Statistics*, 6(2):422–433.

Marchaim, D., Kaye, K., Fowler, V., Anderson, D., Chawla, V., Golan, Y., Karchmer, A., and Carmeli, Y. (2010). Case–control study to identify factors associated with mortality among patients with methicillin-resistant staphylococcus aureus bacteraemia. *Clinical Microbiology and Infection*, 16(6):747–752.

Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74.

Martin, P. (2015). Alcohol based handrubs and infection control. http://www.sehd.scot.nhs.uk/publications/cno(2005)01.pdf. (Accessed: 5-1-2015).

Martinez-Beneito, M. A., García-Donato, G., and Salmerón, D. (2011). A bayesian joinpoint regression model with an unknown number of break-points. *The Annals of Applied Statistics*, 5(3):2150–2168.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

McCord, J., Prewitt, M., Dyakova, E., Mookerjee, S., and Otter, J. (2016). Reduction in clostridium difficile infection associated with the introduction of hydrogen peroxide vapour automated room disinfection. *Journal of Hospital Infection*, 94(2):185–187.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

Mian, M. R. I. (2016). *Analysis of Zero Inflated Over dispersed Count Data Regression Models with Missing Values*. PhD thesis, University of Windsor.

Min, S. and Park, T. (2016). Bayesian variable selection in Poisson change-point regression analysis. *Communications in Statistics-Simulation and Computation*, 46(3):2267–2282.

Mitchell, B. G., Digney, W., Locket, P., and Dancer, S. J. (2014). Controlling methicillin-resistant staphylococcus aureus (MRSA) in a hospital and the role of hydrogen peroxide decontamination: an interrupted time series analysis. *BMJ open*, 4(4):e004522.

Mohammadi, G., Akbari, M. E., Mehrabi, Y., Motlagh, A. G., Heidari, M., Ghanbari, S., et al. (2016). Analysis of cancer incidence and mortality in

Iran using joinpoint regression analysis. *Iranian Red Crescent Medical Journal*, 19(3):e42071.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. New York, NY:McGraw-Hill.

Morton, A., Mengersen, K., Rajmokan, M., Whitby, M., Playford, E., and Jones, M. (2011). Funnel plots and risk-adjusted count data adverse events. a limitation of indirect standardisation. *Journal of Hospital Infection*, 78(4):260–263.

Moxnes, J. F., Moen, A. E. F., and Leegaard, T. M. (2015). Studying the time trend of methicillin-resistant staphylococcus aureus (MRSA) in Norway by use of non-stationary $\gamma$-poisson distributions. *BMJ open*, 5(10):e007163.

Muggeo, V. M., Attanasio, M., and Porcu, M. (2009). A segmented regression model for event history data: an application to the fertility patterns in Italy. *Journal of Applied Statistics*, 36(9):973–988.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365.

Müller, H.-G. and Wang, J.-L. (1990). Nonparametric analysis of changes in hazard rates for censored survival data: An alternative to change-point models. *Biometrika*, 77(2):305–314.

Naber, C. K. (2009). Staphylococcus aureus bacteremia: epidemiology, pathophysiology, and management strategies. *Clinical infectious diseases*, 48(Supplement 4):S231–S237.

Natarajan, R. and Pednault, E. (2002). Segmented regression estimators for massive data sets. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 566–582. SIAM.

NCISR (2017). Joinpoint regression program, statistical methodology and applications branch, surveillance research program, national cancer institute, version 4.5.0.1. https://surveillance.cancer.gov/joinpoint/. (Accessed: 13-6-2017).

Neumark, S. (2014). *Solution of cubic and quartic equations*. Elsevier.

Newitt, S., Myles, P. R., Birkin, J., Maskell, V., Slack, R., Nguyen-Van-Tam, J., and Szatkowski, L. (2015). Impact of infection control interventions on rates of staphylococcus aureus bacteraemia in National Health Service acute hospitals, East Midlands, UK, using interrupted time series analysis. *Journal of Hospital Infection*, 90(1):28–37.

Ng, H. K. T. and Tang, M. L. (2005). Testing the equality of two Poisson means using the rate ratio. *Statistics in Medicine*, 24(6):955–965.

NHS (2016a). National Health Service. http://www.nhs.uk/pages/home.aspx. (Accessed: 20-9-2016).

NHS (2016b). Who's most at risk of c. difficile? Available from: National Health Service. http://www.nhs.uk/conditions/Clostridium-difficile/Pages/Introduction.aspx#risk-factors. (Accessed: 15-10-2016).

Nickalls, R. (1993). A new approach to solving the cubic: Cardan's solution revealed. *The Mathematical Gazette*, 77(480):354–359.

O'Brien, S. M. (2004). Cutpoint selection for categorizing a continuous predictor. *Biometrics*, 60(2):504–509.

Oliveira, N. F. d., Santana, V. S., and Lopes, A. A. (1997). Ratio of proportions and the use of the delta method for confidence interval estimation in logistic regression. *Revista de Saúde Pública*, 31(1):90–99.

Oman Evans II, M., Starley, B., Galagan, J. C., Yabes, J. M., Evans, S., and Salama, J. J. (2016). Tea and recurrent clostridium difficile infection. *Gastroenterology Research and Practice*, 2016:1–5.

Orváth, L. and Kokoszka, P. (2002). Change-point detection with non-parametric regression. *Statistics: A Journal of Theoretical and Applied Statistics*, 36(1):9–31.

Özgen, C. (2008). *Antibiotic resistant staphylococcus aureus infection studies in hospitals*. PhD thesis, Middle east technical university.

Page, E. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Park, H. M. (2008). Hypothesis testing and statistical power of a test. *The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University.* `https://scholarworks.iu.edu/dspace/handle/2022/19738`. (Accessed: 20-5-2017).

Park, T., Krafty, R. T., and Sánchez, A. I. (2012). Bayesian semi-parametric analysis of Poisson change-point regression models: application to policy-making in Cali, Colombia. *Journal of applied statistics*, 39(10):2285–2298.

Pastor-Barriuso, R., Guallar, E., and Coresh, J. (2003). Transition models for change-point estimation in logistic regression. *Statistics in medicine*, 22(7):1141–1162.

Perovic, O., Iyaloo, S., Kularatne, R., Lowman, W., Bosman, N., Wadula, J., Seetharam, S., Duse, A., Mbelle, N., Bamford, C., et al. (2015). Prevalence and trends of staphylococcus aureus bacteraemia in hospitalized patients in South Africa, 2010 to 2012: laboratory-based surveillance mapping of antimicrobial resistance and molecular epidemiology. *PloS one*, 10(12):e0145429.

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, 28(2):126–135.

Phatharacharukul, P., Thongprayoon, C., Cheungpasitporn, W., Edmonds, P. J., Mahaparn, P., and Bruminhent, J. (2015). The risks of incident and recurrent clostridium difficile-associated diarrhea in chronic kidney disease and end-stage kidney disease patients: a systematic review and meta-analysis. *Digestive diseases and sciences*, 60(10):2913–2922.

Piegorsch, W. W. (1982). *A modification of the least squares join point estimator in bilinear segmented regression*. Cornell University, May.

Piegorsch, W. W. et al. (1982). Confidence intervals on the join point in segmented regression. *Biometrics Unit, Cornell University, Ithaca, NY, USA*.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986.

Pinlac, P. A. V., Silawan, T., Tempongko, M. S. B., Tolabing, M. C. C., and Soonthornworasiri, N. (2016). Interrupted time series analysis using segmented regression of premature mortality from noncommunicable disease among Filipinos. *Southest Asian Journal of Tropical Medicine and Public Health*, 47(4):810–821.

Plata, K., Rosato, A. E., Wegrzyn, G., et al. (2009). Staphylococcus aureus as an infectious agent: overview of biochemistry and molecular genetics of its pathogenicity. *Acta Biochimica Polonica*, 56(4):597–612.

Pradhan, V. and Banerjee, T. (2008). Confidence interval of the difference of two independent binomial proportions using weighted profile likelihood. *Communications in Statistics-Simulation and Computation®*, 37(4):645–659.

Pradhan, V., Menon, S., and Das, U. (2013). Corrected profile likelihood confidence interval for binomial paired incomplete data. *Pharmaceutical statistics*, 12(1):48–58.

Procopio, M. and Marriott, P. K. (1998). Seasonality of birth in epilepsy: a Danish study. *Acta neurologica scandinavica*, 98(5):297–301.

Qi, J.-P., Zhang, Q., Zhu, Y., and Qi, J. (2014). A novel method for fast change-point detection on simulated time series and electrocardiogram data. *PloS one*, 9(4):e93365.

Qiu, D., Katanoda, K., Marugame, T., and Sobue, T. (2009). A joinpoint regression analysis of long-term trends in cancer mortality in Japan (1958–2004). *International journal of cancer*, 124(2):443–448.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the american statistical association*, 53(284):873–880.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raza, S. (2011). Healthcare associated infections. Available from: Scottish Parliament Information Centre `http://www.parliament.scot/ResearchBriefingsAndFactsheets/Factsheets/SB_11-80.pdf`. (Accessed: 06-10-2017).

RCN, R. C. N. (2005). Methicillin-resistant staphylococcus aureus (MRSA). Available from: National Health Service. `http://www.nhs.uk/conditions/mrsa/documents/rcn%20mrsa%20guidelines.pdf`. (Accessed: 5-1-2015).

Read, C. B. (2003). Fieller's theorem. *Encyclopedia of Statistical Sciences*, 3:86–88.

Reed, S. D., Friedman, J. Y., Engemann, J. J., Griffiths, R. I., Anstrom, K. J., Kaye, K. S., Stryjewski, M. E., Szczech, L. A., Reller, L. B., Corey, G. R., et al. (2005). Costs and outcomes among hemodialysis-dependent patients with methicillin-resistant or methicillin-susceptible staphylococcus aureus bacteremia. *Infection Control and Hospital Epidemiology*, 26(2):175–183.

Reil, M., Hensgens, M., Kuijper, E., Jakobiak, T., Gruber, H., Kist, M., and Borgmann, S. (2012). Seasonality of clostridium difficile infections in Southern Germany. *Epidemiology and infection*, 140(10):1787.

Robotham, J. V., Deeny, S. R., Fuller, C., Hopkins, S., Cookson, B., and Stone, S. (2016). Cost-effectiveness of national mandatory screening of all admissions to English National Health Service hospitals for meticillin-resistant staphylococcus aureus: a mathematical modelling study. *The Lancet Infectious Diseases*, 16(3):348–356.

Rodriguez-Palacios, A., Reid-Smith, R. J., Staempfli, H. R., Daignault, D., Janecko, N., Avery, B. P., Martin, H., Thomspon, A. D., McDonald, L. C., Limbago, B., et al. (2009). Possible seasonality of clostridium difficile in retail meat, Canada. *Emerging infectious diseases*, 15(5):802–805.

Roser, C. and Nakano, M. (2002). Single simulation confidence intervals using the delta method. In *International Symposium on Scheduling*.

Rothmann, M., Li, N., Chen, G., Chi, G. Y., Temple, R., and Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in medicine*, 22(2):239–264.

Rothmann, M. D. and Tsou, H.-H. (2003). On non-inferiority analysis based on delta-method confidence intervals. *Journal of biopharmaceutical statistics*, 13(3):565–583.

Royston, P. et al. (2007). Profile likelihood for estimation and confidence intervals. *Stata Journal*, 7(3):376–387.

Rubin, R. J., Harrington, C. A., Poon, A., Dietrich, K., Greene, J. A., and Moiduddin, A. (1999). The economic impact of staphylococcus aureus infection in New York city hospitals. *Emerging infectious diseases*, 5(1):9–17.

Saffaria, S. E., Adnana, R., Greeneb, W., and Ahmada, M. H. (2013). A Poisson regression model for analysis of censored count data with excess zeroes. *Journal Teknologi (Sciences and Engineering)*, 63(2):71–74.

Saha, K. K., Sen, D., and Jin, C. (2012). Profile likelihood-based confidence interval for the dispersion parameter in count data. *Journal of Applied Statistics*, 39(4):765–783.

Sanjay, P., Fawzi, A., Kulli, C., Polignano, F. M., and Tait, I. S. (2010). Impact of methicillin-resistant staphylococcus aureus (MRSA) infection on patient outcome after pancreatoduodenectomy (pd)-a cause for concern? *Pancreas*, 39(8):1211–1214.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott Long, J. (1997). *Regression models for categorical and limited dependent variables*, volume 7. Sage Publications.

SE (2005). Antimicrobial prescribing policy and practice in Scotland. Available from: Scottish Executive. `http://www.sehd.scot.nhs.uk/cmo/CMO(2005)8report.pdf`. (Accessed: 5-1-2015).

Seyoum, A. and Zewotir, T. (2016). Quasi-Poisson versus negative binomial regression models in identifying factors affecting initial CD4 cell count change due to antiretroviral therapy administered to HIV-positive adults in North–West Ethiopia (Amhara region). *AIDS research and therapy*, 13(1):36–46.

SGHAI (2008). Scottish government healthcare associated infection (HAI) task force report on delivery programme: 2008-2011. Available from: Scottish Government. `http://www.gov.scot/Publications/2011/03/09143800/6`. (Accessed: 15-6-2016).

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3):591–611.

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review/Revue Internationale de Statistique*, 56(1):31–48.

Simpsona, G. L. and Anderson, N. (2009). Deciphering the effect of climate change and separating the influence of confounding factors in sediment core records using additive models. *Limnology and Oceanography*, 54(6part2):2529–2541.

Solberg, C. O. (2000). Spread of staphylococcus aureus in hospitals: causes and prevention. *Scandinavian journal of infectious diseases*, 32(6):587–595.

Solis-Trapala, I. L. and Farewell, V. T. (2005). Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in medicine*, 24(16):2557–2575.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8):1185–1202.

SPSP (2007). Scottish patient safty programme. Available from: Healthcare improvment Scotland. `http://www.healthcareimprovementscotland.org/our_work/patient_safety/spsp.aspx`. (Accessed: 5-1-2015).

SPSP (2008). Improving the safety and reliability of care across NHSScotland. `http://www.scottishpatientsafetyprogramme.scot.nhs.uk/programmes`. (Accessed: 27-6-2015).

Stasinopoulos, D. and Rigby, R. (1992). Detecting break points in generalised linear models. *Computational Statistics and Data Analysis*, 13(4):461–471.

Steward, R. M., Rigdon, S. E., and Pan, R. (2016). A bayesian approach to diagnostics for multivariate control charts. *Journal of Quality Technology*, 48(4):303–325.

Stone, S. P., Fuller, C., Savage, J., Cookson, B., Hayward, A., Cooper, B., Duckworth, G., Michie, S., Murray, M., Jeanes, A., et al. (2012). Evaluation of the national cleanyourhands campaign to reduce staphylococcus aureus bacteraemia and clostridium difficile infection in hospitals in England and Wales by improved hand hygiene: four year, prospective, ecological, interrupted time series study. *Bmj*, 344(1):e3005.

Storr, J., Twyman, A., Zingg, W., Damani, N., Kilpatrick, C., Reilly, J., Price, L., Egger, M., Grayson, M. L., Kelley, E., et al. (2017). Core components for effective infection prevention and control programmes: new who evidence-based recommendations. *Antimicrobial Resistance & Infection Control*, 6(1):6.

Stracci, F., Canosa, A., Minelli, L., Petrinelli, A. M., Cassetti, T., Romagnoli, C., and La Rosa, F. (2007). Cancer mortality trends in the Umbria region of Italy 1978–2004: A joinpoint regression analysis. *BMC cancer*, 7(1):10–19.

Stryhn, H. and Christensen, J. (2003). Confidence intervals by the profile likelihood method, with applications in veterinary epidemiology. In *Proceedings of the 10th International Symposium on Veterinary Epidemiology and Economics, Vina del Mar*, page 208.

Tang, Y.-c. and Fei, H.-l. (2004). Detecting change points in polynomial regression models with an application to cable data sets. *Acta Mathematicae Applicatae Sinica*, 20(4):541–546.

Taylor, W. A. (2000). Change-point analysis: a powerful new tool for detecting changes. Pre-print, `http://www.variation.com/cpa/tech/changepoint.html`. (Accessed: 30-6-2017).

Thode, H. C. (2002). *Testing for normality*, volume 164. CRC press.

Thomas, G. (2014). WHO's first global report on antibiotic resistance reveals serious, worldwide threat to public health. Available from: World Health organization. `http://www.who.int/mediacentre/news/releases/2014/amr-report/en/`. (Accessed: 05-10-2017).

Thwaites, G. E., Edgeworth, J. D., Gkrania-Klotsas, E., Kirby, A., Tilley, R., Török, M. E., Walker, S., Wertheim, H. F., Wilson, P., Llewelyn, M. J., et al. (2011). Clinical management of staphylococcus aureus bacteraemia. *The Lancet infectious diseases*, 11(3):208–222.

Tiwari, R. C., Cronin, K. A., Davis, W., Feuer, E. J., Yu, B., and Chib, S. (2005). Bayesian model selection for join point regression with application to age-adjusted cancer rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(5):919–939.

Tong, E., Clements, A., Haynes, M., Jones, M., Morton, A., and Whitby, M. (2009). Improved hospital-level risk adjustment for surveillance of healthcare-associated bloodstream infections: a retrospective cohort study. *BMC infectious diseases*, 9(1):145–153.

Tzeng, I.-S., Liu, S.-H., Chen, K.-F., Wu, C.-C., and Chen, J.-C. (2016). Impact of performance grading on annual numbers of acute myocardial infarction-associated emergency department visits in Taiwan: Results of segmented regression analysis. *Medicine*, 95(42):e4937.

Van Hal, S. J., Jensen, S. O., Vaska, V. L., Espedido, B. A., Paterson, D. L., and Gosbell, I. B. (2012). Predictors of mortality in staphylococcus aureus bacteremia. *Clinical microbiology reviews*, 25(2):362–386.

Van Velzen, E., Reilly, J., Kavanagh, K., Leanord, A., Edwards, G., Girvan, E., Gould, I., Mackenzie, F., and Masterton, R. (2011). A retrospective cohort

study into acquisition of MRSA and associated risk factors after implementation of universal screening in Scottish hospitals. *Infection Control and Hospital Epidemiology*, 32(9):889–896.

Venzon, D. and Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):87–94.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114(1):106–115.

Vexler, A. and Gurevich, G. (2009). Average most powerful tests for a segmented regression. *Communications in Statistics-Theory and Methods*, 38(13):2214–2231.

Vonberg, R.-P., Kuijper, E., Wilcox, M., Barbut, F., Tüll, P., Gastmeier, P., Van Den Broek, P., Colville, A., Coignard, B., Daha, T., et al. (2008). Infection control measures to limit the spread of clostridium difficile. *Clinical Microbiology and Infection*, 14(s5):2–20.

Vos, P. W. and Hudson, S. (2003). Simulation study of conditional, bootstrap, and t confidence intervals in linear regression. *Communications in Statistics-Simulation and Computation*, 32(3):697–715.

Wagner, A. K., Soumerai, S. B., Zhang, F., and Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309.

Wahrendorf, J., Becher, H., and Brown, C. C. (1987). Bootstrap comparison of non-nested generalized linear models: applications in survival analysis and epidemiology. *Applied statistics*, 36(1):72–81.

Walkowiak, R. and Kala, R. (2000). Two-phase nonlinear regression with smooth transition. *Communications in Statistics-Simulation and Computation*, 29(2):385–397.

Wertheim, H. F., Vos, M. C., Ott, A., van Belkum, A., Voss, A., Kluytmans, J. A., van Keulen, P. H., Vandenbroucke-Grauls, C. M., Meester, M. H., and Verbrugh, H. A. (2004). Risk and outcome of nosocomial staphylococcus aureus bacteraemia in nasal carriers versus non-carriers. *The Lancet*, 364(9435):703–705.

Wertheim, H. F. L. (2005). *Staphylococcus aureus infections; Lead by the nose*. PhD thesis, Erasmus MC: University Medical Center Rotterdam.

Western, B. and Kleykamp, M. (2004). A bayesian change point model for historical time series analysis. *Political Analysis*, 12(4):354–374.

Winter, B. and Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution*, 1(1):7–18.

Wold, S. (1974). Spline functions in data analysis. *Technometrics*, 16(1):1–11.

Won, K. S. C. and Park, E.-C. (2010). Long-term trends in cancer mortality in Korea (1983-2007): A joinpoint regression analysis. *Asian Pacific Journal of Cancer Prevention*, 11:1451–1457.

Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.

Wood, S. and Wood, M. S. (2017). Package 'mgcv'. *R package version*, pages 1–7.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38(2):89–104.

Worsley, K. (1983). Testing for a two-phase multiple regression. *Technometrics*, 25(1):35–42.

Worth, L., Spelman, T., Bull, A., Brett, J., and Richards, M. (2016). Epidemiology of clostridium difficile infections in Australia: enhanced surveillance

to evaluate time trends and severity of illness in Victoria, 2010–2014. *Journal of Hospital Infection*, 93(3):280–285.

Wu, W.-H. and Hsieh, H.-N. (2014). Generalized confidence interval estimation for the mean of delta-lognormal distribution: an application to New Zealand trawl survey data. *Journal of Applied Statistics*, 41(7):1471–1485.

Wu*, Z. and Tian, Y. (2005). Weighted-loss-function CUSUM chart for monitoring mean and variance of a production process. *International Journal of Production Research*, 43(14):3027–3044.

Xu, J. and Long, J. S. (2005). Using the delta method to construct confidence intervals for predicted probabilities, rates, and discrete changes. Manuscript, `http://www.indiana.edu/~jslsoc/stata/ci_computations/spost_deltaci.pdf`. (Accessed: 30-6-2017).

Xu, J., Long, J. S., et al. (2005). Confidence intervals for predicted outcomes in regression models for categorical outcomes. *Stata Journal*, 5(4):537.

Yu, B., Barrett, M. J., Kim, H.-J., and Feuer, E. J. (2007). Estimating joinpoints in continuous time scale for multiple change-point models. *Computational Statistics and Data Analysis*, 51(5):2420–2427.

Zand, A., Yazdanshenas, N., and Amiri, A. (2013). Change point estimation in phase I monitoring of logistic regression profile. *The International Journal of Advanced Manufacturing Technology*, 67(9-12):2301–2311.

Zhang, F., Wagner, A., Soumerai, S. B., and Ross-Degnan, D. (2002). Estimating confidence intervals around relative changes in outcomes in segmented regression analyses of time series data. In *NESUG 15th Annual Conference. Buffalo, NY*.

Zhang, F., Wagner, A. K., Soumerai, S. B., and Ross-Degnan, D. (2009). Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *Journal of clinical epidemiology*, 62(2):143–148.

Zhang, S., Palazuelos-Munoz, S., Balsells, E. M., Nair, H., Chit, A., and Kyaw, M. H. (2016). Cost of hospital management of clostridium difficile infection in United States a meta-analysis and modelling study. *BMC Infectious Diseases*, 16(1):447–465.

Zhou, H., Liang, K.-Y., et al. (2008). On estimating the change point in generalized linear models. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, pages 305–320. Institute of Mathematical Statistics.

Zou, C., Liu, Y., Qin, P., and Wang, Z. (2007). Empirical likelihood ratio test for the change-point problem. *Statistics and probability letters*, 77(4):374–382.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York ; London : Springer.

# Appendices

# Appendix A

# Modelling Count Data - Chapter 3

## A.1 R code for modelling rate

```
# To read xlsx files
library(XLConnect)

# Need for dispersiontest
library(AER)

library(lmtest)

d1 <- readWorksheetFromFile("joinpointDATA-uptoSep2014.xlsx",
sheet=1, header=T)

d1$t <- d1$time-2003
d1$t2 <- d1$t^2
d1$t3 <- d1$t^3
d1$t4 <- d1$t^4
d1$t5 <- d1$t^5
# To make factors
d1$Qu <- factor(d1$Qu)
d1$HB <- factor(d1$HB)

d1m1 <- glm(no.MRSA1 ~ offset(log(aobd))+t+t2+t3+t4+Qu ,
family=poisson,data=d1)
summary(d1m1)
dispersiontest(d1m1)

d1m2 <- glm(no.MRSA1 ~ offset(log(aobd))+t+t2+t3+Qu ,
family=poisson,data=d1)

lrtest(d1m1,d1m2)

### Check autocorrelation of the residuals of fitted model
# Construct Pearson residuals from the model
z.res <- residuals(d1m1, type="pearson")
z.res <- z.res-mean(z.res)
```

```
plot(density(z.res)) # density function of residuals

# Test autocorrelation for residual
dw<- dwtest(z.res ~ 1, alt="two.sided")

# Autocorrelation function
ac <- acf(z.res, type = "correlation")$acf
print(acf(z.res,plot=F,lag.max=2),digits=4)
ro.ac <- ac[2] # autocorrelation value at lag1

# Fisher's z-transformation of (ro.ac)
x<- (1+ro.ac)/(1-ro.ac)
z.ro<- 0.5 * log(x, base = exp(1))
length(d1$no.MRSA1)
seg.z<- 1/sqrt(length(d1$no.MRSA1) - 3) #standard error
lcl.z<- z.ro-1.96*seg.z
ucl.z<- z.ro+1.96*seg.z

# Confidence interval of (ro.ac)
lcl.ro<- ((exp(2*lcl.z))-1)/((exp(2*lcl.z))+1)
ucl.ro<- ((exp(2*ucl.z))-1)/((exp(2*ucl.z))+1)

##########################

# To get the predicted values
d1.p.MRSA1 <- predict(d1m1,type="response" ,
interval = "confidence" ,level = 0.95, se=T)
d1.p.MRSA1.rate <- d1.p.MRSA1$fit/d1$aobd*100000

# 95% CI for predicted rates using Byar's method
lp.MRSA1 <- d1.p.MRSA1$fit * (1-(1/(9*d1.p.MRSA1$fit))-
(1.96/(3*sqrt(d1.p.MRSA1$fit))))^3
lrp.MRSA1 <- lp.MRSA1/d1$aobd*100000 # lower CI
up.MRSA1 <- (1+d1.p.MRSA1$fit) * (1-(1/(9*(1+d1.p.MRSA1$fit)))+
(1.96/(3*sqrt(1+d1.p.MRSA1$fit))))^3
urp.MRSA1 <- up.MRSA1/d1$aobd*100000 # upper CI
```

## A.2 Modelling HAIs by health boards

```
# Fitting model with health boards.
m1<- glm(no.MRSA1 ~ offset(log(aobd))+t+t2+t3+t4+Qu +
HB +t*HB + Qu*HB  , family=quasipoisson,data=d1)
```

### A.2.1 The coefficients of MRSA bacteraemia model with HBs effect

Table A.1: The coefficients of MRSA bacteraemia model with HBs effect.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -8.4660 | 0.0699 | -121.0403 | 0.0000* |
| t | -0.0464 | 0.0734 | -0.6327 | 0.5272 |
| t2 | 0.0682 | 0.0306 | 2.2283 | 0.0263* |
| t3 | -0.0191 | 0.0046 | -4.1067 | 0.0000* |
| t4 | 0.0011 | 0.0002 | 4.7327 | 0.0000* |
| Qu2 | -0.1439 | 0.0678 | -2.1233 | 0.0342* |
| Qu3 | -0.2382 | 0.0706 | -3.3746 | 0.0008* |
| Qu4 | -0.0282 | 0.0670 | -0.4208 | 0.6741 |
| HB11LO | 0.1046 | 0.0917 | 1.1410 | 0.2543 |
| HB12NWTC | 3.5576 | 4.3258 | 0.8224 | 0.4112 |
| HB13ORK | -18.2676 | 1158.8934 | -0.0158 | 0.9874 |
| HB14SH | -0.5906 | 0.7574 | -0.7798 | 0.4358 |
| HB15TAY | -0.0565 | 0.1190 | -0.4749 | 0.6350 |
| HB16WI | -1.1783 | 0.5929 | -1.9873 | 0.0474* |
| HB1LA | -0.0453 | 0.1155 | -0.3927 | 0.6947 |
| HB2A.A | -0.7334 | 0.1626 | -4.5099 | 0.0000* |
| HB3BOR | -0.9131 | 0.2634 | -3.4658 | 0.0006* |
| HB4DG | -0.0931 | 0.2191 | -0.4249 | 0.6711 |
| HB5Fife | -0.1480 | 0.1496 | -0.9893 | 0.3229 |
| HB6FV | -0.4449 | 0.1702 | -2.6149 | 0.0092* |
| HB7GR | -0.6777 | 0.1310 | -5.1746 | 0.0000* |
| HB9HI | -1.0260 | 0.1969 | -5.2098 | 0.0000* |
| t:HB11LO | 0.0047 | 0.0154 | 0.3016 | 0.7630 |
| t:HB12NWTC | -0.5704 | 0.5512 | -1.0347 | 0.3012 |
| t:HB13ORK | 0.3175 | 0.1744 | 1.8208 | 0.0692 |
| t:HB14SH | 0.1506 | 0.1240 | 1.2148 | 0.2249 |
| t:HB15TAY | -0.0069 | 0.0197 | -0.3530 | 0.7242 |
| t:HB16WI | 0.0882 | 0.1000 | 0.8823 | 0.3780 |
| t:HB1LA | 0.0235 | 0.0201 | 1.1655 | 0.2443 |
| t:HB2A.A | 0.0637 | 0.0264 | 2.4071 | 0.0164* |
| t:HB3BOR | 0.1947 | 0.0411 | 4.7372 | 0.0000* |
| t:HB4DG | 0.0019 | 0.0391 | 0.0483 | 0.9615 |
| t:HB5Fife | 0.0312 | 0.0239 | 1.3032 | 0.1931 |
| t:HB6FV | 0.0862 | 0.0258 | 3.3476 | 0.0009* |
| t:HB7GR | 0.0617 | 0.0216 | 2.8490 | 0.0045* |
| t:HB9HI | 0.0941 | 0.0321 | 2.9339 | 0.0035* |
| Qu2:HB11LO | 0.1829 | 0.1082 | 1.6905 | 0.0915 |
| Qu3:HB11LO | 0.2288 | 0.1115 | 2.0515 | 0.0407* |
| Qu4:HB11LO | 0.0515 | 0.1090 | 0.4729 | 0.6365 |

*Continued on next page*

Table A.1 – *Continued from previous page*

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Qu2:HB12NWTC | 0.0513 | 1.4492 | 0.0354 | 0.9718 |
| Qu3:HB12NWTC | -15.4523 | 1357.0891 | -0.0114 | 0.9909 |
| Qu4:HB12NWTC | 0.4112 | 1.3953 | 0.2947 | 0.7684 |
| Qu2:HB13ORK | 0.1650 | 1640.8310 | 0.0001 | 0.9999 |
| Qu3:HB13ORK | 15.0964 | 1158.8936 | 0.0130 | 0.9896 |
| Qu4:HB13ORK | 16.2395 | 1158.8932 | 0.0140 | 0.9888 |
| Qu2:HB14SH | -0.3405 | 0.8318 | -0.4093 | 0.6825 |
| Qu3:HB14SH | -1.2965 | 1.2455 | -1.0409 | 0.2984 |
| Qu4:HB14SH | -1.5362 | 1.2472 | -1.2317 | 0.2186 |
| Qu2:HB15TAY | 0.4293 | 0.1369 | 3.1371 | 0.0018* |
| Qu3:HB15TAY | 0.4132 | 0.1423 | 2.9031 | 0.0038* |
| Qu4:HB15TAY | 0.1111 | 0.1439 | 0.7721 | 0.4404 |
| Qu2:HB16WI | -0.9228 | 0.9288 | -0.9935 | 0.3209 |
| Qu3:HB16WI | 0.1780 | 0.6915 | 0.2574 | 0.7970 |
| Qu4:HB16WI | 0.1093 | 0.6600 | 0.1656 | 0.8685 |
| Qu2:HB1LA | -0.1699 | 0.1412 | -1.2029 | 0.2295 |
| Qu3:HB1LA | -0.1590 | 0.1466 | -1.0845 | 0.2786 |
| Qu4:HB1LA | -0.2164 | 0.1401 | -1.5443 | 0.1231 |
| Qu2:HB2A.A | 0.2435 | 0.1818 | 1.3392 | 0.1810 |
| Qu3:HB2A.A | -0.2748 | 0.2139 | -1.2845 | 0.1995 |
| Qu4:HB2A.A | -0.0835 | 0.1916 | -0.4360 | 0.6630 |
| Qu2:HB3BOR | -0.4560 | 0.3074 | -1.4834 | 0.1385 |
| Qu3:HB3BOR | -0.4924 | 0.3286 | -1.4983 | 0.1346 |
| Qu4:HB3BOR | -0.2336 | 0.2812 | -0.8305 | 0.4066 |
| Qu2:HB4DG | -0.1409 | 0.2622 | -0.5373 | 0.5912 |
| Qu3:HB4DG | -0.4070 | 0.2955 | -1.3771 | 0.1690 |
| Qu4:HB4DG | -0.3669 | 0.2747 | -1.3358 | 0.1822 |
| Qu2:HB5Fife | 0.2987 | 0.1706 | 1.7512 | 0.0805 |
| Qu3:HB5Fife | 0.1480 | 0.1831 | 0.8082 | 0.4193 |
| Qu4:HB5Fife | 0.1888 | 0.1725 | 1.0949 | 0.2740 |
| Qu2:HB6FV | 0.0768 | 0.2012 | 0.3816 | 0.7029 |
| Qu3:HB6FV | 0.3505 | 0.1957 | 1.7913 | 0.0738 |
| Qu4:HB6FV | 0.2062 | 0.1921 | 1.0733 | 0.2836 |
| Qu2:HB7GR | 0.0176 | 0.1541 | 0.1143 | 0.9090 |
| Qu3:HB7GR | -0.0051 | 0.1606 | -0.0319 | 0.9746 |
| Qu4:HB7GR | -0.0527 | 0.1540 | -0.3422 | 0.7323 |
| Qu2:HB9HI | -0.1186 | 0.2330 | -0.5088 | 0.6111 |
| Qu3:HB9HI | -0.0204 | 0.2354 | -0.0868 | 0.9308 |
| Qu4:HB9HI | -0.1839 | 0.2333 | -0.7884 | 0.4308 |

* : Significant coefficient at $\alpha = 0.05$.

### A.2.2 The coefficients of linear model with HBs effect of MSSA bacteraemia

Table A.2: The coefficients of MSSA bacteraemia model with HBs effect.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -7.9789 | 0.0392 | -203.7174 | 0.0000* |
| t | -0.0414 | 0.0074 | -5.5931 | 0.0000* |
| Qu2 | -0.0154 | 0.0293 | -0.5250 | 0.5999 |
| Qu3 | 0.0773 | 0.0288 | 2.6814 | 0.0076* |
| Qu4 | 0.0157 | 0.0292 | 0.5367 | 0.5917 |
| HB11LO | -0.0444 | 0.0590 | -0.7522 | 0.4523 |
| HB12NWTC | 2.0463 | 0.6924 | 2.9552 | 0.0033* |
| HB13ORK | -0.8452 | 0.4832 | -1.7493 | 0.0809 |
| HB14SH | 0.0059 | 0.4033 | 0.0145 | 0.9884 |
| HB15TAY | -0.1883 | 0.0763 | -2.4687 | 0.0139* |
| HB16WI | -2.2810 | 0.5661 | -4.0291 | 0.0001* |
| HB1LA | -0.1782 | 0.0734 | -2.4271 | 0.0156* |
| HB2A.A | -0.0781 | 0.0808 | -0.9656 | 0.3348 |
| HB3BOR | -0.9036 | 0.1739 | -5.1951 | 0.0000* |
| HB4DG | 0.0555 | 0.1229 | 0.4515 | 0.6518 |
| HB5Fife | 0.0678 | 0.0843 | 0.8042 | 0.4217 |
| HB6FV | -0.0713 | 0.0987 | -0.7219 | 0.4707 |
| HB7GR | -0.2632 | 0.0705 | -3.7342 | 0.0002* |
| HB9HI | -0.6979 | 0.1095 | -6.3741 | 0.0000* |
| t:HB11LO | 0.0177 | 0.0124 | 1.4202 | 0.1562 |
| t:HB12NWTC | -0.3625 | 0.1218 | -2.9761 | 0.0031* |
| t:HB13ORK | 0.0259 | 0.1021 | 0.2532 | 0.8002 |
| t:HB14SH | -0.0442 | 0.0871 | -0.5078 | 0.6118 |
| t:HB15TAY | 0.0463 | 0.0157 | 2.9487 | 0.0033* |
| t:HB16WI | 0.2376 | 0.1000 | 2.3766 | 0.0179* |
| t:HB1LA | 0.0303 | 0.0153 | 1.9867 | 0.0475* |
| t:HB2A.A | 0.0114 | 0.0172 | 0.6625 | 0.5080 |
| t:HB3BOR | 0.1444 | 0.0340 | 4.2521 | 0.0000* |
| t:HB4DG | -0.0233 | 0.0265 | -0.8780 | 0.3804 |
| t:HB5Fife | 0.0517 | 0.0171 | 3.0240 | 0.0026* |
| t:HB6FV | 0.0399 | 0.0201 | 1.9813 | 0.0481* |
| t:HB7GR | 0.0404 | 0.0147 | 2.7461 | 0.0063* |
| t:HB9HI | 0.0658 | 0.0224 | 2.9320 | 0.0035* |

* : Significant coefficient at $\alpha = 0.05$.

### A.2.3 The coefficients of CDI in patients over 65 years model with HBs effect

Table A.3: The coefficients of CDI in patients over 65 years model with HBs effect.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -6.7518 | 0.0699 | -96.6054 | 0.0000* |
| t | 0.8038 | 0.1445 | 5.5627 | 0.0000* |
| t2 | -0.6716 | 0.0875 | -7.6754 | 0.0000* |
| t3 | 0.1307 | 0.0186 | 7.0097 | 0.0000* |
| t4 | -0.0078 | 0.0013 | -6.1428 | 0.0000* |
| Qu2 | -0.0053 | 0.0291 | -0.1811 | 0.8564 |
| Qu3 | -0.0122 | 0.0301 | -0.4055 | 0.6853 |
| Qu4 | -0.0786 | 0.0293 | -2.6836 | 0.0076* |
| HB11LO | 0.3046 | 0.1066 | 2.8579 | 0.0045* |
| HB12NWTC | 3.6747 | 15.9515 | 0.2304 | 0.8179 |
| HB13ORK | -1.2555 | 0.9657 | -1.3001 | 0.1943 |
| HB14SH | -0.4402 | 0.7183 | -0.6129 | 0.5403 |
| HB15TAY | 0.1611 | 0.1326 | 1.2145 | 0.2253 |
| HB16WI | 0.1020 | 0.4692 | 0.2173 | 0.8281 |
| HB1LA | 0.2669 | 0.1193 | 2.2370 | 0.0258* |
| HB2A.A | 0.3095 | 0.1325 | 2.3365 | 0.0200* |
| HB3BOR | -1.3524 | 0.3751 | -3.6056 | 0.0004* |
| HB4DG | 0.1709 | 0.2148 | 0.7956 | 0.4267 |
| HB5Fife | -0.2211 | 0.1710 | -1.2930 | 0.1968 |
| HB6FV | -0.0939 | 0.1765 | -0.5318 | 0.5952 |
| HB7GR | -0.3504 | 0.1414 | -2.4786 | 0.0136* |
| HB9HI | 0.1292 | 0.1810 | 0.7135 | 0.4759 |
| t:HB11LO | -0.5127 | 0.2232 | -2.2972 | 0.0221* |
| t:HB12NWTC | -3.4911 | 17.6536 | -0.1978 | 0.8433 |
| t:HB13ORK | 1.3351 | 1.6395 | 0.8143 | 0.4159 |
| t:HB14SH | 0.4060 | 1.9316 | 0.2102 | 0.8336 |
| t:HB15TAY | -0.0422 | 0.2766 | -0.1527 | 0.8787 |
| t:HB16WI | -1.9513 | 1.0981 | -1.7769 | 0.0764 |
| t:HB1LA | 0.2535 | 0.2538 | 0.9989 | 0.3185 |
| t:HB2A.A | -0.3980 | 0.2703 | -1.4722 | 0.1418 |
| t:HB3BOR | 1.4851 | 0.6829 | 2.1748 | 0.0302* |
| t:HB4DG | -0.8115 | 0.4344 | -1.8682 | 0.0625 |
| t:HB5Fife | 0.7209 | 0.3544 | 2.0340 | 0.0426* |
| t:HB6FV | 0.6743 | 0.3731 | 1.8074 | 0.0715 |
| t:HB7GR | 0.6261 | 0.2793 | 2.2418 | 0.0255* |
| t:HB9HI | -0.7091 | 0.3879 | -1.8281 | 0.0683 |
| t2:HB11LO | 0.3511 | 0.1331 | 2.6382 | 0.0087* |
| t2:HB12NWTC | 0.9171 | 6.7938 | 0.1350 | 0.8927 |

*Continued on next page*

Table A.3 – *Continued from previous page*

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| t2:HB13ORK | -0.2513 | 0.8766 | -0.2867 | 0.7745 |
| t2:HB14SH | -1.2558 | 1.3628 | -0.9215 | 0.3573 |
| t2:HB15TAY | 0.0666 | 0.1658 | 0.4017 | 0.6881 |
| t2:HB16WI | 0.9452 | 0.6342 | 1.4904 | 0.1369 |
| t2:HB1LA | -0.1881 | 0.1542 | -1.2196 | 0.2233 |
| t2:HB2A.A | 0.4670 | 0.1597 | 2.9241 | 0.0037* |
| t2:HB3BOR | -0.6419 | 0.3800 | -1.6893 | 0.0920 |
| t2:HB4DG | 0.7815 | 0.2571 | 3.0398 | 0.0025* |
| t2:HB5Fife | -0.4106 | 0.2143 | -1.9166 | 0.0560 |
| t2:HB6FV | -0.2307 | 0.2326 | -0.9917 | 0.3219 |
| t2:HB7GR | 0.0564 | 0.1656 | 0.3405 | 0.7337 |
| t2:HB9HI | 0.4513 | 0.2334 | 1.9335 | 0.0539 |
| t3:HB11LO | -0.0733 | 0.0280 | -2.6153 | 0.0093* |
| t3:HB12NWTC | -0.0982 | 1.0837 | -0.0906 | 0.9278 |
| t3:HB13ORK | 0.0100 | 0.1760 | 0.0567 | 0.9548 |
| t3:HB14SH | 0.3443 | 0.2979 | 1.1558 | 0.2485 |
| t3:HB15TAY | -0.0142 | 0.0352 | -0.4025 | 0.6875 |
| t3:HB16WI | -0.1504 | 0.1281 | -1.1746 | 0.2409 |
| t3:HB1LA | 0.0441 | 0.0330 | 1.3377 | 0.1818 |
| t3:HB2A.A | -0.1114 | 0.0337 | -3.3049 | 0.0010* |
| t3:HB3BOR | 0.1191 | 0.0776 | 1.5346 | 0.1257 |
| t3:HB4DG | -0.1837 | 0.0545 | -3.3727 | 0.0008* |
| t3:HB5Fife | 0.0740 | 0.0456 | 1.6223 | 0.1055 |
| t3:HB6FV | 0.0039 | 0.0512 | 0.0770 | 0.9386 |
| t3:HB7GR | -0.0695 | 0.0353 | -1.9680 | 0.0498* |
| t3:HB9HI | -0.0965 | 0.0493 | -1.9577 | 0.0510* |
| t4:HB11LO | 0.0048 | 0.0019 | 2.5334 | 0.0117* |
| t4:HB12NWTC | 0.0030 | 0.0610 | 0.0494 | 0.9606 |
| t4:HB13ORK | 0.0004 | 0.0117 | 0.0372 | 0.9703 |
| t4:HB14SH | -0.0244 | 0.0199 | -1.2226 | 0.2222 |
| t4:HB15TAY | 0.0008 | 0.0024 | 0.3338 | 0.7387 |
| t4:HB16WI | 0.0081 | 0.0083 | 0.9777 | 0.3288 |
| t4:HB1LA | -0.0032 | 0.0023 | -1.4130 | 0.1584 |
| t4:HB2A.A | 0.0075 | 0.0023 | 3.2733 | 0.0012* |
| t4:HB3BOR | -0.0077 | 0.0052 | -1.4820 | 0.1391 |
| t4:HB4DG | 0.0125 | 0.0037 | 3.3768 | 0.0008* |
| t4:HB5Fife | -0.0042 | 0.0031 | -1.3523 | 0.1770 |
| t4:HB6FV | 0.0019 | 0.0036 | 0.5226 | 0.6016 |
| t4:HB7GR | 0.0069 | 0.0024 | 2.8689 | 0.0043* |
| t4:HB9HI | 0.0065 | 0.0033 | 1.9649 | 0.0501* |

* : Significant coefficient at $\alpha = 0.05$.

### A.2.4 The coefficients of CDI in patients aged 15-64 years model with HBs effect

Table A.4: The coefficients of CDI in patients aged 15-64 years model with HBs effect.

| Coefficient | Estimate | Standard error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -7.6324 | 0.0885 | -86.2632 | 0.0000* |
| t | -0.6527 | 0.0974 | -6.7033 | 0.0000* |
| t2 | 0.2024 | 0.0451 | 4.4848 | 0.0000* |
| t3 | -0.0203 | 0.0058 | -3.4846 | 0.0006* |
| Qu2 | 0.0524 | 0.0566 | 0.9270 | 0.3547 |
| Qu3 | 0.3146 | 0.0536 | 5.8666 | 0.0000* |
| Qu4 | 0.0953 | 0.0575 | 1.6570 | 0.0986 |
| HB11LO | 0.7359 | 0.0905 | 8.1315 | 0.0000* |
| HB12NWTC | -1.3805 | 0.6742 | -2.0475 | 0.0415* |
| HB13ORK | 1.2201 | 0.4793 | 2.5457 | 0.0114* |
| HB14SH | -5.0953 | 4.0014 | -1.2734 | 0.2039 |
| HB15TAY | 0.6638 | 0.1171 | 5.6702 | 0.0000* |
| HB16WI | 0.7721 | 0.3435 | 2.2478 | 0.0253* |
| HB1LA | 0.0227 | 0.1302 | 0.1745 | 0.8616 |
| HB2A.A | 1.0890 | 0.1111 | 9.8021 | 0.0000* |
| HB3BOR | 0.3107 | 0.2493 | 1.2464 | 0.2136 |
| HB4DG | 0.9322 | 0.1682 | 5.5413 | 0.0000* |
| HB5Fife | 0.0625 | 0.1779 | 0.3510 | 0.7258 |
| HB6FV | -1.2275 | 0.3182 | -3.8581 | 0.0001* |
| HB7GR | 0.8203 | 0.1049 | 7.8165 | 0.0000* |
| HB9HI | 0.2499 | 0.1627 | 1.5365 | 0.1255 |
| t:HB11LO | -0.0172 | 0.0325 | -0.5304 | 0.5963 |
| t:HB12NWTC | -0.1120 | 0.2542 | -0.4407 | 0.6598 |
| t:HB13ORK | 0.0617 | 0.1652 | 0.3734 | 0.7091 |
| t:HB14SH | 1.1111 | 0.8733 | 1.2724 | 0.2042 |
| t:HB15TAY | -0.0821 | 0.0436 | -1.8829 | 0.0607* |
| t:HB16WI | 0.1589 | 0.1135 | 1.3999 | 0.1626 |
| t:HB1LA | 0.0611 | 0.0450 | 1.3582 | 0.1754 |
| t:HB2A.A | -0.1772 | 0.0439 | -4.0323 | 0.0001* |
| t:HB3BOR | 0.0686 | 0.0887 | 0.7735 | 0.4398 |
| t:HB4DG | 0.0043 | 0.0597 | 0.0723 | 0.9424 |
| t:HB5Fife | -0.0379 | 0.0660 | -0.5739 | 0.5665 |
| t:HB6FV | 0.1964 | 0.0968 | 2.0291 | 0.0433* |
| t:HB7GR | -0.2297 | 0.0437 | -5.2601 | 0.0000* |
| t:HB9HI | 0.0412 | 0.0549 | 0.7504 | 0.4536 |

* : Significant coefficient at $\alpha = 0.05$.

## A.3 R code for power and sample size analysis

### A.3.1 Poisson ration test calculation

```
#Poisson test for MRSA in Scotland
x1 <- 985 # observed count in Scotland in 2005
x2 <- 965 # observed count in Scotland in 2006
t1 <- 5377870 # population in Scotland in 2005
t2 <- 5421267 # population in Scotland in 2006
d <- t2/t1
R <- 1
roo <- R/d
PRTtest <- (x1-x2*roo)/sqrt(x1+x2*(roo)^2)

# If PRTtest >= 1.65 (z_alpha),
# there is significant difference between two rates
pv <- 1-pnorm(PRTtest) # p-value

# Ratio of two rates in two years (2005, 2006) in Scotland
Rdash <- 18.3/17.8
```

### A.3.2 Statistical power calculation

```
#To calculate the power in Scotland and different boards
#using different effect sizes (Rdash)
R<-1
gama2005<- 0.00018312  # gama1
gama2006<- 0.000177758 # gama2
Rdash <- 18.3/17.8 # in case of MRSA (real effect size)
#Rdash<- seq(1,2.5,0.01) # sequence of Rdashs

c<- R/Rdash
z1_alpha<- 1.645

# Scotland
t2005<-5377870 # Population in Scotland in 2005  (t1)
t2006<- 5421267 # Population in Scotland in 2006 (t2)
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y1<-pnorm(fi, mean = 0, sd = 1, lower.tail = F, log.p = FALSE)

# Glasgow
t2005<-1560477
t2006<- 1588897
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y2<- pnorm(fi, mean = 0, sd = 1, lower.tail = F, log.p = FALSE)
```

```
# Lothian
t2005<-805606
t2006<- 810988
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y3<- pnorm(fi, mean = 0, sd = 1, lower.tail = F, log.p = FALSE)

#Grampian
t2005<-592996
t2006<- 598488
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y4<- pnorm(fi, mean = 0, sd = 1, lower.tail = F, log.p = FALSE)

#Tayside
t2005<-447053
t2006<- 449644
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y5<- pnorm(fi, mean = 0, sd = 1, lower.tail = F, log.p = FALSE)

#Fife
t2005<-260199
t2006<- 264053
d<-t2006/t2005
roo<- R/d
fi<- (z1_alpha*sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash)
-(((roo/c)-roo)*t2006*gama2005*Rdash))/
(sqrt(((roo/c)+roo^2)*t2006*gama2005*Rdash))
y6<- pnorm(fi, mean = 0, sd = 1, lower.tail =F, log.p = FALSE)

x11()
plot(Rdash,y1, type="l", xlab="Effect size" , ylab="Power", col="red",
 lty=1, lwd=2)

lines (Rdash,y2,col="blue" , lty=2, lwd=2)
lines (Rdash,y3, col="black", lty=3, lwd=2)
lines (Rdash,y4, col="purple" , lty=1, lwd=2)
lines (Rdash,y5,col="cyan4", lty=2, lwd=2)
lines (Rdash,y6, col="deeppink4" , lty=3, lwd=2)

legend(2,0.9,legend = c("Scotland","Glasgow", "Lothian", "Grampian",
"Tayside", "Fife"),
col=c("red","blue","black","purple", "cyan4", "deeppink4"),
lwd=2, lty=c(1,2,3,1,2,3))

#3# The detectable effect size with 80%, 90% and 95% power
R<-1
Rdash<- seq(1,3,0.1)
c<- R/Rdash
```

```
t2005<-5377870 # Population in Scotland in 2005
t2006<- 5421267 # Population in Scotland in 2006
d<-t2006/t2005
roo<- R/d

lamda1.80<- ((c/roo)+c^2)*(qnorm(0.95)+qnorm(0.80))^2/(1-c)^2
lamda1.90<- ((c/roo)+c^2)*(qnorm(0.95)+qnorm(0.90))^2/(1-c)^2
lamda1.95<- ((c/roo)+c^2)*(qnorm(0.95)+qnorm(0.95))^2/(1-c)^2

plot(Rdash,lamda1.80, type="l", xlab="Effect size" ,
ylab="Number of cases", lwd=2 )
lines (Rdash,lamda1.90,col="blue", lty=2, lwd=2 )
lines (Rdash,lamda1.95, col="red", lty=3, lwd=2 )

legend(2.4,1200,legend = c("80% power","90% power", "95% power"),
col=c("black","blue","red"),
lwd=2, lty=c(1,2,3))
```

## A.4 Risk and risk adjusted funnel plots

### A.4.1 R code for funnel plot

```
library(XLConnect)

# To read data from individual health boards
di <- readWorksheetFromFile("NHSNEWn1.xlsx",
sheet=i, header=T) # , i= 2,3,4,...,16

# aobds in individual health boards in Qu4, 2013.
aobds <- c(d2[44,]$aobd,d3[44,]$aobd,d4[44,]$aobd,d5[44,]$aobd,
d6[44,]$aobd,d7[44,]$aobd,d8[44,]$aobd,d9[44,]$aobd,d10[44,]$aobd,
d11[44,]$aobd,d12[19,]$aobd,d13[44,]$aobd,d14[44,]$aobd,
d15[44,]$aobd,d16[44,]$aobd)

N.MRSAs <- c(d2[44,]$no.MRSA2,d3[44,]$no.MRSA3,d4[44,]$no.MRSA4,
d5[44,]$no.MRSA5,d6[44,]$no.MRSA6,d7[44,]$no.MRSA7,d8[44,]$no.MRSA8,
d9[44,]$no.MRSA9,d10[44,]$no.MRSA10,d11[44,]$no.MRSA11,
d12[19,]$no.MRSA12,d13[44,]$no.MRSA13,d14[44,]$no.MRSA14,
d15[44,]$no.MRSA15,d16[44,]$no.MRSA16)

R.MRSAs <- c(d2[44,]$r.MRSA2,d3[44,]$r.MRSA3,d4[44,]$r.MRSA4,
d5[44,]$r.MRSA5,d6[44,]$r.MRSA6,d7[44,]$r.MRSA7,d8[44,]$r.MRSA8,
d9[44,]$r.MRSA9,d10[44,]$r.MRSA10,d11[44,]$r.MRSA11,
d12[19,]$r.MRSA12,d13[44,]$r.MRSA13,d14[44,]$r.MRSA14,
d15[44,]$r.MRSA15,d16[44,]$r.MRSA16)

ASP <- c(d2[44,]$ASP,d3[44,]$ASP,d4[44,]$ASP,d5[44,]$ASP,
d6[44,]$ASP,d7[44,]$ASP,d8[44,]$ASP,d9[44,]$ASP,d10[44,]$ASP,
d11[44,]$ASP,NA,d13[44,]$ASP,d14[44,]$ASP,
d15[44,]$ASP,d16[44,]$ASP)

TH <- c(0,0,0,0,0,1,1,0,0,1,0,0,0,1,0)

HB <- c("AA", "BOR", "DG", "Fife", "FV", "GR", "GGC", "HI",
"LA", "LO", "NWTC", "ORK", "SH", "TAY", "WI")

data <- cbind(aobds,N.MRSAs,R.MRSAs,ASP,TH)
d <- data.frame(data)
row.names(d) <- HB

d$TH <- factor(d$TH)
d$asp <- d$ASP-25

# To give subset date according to ASP
dsub <- subset(d, ! is.na(ASP))

# To sort data=d according to aobds
dsubb <- dsub[with(dsub, order(aobds)), ]

## Data including NWTC HB
www <- d[with(d, order(aobds)), ]

## To build 95% control limit using Byar's method
ob <- www$aobds*d1[44,]$r.MRSA1/100000
ol <- ob * (1-(1/(9*ob))-(1.96/(3*sqrt(ob))))^3
```

```
rl <- ol/www$aobds
ou <-(1+ob) * (1-(1/(9*(1+ob)))+(1.96/(3*sqrt(1+ob))))^3
ru <- ou/www$aobds

plot(www$aobds/100000,www$R.MRSAs , xlim=c(0,4.2),ylim=c(0,30),
col = "blue",pch = 12,
xlab="Acute occupied bed days (100,000)" ,ylab="Observed Rates of MRSA")
text(www$aobds/100000,www$R.MRSAs, row.names(www),
cex=0.6, pos=4, col = "blue")

# To plot line observes overall rate of MRSA
abline(h = d1[44,]$r.MRSA1, col = "red")

# To plot lower and upper limits of funnel plot
lines(www$aobds/100000, rl*100000)
lines(www$aobds/100000, ru*100000)

# To show components of funnel plot
arrows(4, 9.5, x1 = 4, y1 = 4.1, length = 0.15, angle = 30, col=3, lwd=2)
text(4, 10.5, "Target", cex = 1)

arrows(0.3, 28, x1 = -0.16, y1 = 28, length = 0.15, angle = 30, col=3, lwd=2)
text(0.6, 28, "Indicator", cex = 1)

arrows(3.5, 1, x1 = 3.5, y1 = -1.2, length = 0.15, angle = 30, col=3, lwd=2)
text(3.9, 0.5, "Precision", cex = 1)

arrows(2.5, 9.5, x1 = 2.5, y1 = 7.8, length = 0.15, angle = 30, col=3, lwd=2)
arrows(2.8, 9.5, x1 = 2.8, y1 = 2, length = 0.15, angle = 30, col=3, lwd=2)
text(2.6, 10.5, "Control limits", cex = 1)

## Use dsubb data to plot observed rates and
#use ds data to plot predict and observed rates

mod1 <- glm(N.MRSAs ~ offset (log(aobds))+asp,family=poisson,data=dsub)

expe <- predict(mod1, type="response" ,interval = "confidence",
level = 0.95, se=T)
R.adj <- dsub$N.MRSAs/expe$fit*d1[44,]$r.MRSA1
x <- cbind(dsub,R.adj)
xx <- data.frame(x)

# To sort data=d according to aobds.
ds <- xx[with(xx, order(aobds)), ]

# To plot predicted rates in the funnel plot
points(ds$aobds/100000,ds$R.adj, col = 'red', pch = 16)
text(ds$aobds/100000,ds$R.adj, row.names(ds), cex=0.6, pos=4, col="red")

# To add legend to the plot
legend(2,30,legend = c("Observed rate","Adjusted rate", "(P=0.186)"),
col = c("blue", "red",NA),pch =c(12,16,NA))
```

### A.4.2 Funnel plots of MRSA bacteraemia for different quarters

The following Figures show funnel plots of unadjusted MRSA bacteraemia rates and adjusted MRSA bacteraemia rates by acute surgical procedure (ASP) in different quarters from Qu2, 2009 to Qu4, 2013.

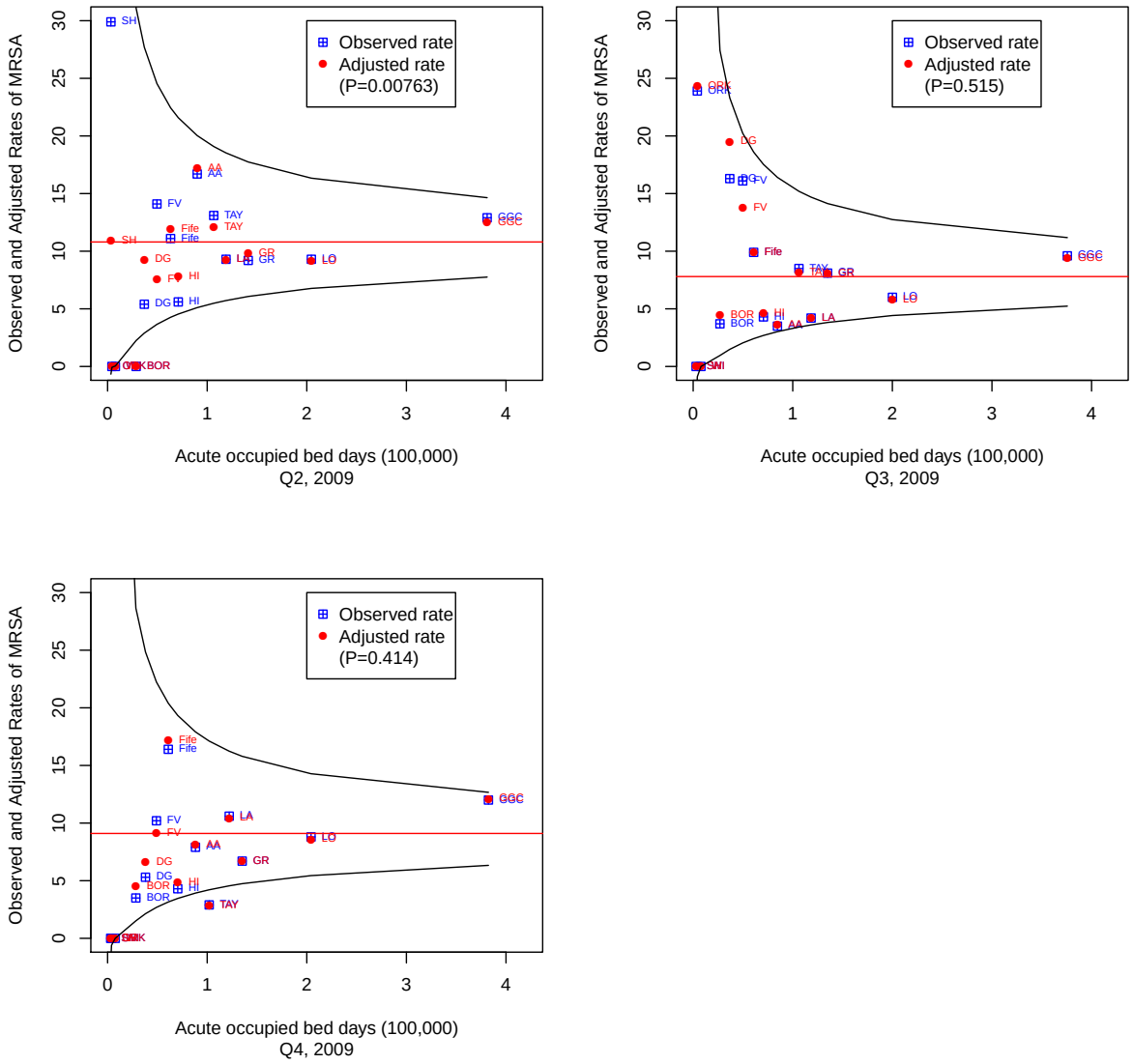



Figure A.1: Funnel plots of adjusted and unadjusted MRSA bacteraemia rates in 2009.

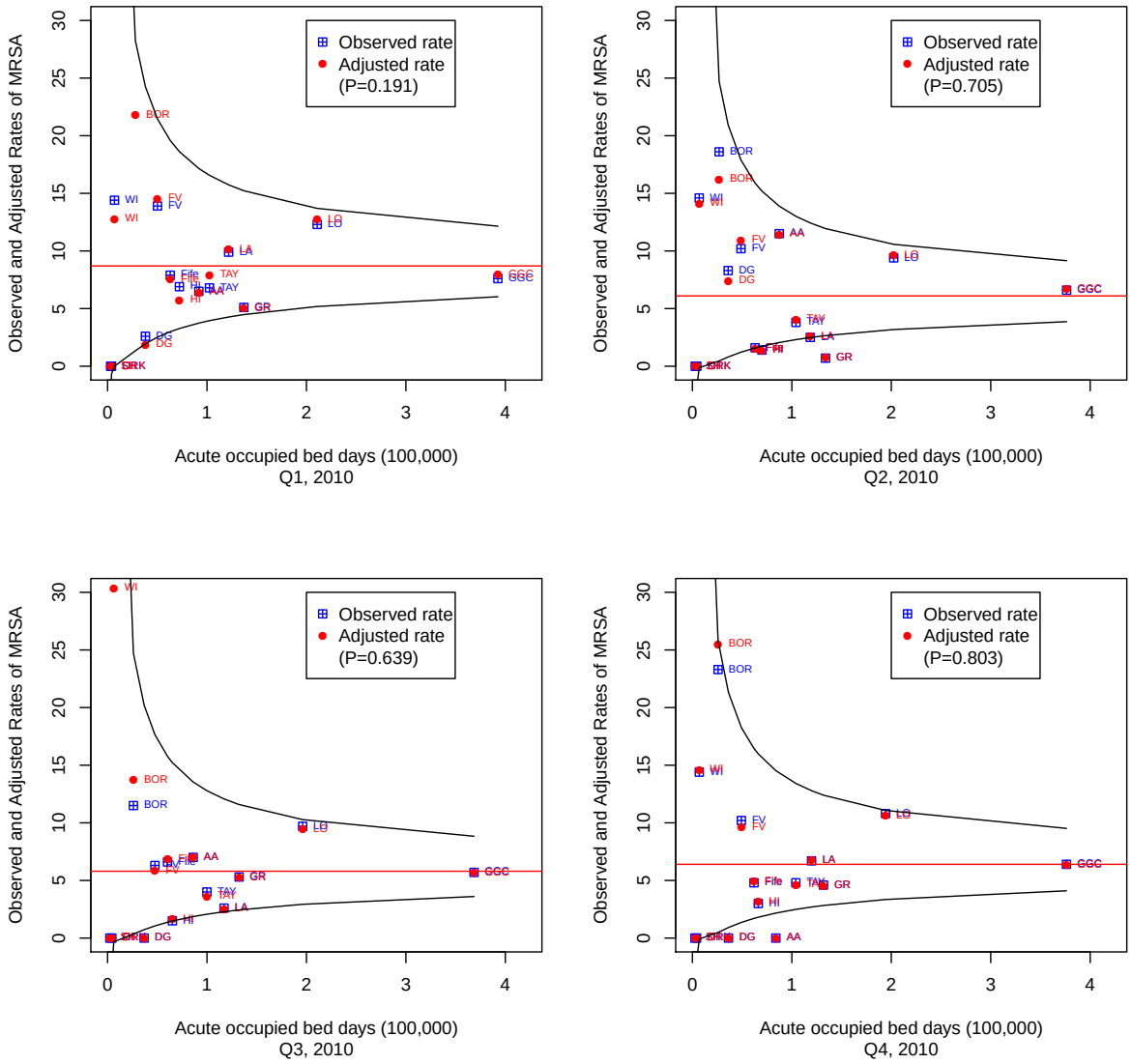



Figure A.2: Funnel plots of adjusted and unadjusted MRSA bacteraemia rates in 2010.

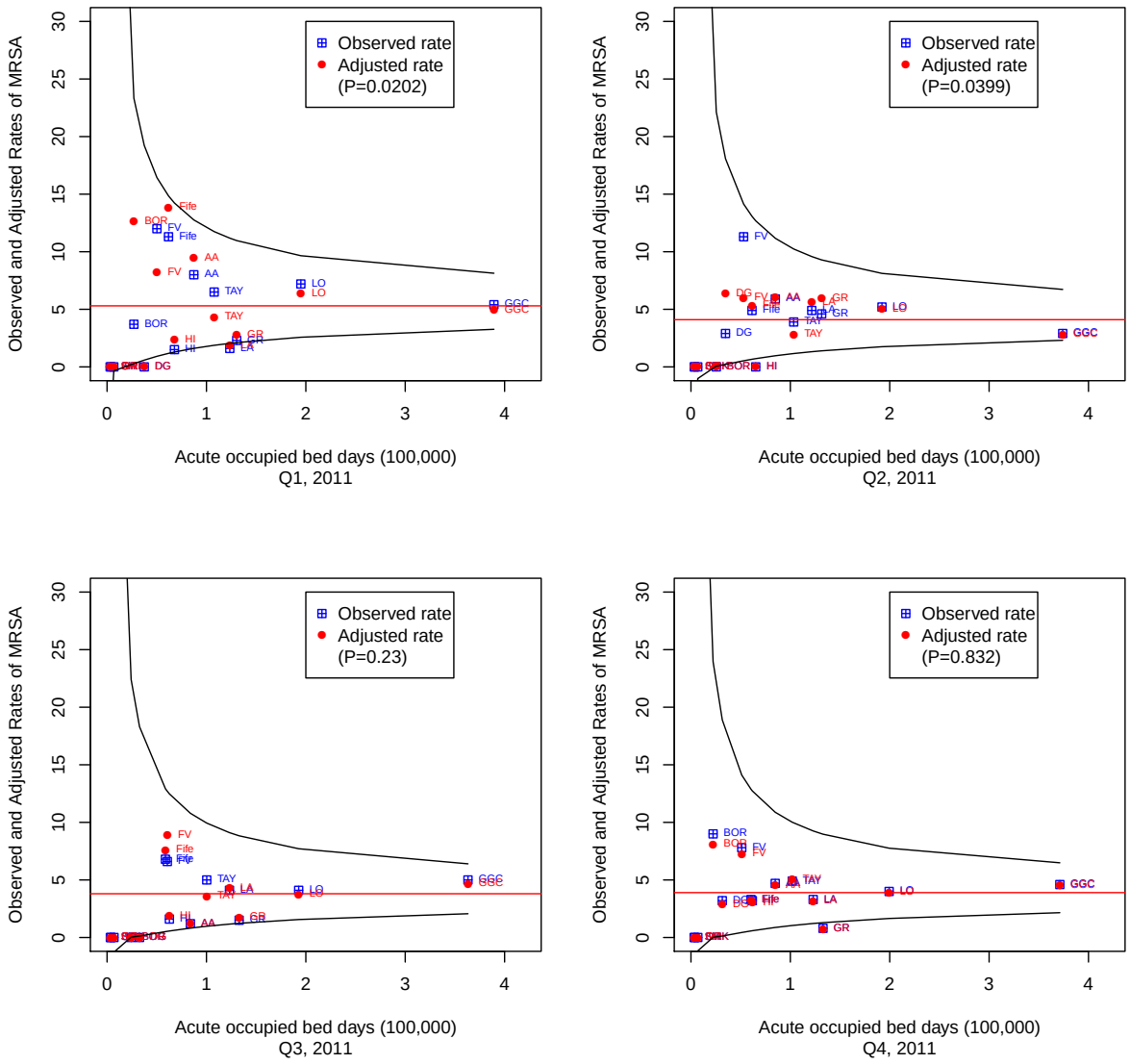



Figure A.3: Funnel plots of adjusted and unadjusted MRSA bacteraemia rates in 2011.

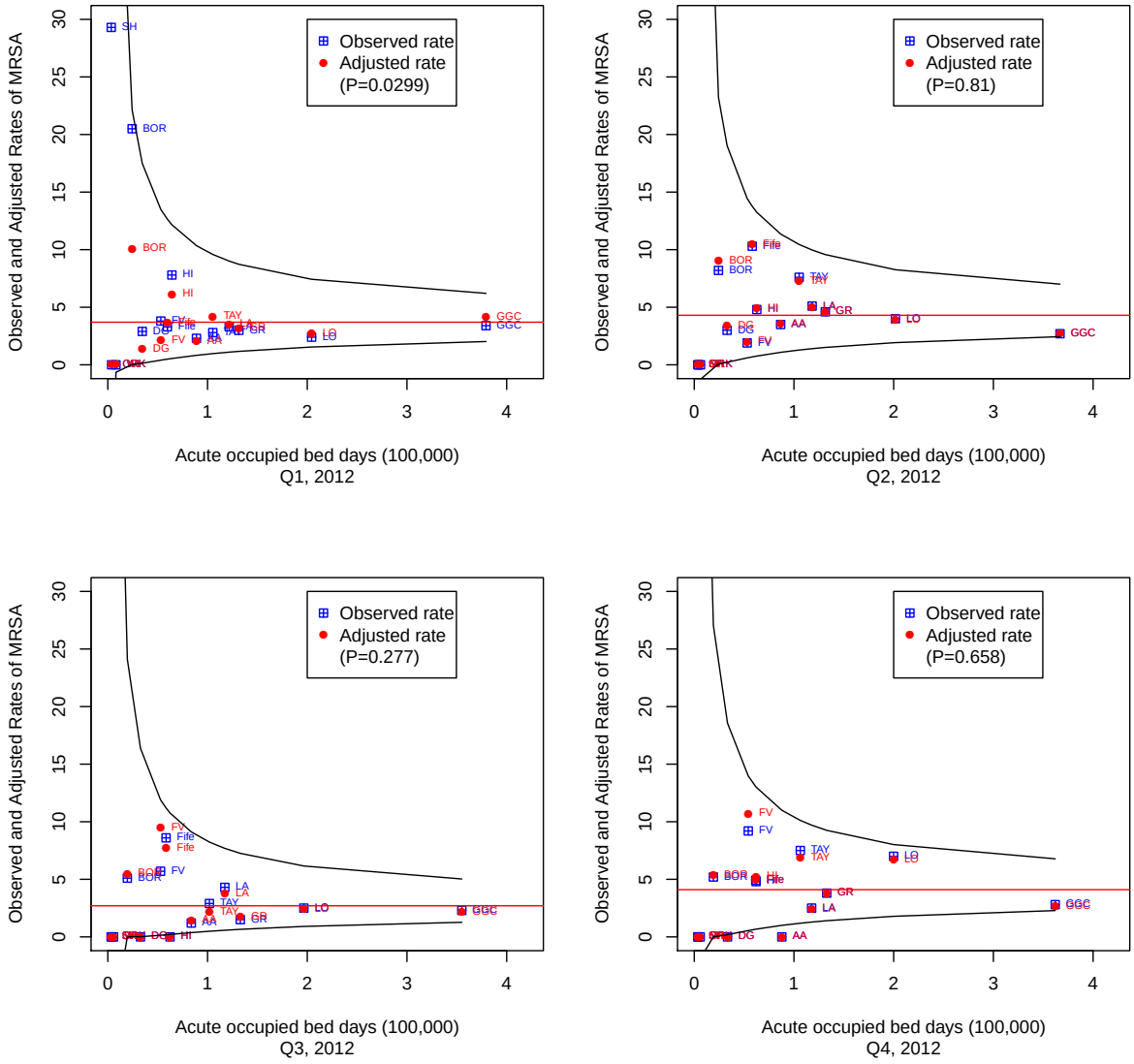



Figure A.4: Funnel plots of adjusted and unadjusted MRSA bacteraemia rates in 2012.

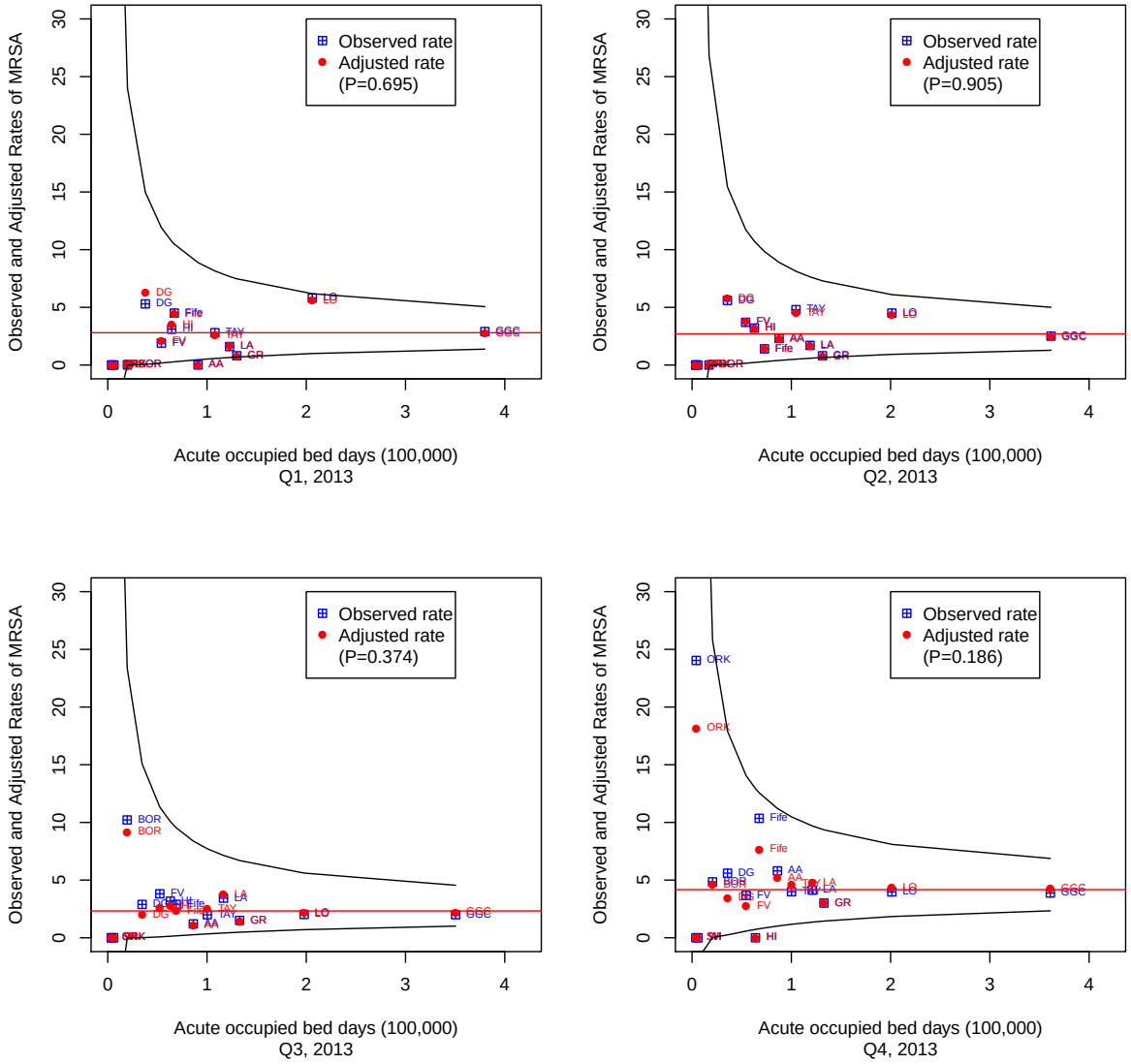



Figure A.5: Funnel plots of adjusted and unadjusted MRSA bacteraemia rates in 2013.

### A.4.3 Funnel plots of MSSA bacteraemia for different quarters

The following Figures show funnel plots of unadjusted MSSA bacteraemia rates and adjusted MSSA bacteraemia rates by acute surgical procedure (ASP) in different quarters from Qu2, 2009 to Qu4, 2013.

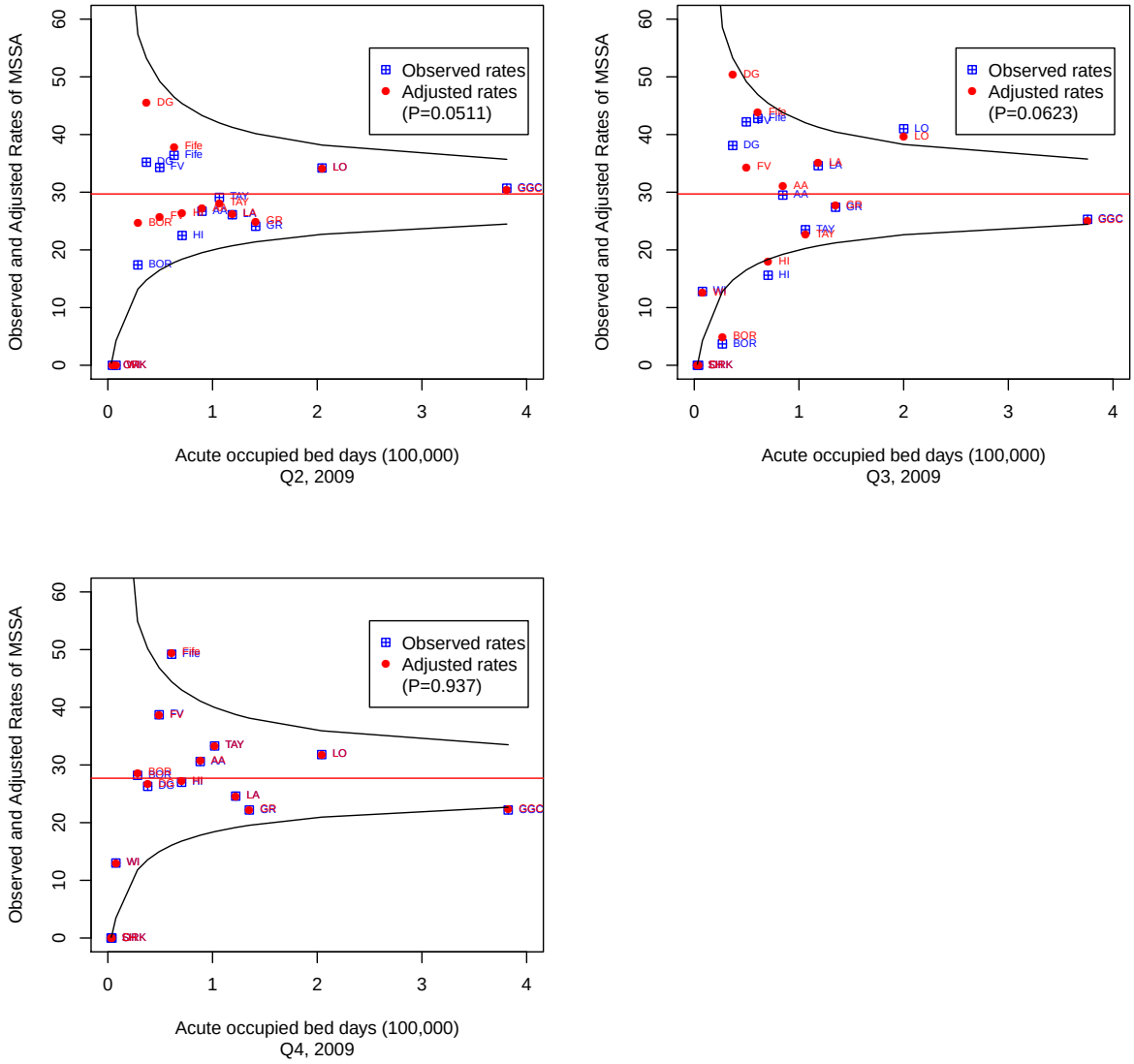



Figure A.6: Funnel plots of adjusted and unadjusted MSSA bacteraemia rates in 2009.

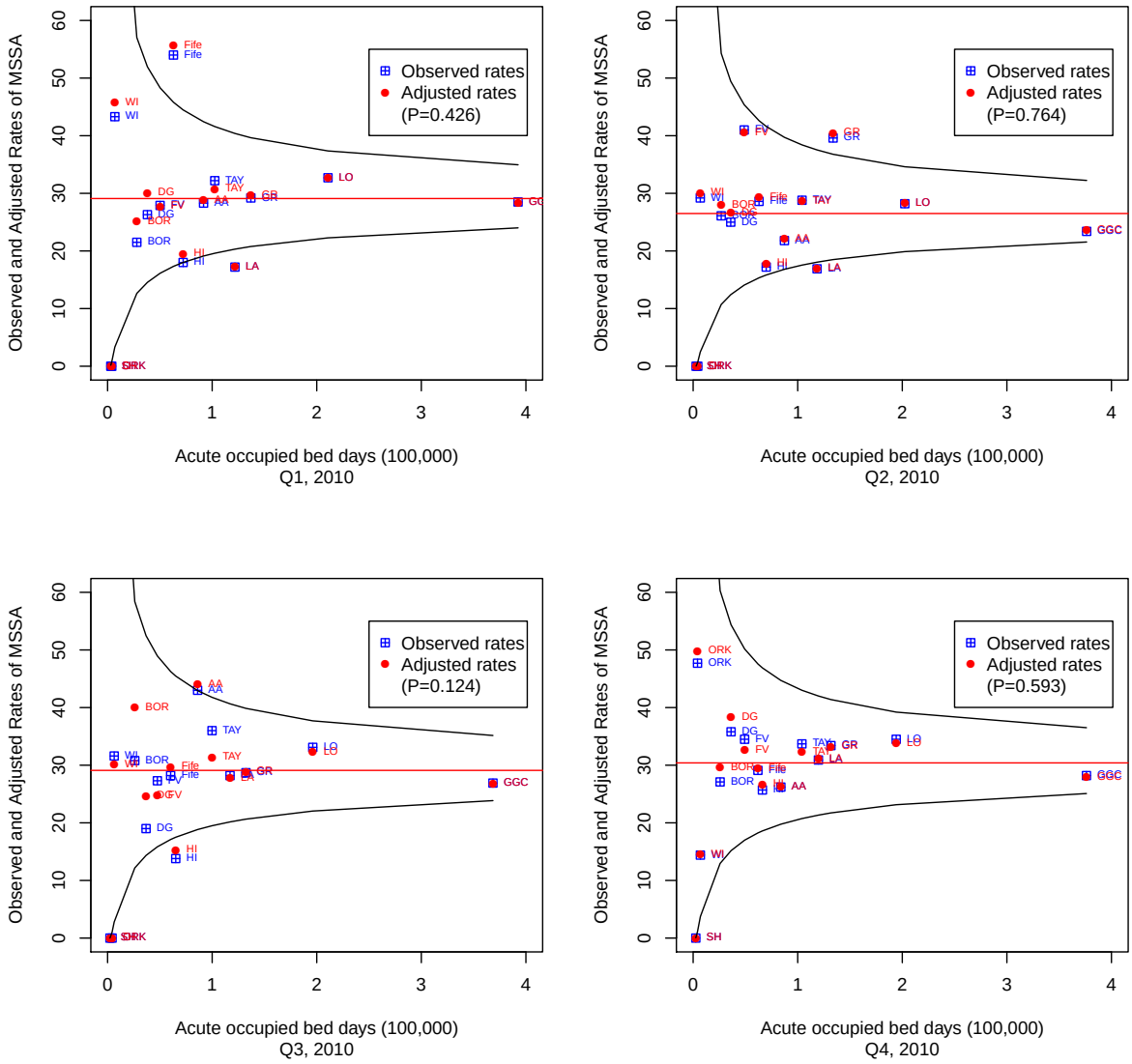



Figure A.7: Funnel plots of adjusted and unadjusted MSSA bacteraemia rates in 2010.

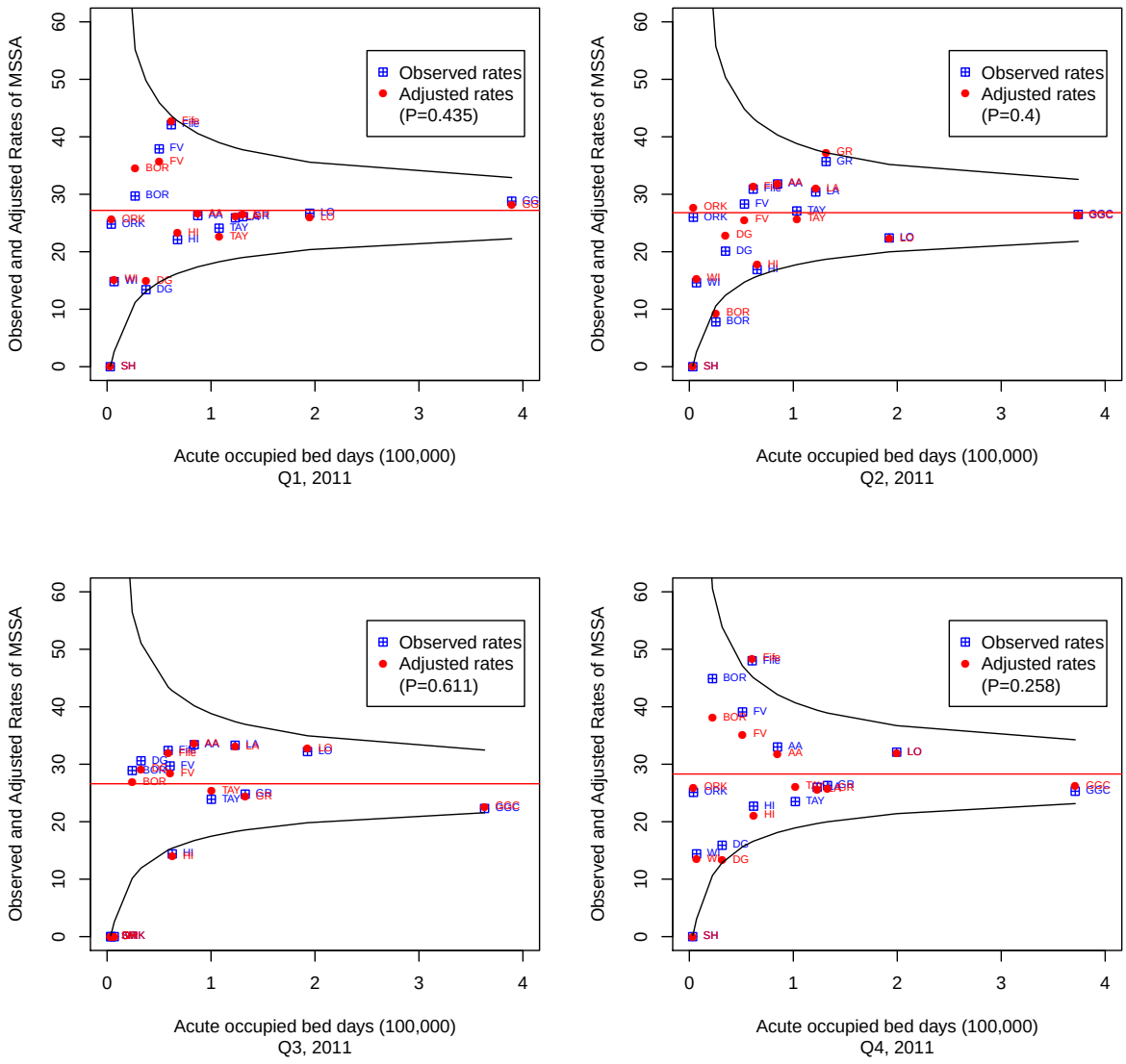



Figure A.8: Funnel plots of adjusted and unadjusted MSSA bacteraemia rates in 2011.

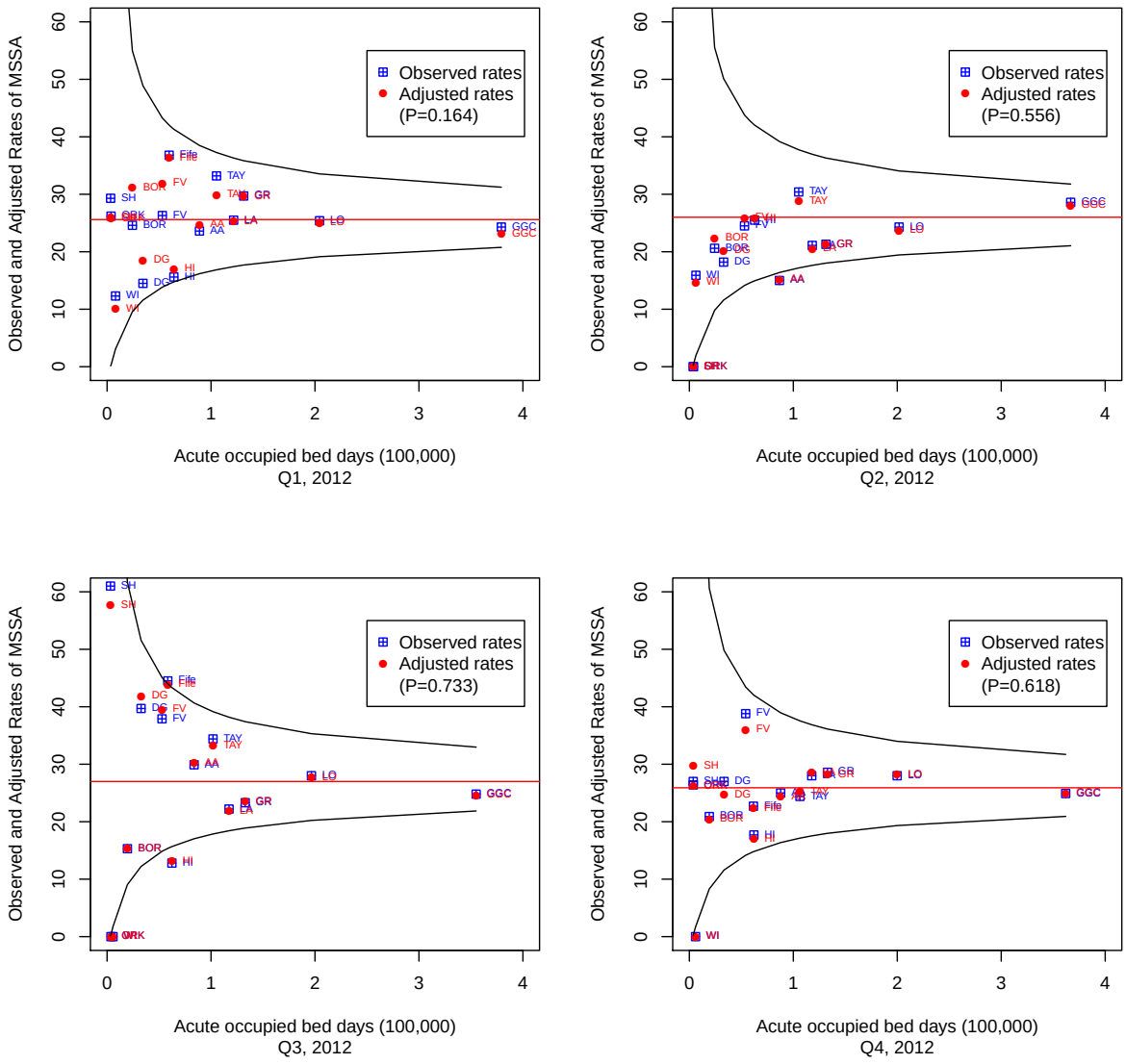



Figure A.9: Funnel plots of adjusted and unadjusted MSSA bacteraemia rates in 2012.

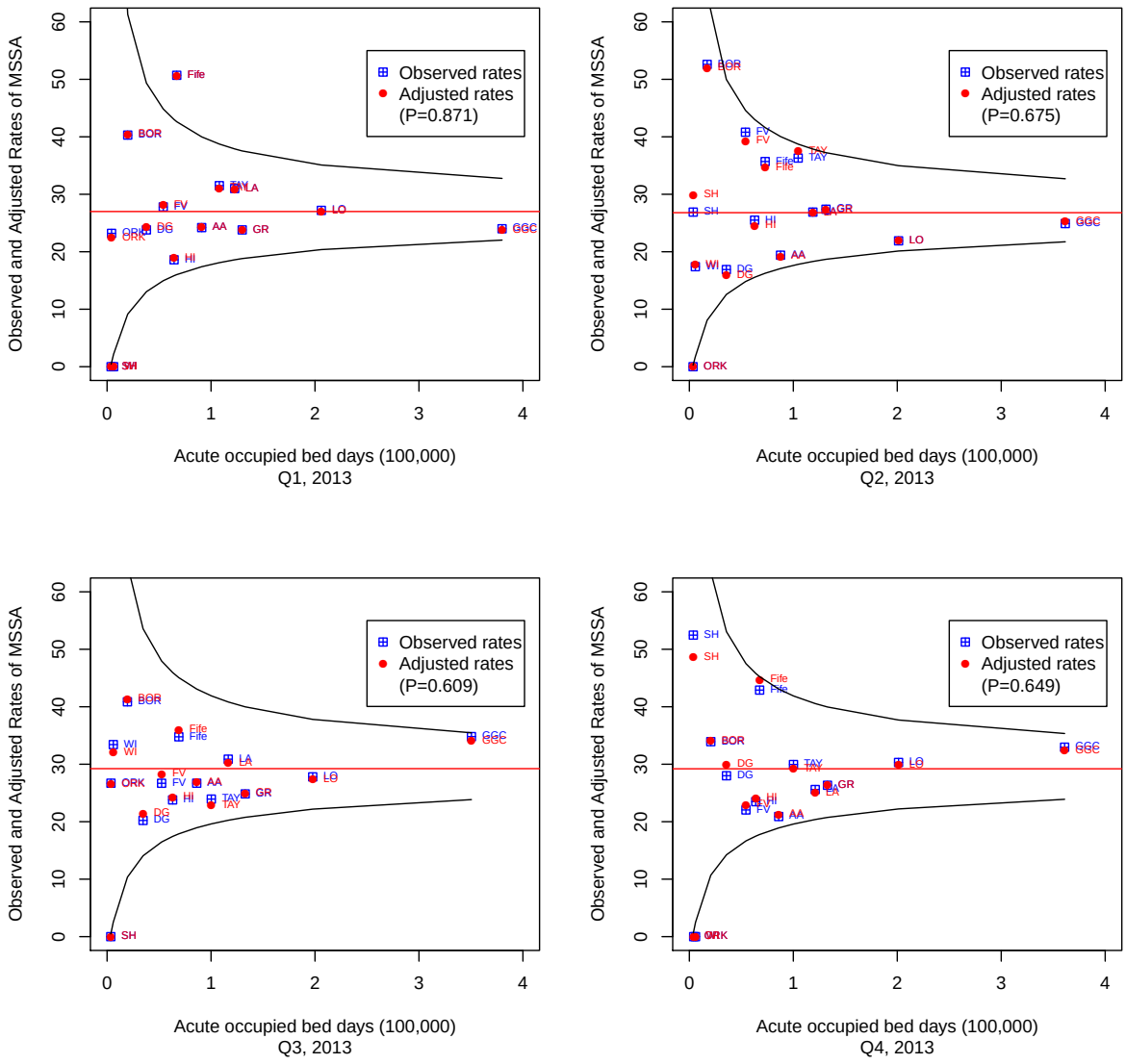



Figure A.10: Funnel plots of adjusted and unadjusted MSSA bacteraemia rates in 2013.

# Appendix B

# Polynomial GLM Regression - Chapter 4

## B.1 R code for simulation study of comparing delta and bootstrap methods

See Section 4.5 for the algorithm and results.

```
library(lmtest)     # to calculate Likelihood ratio test
library(boot)       # for bootstrapping
library(xtable)     # to print R code to latex file
library(XLConnect) # to read .xlsx files

segm <- readWorksheetFromFile("joinpointDATA-uptoSep2014.xlsx",
sheet=5, header=T) # data includes 50 data points
attach(segm)
x <- segm$x

b0 <- 5      # Can be 1.5 ,3 or 5
b1 <- 0.001
b2 <- -0.003
x0 <- 25 # Can be 25, 13 or 38 as true turning point

# Quadratic model is a started model with true turning point.
xq <- b0+ b1*(x-x0) + b2*((x-x0)^2)
xqr<- exp(xq)
xd <- rpois(length(xq),xqr)
plot(x,log(xd), xlab="Time", ylab="Rate" )
lines(x,xq)
abline(v=x0)

# To do 500 simulation for the final result of CI
n.sim <- 500
```

```
resultCI <- matrix(NA, nrow=n.sim, ncol=6)
resultCId <- matrix(NA, nrow=n.sim, ncol=10)

# Bootstrap function
bs <- function(data, indices) {
data$new.n <- round(data$fit + data$resid[indices]*sqrt(data$fit),0)
data$new.n[data$new.n<0] <- 0
rz <- glm(new.n ~ x + x2, data=data,family=poisson)
z.a<- rz$coefficients["x2"]
z.b<- rz$coefficients["x"]
z.c <- rz$coefficients["(Intercept)"]
A<- 2*z.a
B<- z.b
C<- z.c
xx <- polyroot(z=c(B,A))
rex <-Re(xx) # to construct a real part of xx
tp1 <- rex[1]
return(cbind(tp1))
}

for(i in 1:n.sim) {

# Simulate data according to quadratic model
new.y <- rpois(length(xqr),xqr)
d <- data.frame(cbind(new.y,segm))
d$x2 <- d$x^2

# Fit quadratic model to simulated data
z <- glm(new.y ~ x + x2, data=d,family=poisson)

# Calculate the root of quadratic equation
rexd <- Re(polyroot(z=c(z$coef[2],2*z$coef[3])))
z.res <- residuals(z,type="pearson")
z.predict <- predict(z,type="response")# predicted counts

# Calculate CI using delta method
s.m <- summary(z)
v.b1 <- (s.m$coefficients[ 2, 2])^2
v.b2 <- (s.m$coefficients[ 3, 2])^2
m.b1 <- s.m$coefficients[ 2, 1]
m.b2 <- s.m$coefficients[ 3, 1]
cov.b1.b2 <- s.m$cov.scaled [3,2]
E.b1.b2 <-  m.b1/ m.b2
V.b1.b2 <- ((m.b1/ m.b2)^2)*((v.b1/ (m.b1)^2)
+(v.b2/(m.b2)^2)-2*(cov.b1.b2)/(m.b1* m.b2))
EX <- - 0.5 *  E.b1.b2
VAR <- 0.25 * V.b1.b2
sq <- sqrt(VAR)
LL <- EX - 1.96 * sq
UL <- EX + 1.96 * sq
CId <- c(rexd, E.b1.b2, EX, VAR, LL, UL, UL-LL ,z$coef)
resultCId[i,] <- CId

## For bootstrapping process
new.d1 <- cbind(d, resid = z.res, fit = z.predict)
data <- new.d1
results <- boot(data=data, statistic=bs,R=500)
LB <- quantile(results$t[,1],0.025)
UB <- quantile(results$t[,1],0.975)
```

```
CI <- c(rexd,LB, UB, z$coef)
resultCI[i,] <- CI
gc()
}

ttp <- x0
# To save the main results in vector
Res.sim<- matrix(NA, nrow=1, ncol=10)
Res.sim[1,1] <- b0
Res.sim[1,2] <- mean(rexd)

# Results from delta method
Res.sim[1,3] <- mean(resultCId[,5]) # mean lower CI
Res.sim[1,4] <- mean(resultCId[,6]) # mean upper CI
Res.sim[1,5] <- mean(resultCId[,7]) # mean width CI
# Percentage of cover true turning point
Res.sim[1,6] <- sum(resultCId[,5] <= ttp & resultCId[,6] >= ttp)/n.sim*100

# Results from bootstrap method
Res.sim[1,7] <- mean(resultCI[,2]) # mean lower CI
Res.sim[1,8] <- mean(resultCI[,3]) # mean upper CI
Res.sim[1,9] <- mean(resultCI[,3]- resultCI[,2]) # mean width CI
# Percentage of cover true turning point
Res.sim[1,10] <- sum(resultCI[,2] <= ttp & resultCI[,3] >= ttp )/n.sim*100

Res.sim
```

## B.2 R code for simulation study of investigating confidence intervals of two change points

```
# True turning points
t1 <- 16
t2 <- 33

para <- function(t1,t2){
bb1=t1*t2;
bb2=-0.5*(t1+t2);
bb3=1/3;
values=c(bb1,bb2,bb3);
return(values);
}
m <- para(t1,t2)
m1<- m/10000
m2<- round(m1,6)
b0<-  5
b1<- m2[1]
b2<- m2[2]
b3<- m2[3]
ttp1<- t1
ttp2<- t2
xq <- b0+ b1*x + b2*x^2 + b3*x^3

# Bootstrap function
```

```
bs3 <- function(data, indices) {
data$new.n <- round(data$fit + data$resid[indices]*sqrt(data$fit),0)
data$new.n[data$new.n<0] <- 0

rz3 <- glm(new.n ~ x+x2+x3 ,family=poisson,data=data)
z.a<- rz3$coefficients["x3"]
z.b<- rz3$coefficients["x2"]
z.c<- rz3$coefficients["x"]
z.d <- rz3$coefficients["(Intercept)"]
A<- 3*z.a
B<- 2*z.b
C<- z.c
xx<- polyroot(z=c(C,B,A))
srex3b<-Re(xx)
rex3b<- sort(srex3b)
tp13<- rex3b[1] # t when the first change appear
tp23<- rex3b[2] # t when the second change appear
return(cbind(tp13,tp23))
}

n.sim <- 500 # no. of simulation

# Output from one simulation with all data
out1 <- matrix(NA, nrow=n.sim, ncol=15)
d <- data.frame(cbind(segm))
d$x2 <- d$x^2
d$x3 <- d$x^3
lenxqr <- length(xqr)

for(i in 1:n.sim) {
# Simulate data according to original model
d$y <- rpois(lenxqr,xqr)

# Cubic model
z3 <- glm(y~  x + x2 +x3 ,data=d,family=poisson)
zz3<- summary(z3)
pvalue3<- zz3$coefficients["x3",c("Pr(>|z|)")]
esb3 <- zz3$coefficients["x3",c("Estimate")]
ABS3<- abs(esb3)
z.a<- z3$coefficients["x3"]
z.b<- z3$coefficients["x2"]
z.c<- z3$coefficients["x"]
z.d <- z3$coefficients["(Intercept)"]
A<- 3*z.a
B<- 2*z.b
C<- z.c
xx<- polyroot(z=c(C,B,A))
srex3 <- Re(xx) # turning points of cubic model
rex3 <- sort(srex3) # sort tps from the smallest

# To get CI for tps by bootstrapping from cubic model
z.res <- residuals(z3,type="pearson")
z.predict <- predict(z3,type="response")# these are predicted counts
data  <- cbind(d, resid = z.res, fit = z.predict)
results3 <- boot(data=data, statistic=bs3,R=500)

# To remove tp1=tp2 (indicating complex numbers) from bootstrap samples
kk <- subset (results3$t, results3$t[,1]!=results3$t[,2])
```

```
CI13 <- quantile(kk[,1],c(0.025, 0.975))
WCI13 <- CI13[2]- CI13[1]
CI23 <- quantile(kk[,2],c(0.025, 0.975))
WCI23 <- CI23[2]- CI23[1]

out3<-  c(rex3[1],sd(kk[,1]), CI13[1],CI13[2],WCI13,
rex3[2],sd(kk[,2]),CI23[1],CI23[2],WCI23,
Im(xx),z3$deviance,pvalue3,ABS3)

result <- c(out3)
out1[i,] <- result
gc()
}

nrow(out1)

# To remove all t1=t2 from simulation samples
OUT1 <- subset(out1, out1[,1]!=out1[,6])
nrow(OUT1)

tp1 <- mean(OUT1[,1])
ltp1 <- mean(OUT1[,3])
utp1 <- mean(OUT1[,4])
wdtp1 <- mean(OUT1[,5])
ci1 <- sum(ttp1 >= OUT1[,3]& ttp1 <= OUT1[,4])/ nrow(OUT1)*100

tp2 <- mean(OUT1[,6])
ltp2 <- mean(OUT1[,8])
utp2 <- mean(OUT1[,9])
wdtp2 <- mean(OUT1[,10])
ci2 <- sum(ttp2 >= OUT1[,8]& ttp2 <= OUT1[,9])/ nrow(OUT1)*100
```

# Appendix C

# Segmented and Joinpoint Regression - Chapter 5

## C.1 R code for segmented regression analysis

Data of MRSA bacteraemia up to September 2014 is used to explain the segmented regression analysis algorithm. See Section 5.1.2.

```
library(XLConnect)

int<-readWorksheetFromFile("Interventions.xlsx",
sheet=7, header=T)

int$t <- int$time-2003
int$Qu <- factor(int$Qu)

### Segmented regression using all data before and after

# Segmented regression for all point with one change point
z.int <- names(int)[7:19]
z.int.s <- names(int)[20:32]

if (exists("z.out")) rm(z.out)
for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.ii <- z.int[z.index]
z.row <- as.numeric(gsub("Int","",z.ii))
z.abline <- int[z.row,"time"]

# Fit segmented at each specific time when interventions took place
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t + Qu + get(z.i),
data=int,family=poisson)
```

```
z1 <- summary(z)
z.time <- z1$coefficients["t",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]
z.res <- matrix(c(z.time,z.change,z$deviance,z$df.residual),nrow=1)

dimnames(z.res) <- list(z.ii,c("Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","DF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
z.out

# Segmented regression for all point with two change points
if (exists("z.out")) rm(z.out)

for (z.index in 5:length(z.int.s)) {
# "5" is the location after the first change point detected

z.i <- z.int.s[z.index]
z.ii <- z.int[z.index]
z.row <- as.numeric(gsub("Int","",z.ii))
z.abline <- int[z.row,"time"]
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+ Qu + Time.PI14 + get(z.i)
,data=int,family=poisson)

z1 <- summary(z)
z.time <- z1$coefficients["t",c("Estimate","Std. Error","Pr(>|z|)")]
z.time14 <- z1$coefficients["Time.PI14",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]
z.res <- matrix(c(z.time,z.time14,z.change,z$deviance,z$df.residual),nrow=1)

dimnames(z.res) <- list(z.ii,c("Slope","Slope.SE","Slope.P", "Slope14",
"Slope14.SE","Slope14.P","Change.Slope","Change.Slope.SE",
"Change.Slope.P","Deviance","DF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
z.out

# Given two change points, test for the third one as previous step.
# Then test if the models are different
z2 <- glm(no.MRSA1 ~ offset (log(aobd)) + t + Qu +
Time.PI14+ Time.PI22 + Time.PI33
 ,data=int,family=poisson)
summary(z2)

z1 <- glm(no.MRSA1 ~ offset (log(aobd)) + t + Qu +
Time.PI14 + Time.PI22
 ,data=int,family=poisson)

anova(z1,z2, test="Chisq")

### Segmented regression with all data before and two years after
if (exists("z.out")) rm(z.out)

for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.ii <- z.int[z.index]
z.row <- as.numeric(gsub("Int","",z.ii))
```

```
z.abline <- int[z.row,"time"]

# To choose dataset all before and two years after
z.df<- subset(int,time <= z.abline +2 )

z.null<- glm(no.MRSA1 ~ offset (log(aobd)) , # fit null model
data=z.df,family=poisson)
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+ Qu + get(z.i),
data=z.df,family=poisson)
z1 <- summary(z)
z.time <- z1$coefficients["t",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]

# Percentage of deviance explained
PDE <- (1- (z$deviance/z.null$deviance))*100
PDEs<- (z$deviance/z.null$deviance)*100

z.res <- matrix(c(z.time,z.change,z$deviance, z.null$deviance , PDE,PDEs,
z$df.residual, z.null$df.residual),nrow=1)
dimnames(z.res) <- list(z.ii,c("Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","NullDev",
"PDE","PDEs","DF","NullDF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
z.out

### Segmented regression of two years of data before and after
if (exists("z.out")) rm(z.out)

for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.ii <- z.int[z.index]
z.row <- as.numeric(gsub("Int","",z.ii))
z.abline <- int[z.row,"time"]

# To choose dataset two years before and after
z.df<- subset(int,time <= z.abline +2 & time > z.abline -2)

z.null<- glm(no.MRSA1 ~ offset(log(aobd)) ,
 data=z.df,family=poisson)
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+ Qu + get(z.i),
 data=z.df,family=poisson)
z1 <- summary(z)
z.time <- z1$coefficients["t",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]
PDE<- (1- (z$deviance/z.null$deviance))*100
PDEs<- (z$deviance/z.null$deviance)*100

z.res <- matrix(c(z.time,z.change,z$deviance, z.null$deviance , PDE,PDEs,
z$df.residual, z.null$df.residual),nrow=1)
dimnames(z.res) <- list(z.ii,c("Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","NullDev",
"PDE","PDEs","DF","NullDF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
z.out
```

## C.2 R codes for joinpoint analysis

### C.2.1 R code for joinpoint detection

Data of MRSA bacteraemia up to June 2016 is used to explain the algorithm of detection two joinpoints. See Section 5.2.1.

```
library(XLConnect)

JDATA<-readWorksheetFromFile("JoinpointDATA-uptoJune2016.xlsx",
sheet=1, header=T)
JDATA$t <- JDATA$time-2003
JDATA$Qu <- factor(JDATA$Qu)
z.int.s <- names(JDATA)[8:59]

########## 1- Fit H0 (null model with 2 jps) ##########
h0 <- 2

if (exists("z.out")) rm(z.out)

#one jp
for (z.index in 1:(length(z.int.s)-1)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("Time.PI","",z.i))
z.abline <- JDATA[z.row,"time"]
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+Qu + get(z.i) ,
        data=JDATA,family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

#two jps
for (z.index2 in (z.index+1): length(z.int.s)) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("Time.PI","",z.j))
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t +Qu + get(z.i) + get(z.j),
        data=JDATA,family=poisson)

z.res <- matrix(c(z.row ,z.row2,z.dev,z.df,z$deviance,z$df.residual),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","jp2","Dev.1","DF.1","Deviance","DF"))
if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
}

MinDev <- min(z.out[,"Deviance"])
z.out[z.out[,"Deviance"] <= MinDev,]
MinDev.1 <- min(z.out[,"Dev.1"])
z.out[z.out[,"Dev.1"] <= MinDev.1,]

z.h0 <- glm(no.MRSA1 ~ offset (log(aobd)) + t +Qu + Time.PI19 + Time.PI37,
data=JDATA,family=poisson)
z.h0.pearson.resid <- residuals(z.h0,type="pearson")
z.h0.mean <- predict(z.h0,type="response")

########## 2- Fit H1 (model with 3 jps to the original data) ##########
h1 <- 3
```

```
if (exists("z.out")) rm(z.out)

#one jp
for (z.index in 1:(length(z.int.s)-2)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("Time.PI","",z.i))
z.abline <- JDATA[z.row,"time"]
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t + Qu + get(z.i) ,
        data=JDATA,family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

#two jps
for (z.index2 in (z.index+1): (length(z.int.s)-1)) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("Time.PI","",z.j))
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t + Qu + get(z.i) + get(z.j),
        data=JDATA,family=poisson)
z.dev2 <- z$deviance
z.df2 <- z$df.residual

#three jps
for (z.index3 in (z.index2+1): length(z.int.s)) {
z.k <- z.int.s[z.index3]
z.row3 <- as.numeric(gsub("Time.PI","",z.k))
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+Qu + get(z.i) + get(z.j) + get(z.k),
         data=JDATA,family=poisson)

z.res <- matrix(c(z.row ,z.row2, z.row3
,z.dev,z.df,z.dev2,z.df2,z$deviance,z$df.residual),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","jp2","jp3","Dev.1","DF.1",
"Dev.2","DF.2","Deviance","DF"))
if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
}
}

MinDev <- min(z.out[,"Deviance"])
Three.CP.Dev <- z.out[z.out[,"Deviance"] <= MinDev,]
MinDev.2 <- min(z.out[,"Dev.2"])
z.out[z.out[,"Dev.2"] <= MinDev.2,]
MinDev.1 <- min(z.out[,"Dev.1"])
z.out[z.out[,"Dev.1"] <= MinDev.1,]

########## 3- Fit H1 model with 3 jps to,#
#permuted sampled data from H0 ##########

if (exists("z.sim.res")) rm(z.sim.res)
for (i.sim in 1:100) { # 100 is number of permutation
if (exists("z.out")) rm(z.out)
z.h0.perm.resid <- sample(z.h0.pearson.resid,length(z.h0.pearson.resid))
z.new.resp <- round(z.h0.mean + z.h0.perm.resid * sqrt(z.h0.mean),0)

#one jp
for (z.index in 1:(length(z.int.s)-2)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("Time.PI","",z.i))
z.abline <- JDATA[z.row,"time"]
z <- glm(z.new.resp ~ offset (log(aobd)) + t + Qu + get(z.i) ,
```

```
        data=JDATA,family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

#two jps
for (z.index2 in (z.index+1): (length(z.int.s)-1)) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("Time.PI","",z.j))
z <- glm(z.new.resp ~ offset (log(aobd)) + t + Qu + get(z.i) + get(z.j),
        data=JDATA,family=poisson)
z.dev2 <- z$deviance
z.df2 <- z$df.residual

#three jps
for (z.index3 in (z.index2+1): length(z.int.s)) {
z.k <- z.int.s[z.index3]
z.row3 <- as.numeric(gsub("Time.PI","",z.k))
z <- glm(z.new.resp ~ offset (log(aobd)) + t+ Qu +
        get(z.i) + get(z.j) + get(z.k),
        data=JDATA,family=poisson)

z.res <- matrix(c(z.row ,z.row2, z.row3
,z.dev,z.df,z.dev2,z.df2,z$deviance,z$df.residual),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","jp2","jp3","Dev.1","DF.1",
"Dev.2","DF.2","Deviance","DF"))
if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}

MinDev <- min(z.out[,"Deviance"])
z.out[z.out[,"Deviance"] <= MinDev,]
}
MinDev.2 <- min(z.out[,"Dev.2"])
z.out[z.out[,"Dev.2"] <= MinDev.2,]
}
MinDev.1 <- min(z.out[,"Dev.1"])
z.out[z.out[,"Dev.1"] <= MinDev.1,]

if (exists("z.sim.res"))
z.sim.res <- rbind(z.sim.res,z.out[z.out[,"Deviance"] <= MinDev,])
 else z.sim.res <- z.out[z.out[,"Deviance"] <= MinDev,]
}

########## 4- permutation test to choose no. of jps ##########
# Calculate change in deviance with data
z.cd.data <- z.h0$deviance - Three.CP.Dev["Deviance"]

# Calculate change in deviances with permuted residuals data
z.cd.prd <- z.h0$deviance - z.sim.res[,"Deviance"]

# P-value of the test
z.permutation.p.value <- sum(z.cd.prd > z.cd.data )/(length(z.cd.prd)+1)

# P-value after using Bonferroni correction for the overall
#significance level of 0.05
if z.permutation.p.value < 0.05/(h1-h0) print(z.permutation.p.value)
```

### C.2.2 R code for confidence intervals of two joinpoints

Data of MRSA bacteraemia up to June 2016 is used to explain the bootstrap confidence intervals of two joinpoints. See Section 5.2.2.2 for the bootstrap confidence interval algorithm.

```
########## Construct CI for two jps ##########
########## Bootstrap method ##########

z.h0 <- glm(no.MRSA1 ~ offset (log(aobd)) + t +Qu + Time.PI19 + Time.PI37,
data=JDATA,family=poisson)
z.h0.pearson.resid <- residuals(z.h0,type="pearson")
z.h0.mean <- predict(z.h0,type="response")

n.boots <- 1000
if (exists("z.sim.res")) rm(z.sim.res)
for (i.sim in 1: n.boots) {
if (exists("z.out")) rm(z.out)
z.h0.perm.resid <- sample(z.h0.pearson.resid,length(z.h0.pearson.resid))
z.new.resp <- round(z.h0.mean + z.h0.perm.resid * sqrt(z.h0.mean),0)

#one jp
for (z.index in 1:(length(z.int.s)-1)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("Time.PI","",z.i))
z.abline <- JDATA[z.row,"time"]
z <- glm(z.new.resp ~ offset (log(aobd)) + t + Qu + get(z.i) ,
        data=JDATA,family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

#two jps
for (z.index2 in (z.index+1): length(z.int.s) ) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("Time.PI","",z.j))
z <- glm(z.new.resp ~ offset (log(aobd)) + t + Qu + get(z.i) + get(z.j),
        data=JDATA,family=poisson)
z.dev2 <- z$deviance
z.df2 <- z$df.residual

z.res <- matrix(c(z.row ,z.row2,z.dev,z.df,z.dev2,z.df2),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","jp2","Dev.1","DF.1",
"Dev.2","DF.2"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}

MinDev.2 <- min(z.out[,"Dev.2"])
z.out[z.out[,"Dev.2"] <= MinDev.2,]
}

MinDev.1 <- min(z.out[,"Dev.1"])
z.out[z.out[,"Dev.1"] <= MinDev.1,]

if (exists("z.sim.res"))
z.sim.res <- rbind(z.sim.res,z.out[z.out[,"Dev.2"] <= MinDev.2,])
```

```
 else z.sim.res <- z.out[z.out[,"Dev.2"] <= MinDev.2,]
}

quantile(z.sim.res[,1],c(0.025, 0.975))
quantile(z.sim.res[,2],c(0.025, 0.975))
```

### C.2.3 R code for confidence interval of one joinpoint using bootstrap and profile likelihood methods

Data of MRSA bacteraemia up to September 2014 is used to explain the bootstrap and profile likelihood methods of constructing confidence interval of one joinpoint. See Section 5.2.2 for the bootstrap and profile likelihood confidence intervals algorithms.

```
########## Construct CI for one jp ##########
########## 1- Bootstrap method ##########
if (exists("z.out")) rm(z.out)
for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("Time.PI","",z.i))
z <- glm(no.MRSA1 ~ offset (log(aobd)) + t+Qu + get(z.i) ,
 data=JDATA,family=poisson)

z.dev1.h0 <- z$deviance
z.df1.h0 <- z$df.residual
z1 <- summary(z)
z.res <- matrix(c(z.row ,z.dev1.h0,z.df1.h0),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","Dev.1.h0","DF.1.h0"))
if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
z.out

MinDev.1.h0 <- min(z.out[,"Dev.1.h0"])
z.out[z.out[,"Dev.1.h0"] <= MinDev.1.h0,]

z.h0 <- glm(no.MRSA1 ~ offset (log(aobd)) + t+Qu   + Time.PI18 ,
 data=JDATA,family=poisson)
z.res <- residuals(z.h0,type="pearson")
z.predict <- predict(z.h0,type="response")
new.d1 <- cbind(JDATA, resid = z.res, fit = z.predict)
data <- new.d1

bs <- function(data, indices) {
data$new.n.mrsa <- round(data$fit + data$resid[indices]*sqrt(data$fit),0)
data$new.n.mrsa[data$new.n.mrsa<0] <- 0
z.int.s <- names(JDATA)[8:52]
if (exists("z1.out")) rm(z1.out)
for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
```

```
z.row <- as.numeric(gsub("Time.PI","",z.i))
z.abline <- JDATA[z.row,"time"]
rz1 <- glm(new.n.mrsa ~ offset (log(aobd)) + t + Qu + get(z.i),
           data=data,family=poisson)
z1r <- summary(rz1)
z.time <- z1r$coefficients["t",c( "Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1r$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]
z.res <- matrix(c(z.row,z.time,z.change,rz1$deviance,rz1$df.residual),nrow=1)
dimnames(z.res) <- list(z.i,c("jp","Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","DF"))

#z1.out has all the results for each possible joinpoint
if (exists("z1.out")) z1.out <- rbind(z1.out,z.res) else z1.out <- z.res
}


a <- z1.out[,8] #deviancs

#jb has 2 outputs: 1- min dev, 2- row corresponding to the min dev (jp)
jb <- c(min(a),which(a==min(a)))
jb[2] <- (as.numeric(gsub("Time.PI","",z.int.s[jb[2]])))

return(cbind(jb[1],jb[2]))
}

results <- boot(data=data, statistic=bs ,R=1000)
lcljp <-  quantile(results$t[,2],c(0.025))
ucljp <-  quantile(results$t[,2],c(0.975))
wdcljp <-  ucljp - lcljp

#########################################################################
########## 2- profile likelihood method ##########

library(sp)    # Provides basic spatial classes/methods, SpatialLines
library(rgeos) # Includes intersection function

z.h0 <- glm(no.MRSA1 ~ offset (log(aobd)) + t+Qu   + Time.PI18 ,
 data=JDATA,family=poisson)
z.h0.pearson.resid <- residuals(z.h0,type="pearson")
z.h0.mean <- predict(z.h0,type="response")

a1 <- z.out[,2] #deviancs
b1 <- rep(MinDev.1.h0 + qchisq(0.95,1),45) # horizontal line
SL1 <- SpatialLines(list(Lines(Line(cbind(seq_along(a1),a1)), "A")))
SL2 <- SpatialLines(list(Lines(Line(cbind(seq_along(b1),b1)), "B")))

# Find intersections
coords <- coordinates(gIntersection(SL1, SL2))

jopt <-  z.out[,1] # joinpoints
plot(jopt , a1 , type="l")
abline(h= MinDev.1.h0 + qchisq(0.95,1), col="red")
abline(v= coords[1]+1, col="red")
abline(v= coords[2]+1, col="red")
lcl <-  coords[1]+1
ucl <-  coords[2]+1
wdcl <- ucl-lcl
points(coords+1, col="red", pch=16)
```

### C.2.4 R code for simulation study of comparing profile likelihood and bootstrap methods of confidence interval for one joinpoint

See Section 5.3 for the algorithm and results.

```
library(XLConnect)
library(boot)
library(sp)    # Provides basic spatial classes/methods, SpatialLines
library(rgeos) # Includes intersection function

segm<-readWorksheetFromFile("joinpointDATA-uptoSep2014.xlsx",
sheet=5 , header=T)
x <- segm$x

dsize <- 50
b0 <- 5
b1 <- 0.001  # for quadratic and combined
#b1 <- 0.005 # for segmented
b2 <- -0.003 # for quadratic and combined
#b2 <- -0.03 # for segmented
b3 <- -0.01
x0 <- 25   # true turning point, it can be 25, 13, 38

# Started models with true turning point.
xq <- b0+ b1*(x-x0)+ b2*((x-x0)^2) # quadratic
#OR
xq <- b0+ b1*(x-x0)+b2*((x-x0)^2)+b3*(ifelse(x<=x0,0,x-x0)) # combined
xq <- b0+ b1*(x-x0)+ b2*(ifelse(x<=x0,0,x-x0)) # segmented

ttp <- x0
xqr <- exp(xq)
xd <- rpois(length(xq),xqr)
plot(x,log(xd), xlab="Time", ylab="Rate" , lwd=2)
lines(x,xq, lwd=2 , col=2)
abline(v=ttp, lwd=2 )

bs <- function(data, indices) {
data$new.n <- round(data$fit + data$resid[indices]*sqrt(data$fit),0)
data$new.n[data$new.n<0] <- 0
z.int.s <- names(segm)[3:dsize]

if (exists("z1.out")) rm(z1.out)
for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]
rz1 <- glm(new.n ~ x + get(z.i),
data=data , family=poisson)
z1r <- summary(rz1)
z.time <- z1r$coefficients["x",c("Estimate","Std. Error","Pr(>|z|)")]

z.change <- z1r$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]
z.res <- matrix(c(z.row,z.time,z.change,rz1$deviance,rz1$df.residual),nrow=1)
dimnames(z.res) <- list(z.i,c("jp","Slope","Slope.SE","Slope.P",
```

```
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","DF"))

if (exists("z1.out")) z1.out <- rbind(z1.out,z.res) else z1.out <- z.res
}
z1.out # has all the results for each possible join point

# Deviances
aa <- z1.out[,8]

#jb 2 entries 1- minimum deviance,
#2- row corresponding to the minimum deviance
jb <- c(min(aa),which(aa==min(aa)))

#jb[2] is the jp time
jb[2] <- (as.numeric(gsub("ts","",z.int.s[jb[2]])))

# To return the turning point
return(cbind(jb[1],jb[2]))

}

n.sim <- 200 # number of simulations

# output from one simulation with all data
out1 <- matrix(NA, nrow=n.sim, ncol=8)

for(i in 1:n.sim) {
y <- rpois(length(xqr),xqr)
d <- data.frame(cbind(segm,y))
z.int.s <- names(segm)[3:dsize]
if (exists("z.out")) rm(z.out)

for (z.index in 1:length(z.int.s)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]
z <- glm(y ~ x + get(z.i), data=d, family=poisson)
z1 <- summary(z)
z.time <- z1$coefficients["x",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]

z.res <- matrix(c(z.row,z.time,z.change,z$deviance,z$df.residual),nrow=1)
dimnames(z.res) <- list(z.i,c("jp","Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","DF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}

z.out

a <- z.out[,8] # deviances
j <- c(min(a),which(a==min(a)))
j[2] <- (as.numeric(gsub("ts","",z.int.s[j[2]])))

z1 <- glm( y ~ x + get(as.character(paste("ts",j[2],sep=""))),
data=d, family=poisson)

#### CI for jp using profile likelihood method ###
# To set horizontal line corresponding to minimum deviance.
```

```
b <- rep(min(z.out[,8]) + qchisq(0.95,1),dsize )

# To get deviances curve
SL1 <- SpatialLines(list(Lines(Line(cbind(seq_along(a),a)), "A")))

# To get line corresponding to minimum deviance.
SL2 <- SpatialLines(list(Lines(Line(cbind(seq_along(b),b)), "B")))

# Find intersections between the curve SL1 and the line SL2
coords <- coordinates(gIntersection(SL1, SL2))
ddf <- coords[,1, drop=FALSE] # all intersection points
x1 <- ddf[ddf < j[2]-1] # all intersection points < joinpoint
x2 <- max(x1)
x3 <- ddf[ddf > j[2]-1] # all intersection points > joinpoint
x4 <- min(x3)
lcljp <- x2+1
ucljp <- x4+1
wdcljp <- ucljp-lcljp


#### CI for jp using bootstrap method ###
z.res <- residuals(z1,type="pearson")
z.predict <- predict(z1,type="response")
new.d1 <- cbind(d, resid = z.res, fit = z.predict)
data <- new.d1

results <- boot(data=data, statistic=bs,R=500)

CIjpb <-  quantile(results$t[,2],c(0.025, 0.975))
wdcljpb <-  CIjpb[2] - CIjpb[1]

result <- c(j, lcljp , ucljp , wdcljp  ,CIjpb[1], CIjpb[2] , wdcljpb)

out1[i,] <- result
gc()
}

dimnames(out1)[[2]]<- c("DV","JP","Lp","Up","WDp","Lb","Ub", "WDb")

WD.pr<- mean(out1[,5])
CI.TP.pr <- sum(x0 > out1[,3] & x0 < out1[,4] )/n.sim*100
WD.bt<- mean(out1[,8])
CI.TP.bt <- sum(x0 > out1[,6] & x0 < out1[,7] )/n.sim*100
```

# Appendix D

# Simulation Study - Chapter 6

## D.1 R code for simulation study of comparing change points methods in case of one change point

See Section 6.3 for the results.

```
library(XLConnect)
library(boot)
library(xtable) # to print R code to latex file

segm <- readWorksheetFromFile("joinpointDATA-uptoSep2014.xlsx",
sheet=5, header=T)
x <- segm$x

# Set up true turning point
x0 <- 25

# Set up the coefficient of the original model
b0 <- 5      # is the no. of cases, can be 5, 3, 1.5
b1 <- 0.001  # for quadratic and combined
#b1 <- 0.005 # for segmented

b2 <- -0.003 # for quadratic and combined
#b2 <- -0.03 # for segmented

b3 <- -0.05  # for combined


# Set up the original model (e.g. combined)
xq <- b0+ b1*(x-x0) + b2*((x-x0)^2) +  b3*(ifelse(x<=x0,0,x-x0))
```

```
# OR original models can be
xq <- b0+ b1*(x-x0) + b2*((x-x0)^2)  # quadratic
xq <- b0+ b1*(x-x0) +  b2*(ifelse(x<=x0,0,x-x0)) # segmented

xqr<- exp(xq)
xd <- rpois(length(xq),xqr)
plot(x,log(xd), xlab="Time", ylab="Rate" , lwd=2)
lines(x,xq, lwd=2 , col=2)
abline(v=x0, lwd=2 )

#############################################################
# Code to do 200 simulation for the final result of CI
#out1 - output from one simulation with all data
#out2 - results from each quadratic model
#out3 - results from each cubic model
#outs - results from segmented models.
#z.out - results for each joinpoint
#jpres - results from all joinpoints
#############################################################

n.sim <- 200
# output from one simulation with all data
out1 <- matrix(NA, nrow=n.sim, ncol=37)
d <- data.frame(cbind(segm))
d$x2 <- d$x^2
d$x3 <- d$x^3
lenxqr <- length(xqr)

########## Bootstrap function to
#construct CI for quadratic model##########
bs2 <- function(data, indices) {
  data$new.n <- round(data$fit +
  data$resid[indices]*sqrt(data$fit),0)
  data$new.n[data$new.n<0] <- 0
  rz2 <- glm(new.n ~ x+x2 ,
             family=poisson,data=data)

  z.a<- rz2$coefficients["x2"]
  z.b<- rz2$coefficients["x"]
  z.c<- rz2$coefficients["(Intercept)"]
  A<- 2*z.a
  B<- z.b

# Calculate the root of quadratic equation
  xx<- polyroot(z=c(B,A))
  rex2b <-Re(xx)
  tp1 <- rex2b[1]
  return(cbind(tp1))
}

########## Bootstrap function to
#construct CI for cubic model##########
bs3 <- function(data, indices) {
data$new.n <- round(data$fit +
                    data$resid[indices]*sqrt(data$fit),0)
data$new.n[data$new.n<0] <- 0
rz3 <- glm(new.n ~ x+x2+x3 ,family=poisson,data=data)

z.a<- rz3$coefficients["x3"]
```

```
z.b<- rz3$coefficients["x2"]
z.c<- rz3$coefficients["x"]
z.d <- rz3$coefficients["(Intercept)"]
A<- 3*z.a
B<- 2*z.b
C<- z.c

xx<- polyroot(z=c(C,B,A))
srex3b<-Re(xx)
rex3b<- sort(srex3b)
tp13<- rex3b[1]
tp23<- rex3b[2]

return(cbind(tp13,tp23))
}

########## Bootstrap function to
#construct CI for joinpoint model##########
bsjp <- function(data, indices) {
data$new <- round(data$fit +
            data$resid[indices]*sqrt(data$fit),0)
data$new[data$new<0] <- 0

z.int.s <- names(segm)[3:lenxqr]
if (exists("z1.out")) rm(z1.out)
for (z.index in 1: (length(z.int.s))) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]

# Fit joinpoint model at all possible jps
rz1 <- glm(new ~ x + get(z.i)
           ,data=data , family=poisson)
z.dev <- rz1$deviance
z.df <- rz1$df.residual

z1r <- summary(rz1)
z.time <- z1r$coefficients["x",c("Estimate","Std. Error","Pr(>|z|)")]
z.change <- z1r$coefficients["get(z.i)",c("Estimate","Std. Error","Pr(>|z|)")]

z.res <- matrix(c(z.row,z.time,z.change,rz1$deviance,rz1$df.residual),nrow=1)
dimnames(z.res) <- list(z.i,c("jp","Slope","Slope.SE","Slope.P",
"Change.Slope","Change.Slope.SE","Change.Slope.P","Deviance","DF"))

if (exists("z1.out")) z1.out <- rbind(z1.out,z.res) else z1.out <- z.res
}

MinDev <- min(z1.out[,"Deviance"])
one <- z1.out[z1.out[,"Deviance"] <= MinDev,]

one[1] # jp
one[8] # deviance

return(cbind(one[1],one[8]))
}

########## Simulation loop ##########

for(i in 1:n.sim) {
```

```
# Simulate data according to original model
d$y <- rpois(lenxqr,xqr)

# Quadratic model #
z2 <- glm(y ~ x + x2  ,data=d, family=poisson)
xx <- polyroot(z=c(z2$coef[2],2*z2$coef[3]))
rex2 <- Re(xx)

zz2<- summary(z2)
pvalue2 <- zz2$coefficients["x2",c( "Pr(>|z|)")] # p-value of coef. x2
esb2 <- zz2$coefficients["x2",c( "Estimate")] # estimation of x2
ABS2<- abs(esb2) # absolute value

z.res <- residuals(z2,type="pearson")
z.predict <- predict(z2,type="response")
data  <- cbind(d, resid = z.res, fit = z.predict)
results <- boot(data=data, statistic=bs2, R=500)

CI2 <- quantile(results$t[,1],c(0.025,0.975))
WCI2 <- CI2[2] - CI2[1]

# out2 - Output from quadratic model for one simulation
out2 <- c(rex2,  sd(results$t),CI2[1],CI2[2],
          WCI2, Im(xx),z2$deviance, pvalue2, ABS2)

#########################################################################
# Cubic model #
z3 <- glm(y~  x + x2 +x3 ,data=d,family=poisson)
zz3<- summary(z3)
pvalue3<- zz3$coefficients["x3",c("Pr(>|z|)")]
esb3 <- zz3$coefficients["x3",c("Estimate")]
ABS3<- abs(esb3)

z.a<- z3$coefficients["x3"]
z.b<- z3$coefficients["x2"]
z.c<- z3$coefficients["x"]
z.d <- z3$coefficients["(Intercept)"]
A<- 3*z.a
B<- 2*z.b
C<- z.c

xx<- polyroot(z=c(C,B,A))
srex3 <- Re(xx)
rex3 <- sort(srex3)

z.res <- residuals(z3,type="pearson")
z.predict <- predict(z3,type="response")
data  <- cbind(d, resid = z.res, fit = z.predict)

sresults3 <- boot(data=data, statistic=bs3, R=500)

# to remove tp1=tp2
results3 <- subset (sresults3$t, sresults3$t[,1]!=sresults3$t[,2])


CI13 <- quantile(results3$t[,1],c(0.025, 0.975))
WCI13 <- CI13[2]- CI13[1]
CI23 <- quantile(results3$t[,2],c(0.025, 0.975))
```

```
WCI23 <- CI23[2]- CI23[1]

out3<-  c(rex3[1],sd(results3$t[,1]), CI13[1],CI13[2],
        WCI13,rex3[2],sd(results3$t[,2]),CI23[1],CI23[2],WCI23,
        Im(xx),z3$deviance,pvalue3,ABS3)

#########################################################################
#Segmented models

# Fit SR at true turning points
zs1 <- glm(y ~ x + (ifelse(x<=x0,0,x-x0)),data=d, family=poisson)
zzs1<- summary(zs1)

pvalues1s<- zzs1$coefficients[3,4]

zs2 <- glm(y ~ x + (ifelse(x<=x0-3,0,x-x0+3)),data=d, family=poisson)
zzs2<- summary(zs2)
pvalues2s<- zzs2$coefficients[3,4]

zs3 <- glm(y ~ x + (ifelse(x<=x0+3,0,x-x0-3)),data=d, family=poisson)
zzs3<- summary(zs3)
pvalues3s<- zzs3$coefficients[3,4]

outs <-  c(zs1$deviance,  pvalues1s,
           zs2$deviance,  pvalues2s,
           zs3$deviance ,  pvalues3s)

#########################################################################
# Joinpoint model #
z.int.s <- names(segm)[3:lenxqr]
if (exists("z.out")) rm(z.out)
for (z.index in 1:(length(z.int.s))) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]
z <- glm(y ~ x + get(z.i) ,data=d, family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

z.res <- matrix(c(z.row ,z.dev,z.df),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","Deviance","DF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}

MinDev <- min(z.out[,"Deviance"])
one.CP.Dev <- z.out[z.out[,"Deviance"] <= MinDev,]

jp.dv <- one.CP.Dev[2] # deviance
jp1 <- one.CP.Dev[1] #  jp

# Fit model with estimated jps
z12 <- glm( y ~ x + get(z.int.s[one.CP.Dev[1]-1]),
data=d, family=poisson)
jpz1 <- summary(z12)
pvalujp1 <- jpz1$coefficients[3,4]
es.jp1 <- jpz1$coefficients[3,1]
ABS2z1 <- abs(es.jp1)
```

```
z.res <- residuals(z12,type="pearson")
z.predict <- predict(z12,type="response")
data <- cbind(d, resid = z.res, fit = z.predict)

results1 <- boot(data=data, statistic=bsjp, R=500)

cljp1 <-  quantile(results1$t[,1],c(0.025, 0.975))
wdcljp1 <- cljp1[2]- cljp1[1]

jpres <- c(jp.dv, jp1,  cljp1[1] ,cljp1[2], wdcljp1, pvalujp1, ABS2z1)

#########################################################################
result <- c(out2, out3, outs, jpres)
out1[i,] <- result
gc()

}

dimnames(out1)[[2]]<-c("Q.TP","Q.SD","Q.LCL","Q.UCL",
                       "Q.WDCL","Q.IM","Q.DV","Q.PV","Q.ES.b2",
                       "C.TP1","C.SD1","C.LCL1","C.UCL1","C.WDCL1",
                       "C.TP2","C.SD2","C.LCL2","C.UCL2","C.WDCL2",
                       "C.IM1","C.IM2","C.DV","C.PV", "C.ES.b3",
                       "zs1$deviance",  "pvalues1s",
                       "zs2$deviance",  "pvalues2s",
                       "zs3$deviance" , "pvalues3s",
                       "JP.DV","JP1","Lcljp1","Ucljp1","WDjp1",
                       "pvalujp1", "ABS2z1")

# To remove all t1=t2 in cubic models from simulation samples
OUT1 <- subset(out1, out1[,10]!=out1[,15])
 nrow(OUT1)

out1 <- OUT1

# deviances of models
dev2 <- mean(out1[,7])
dev3 <- mean(out1[,22])
devs <- mean(out1[,25])
devs_3 <- mean(out1[,27])
devs3 <- mean(out1[,29])
devj <- mean(out1[,31])

# to save the main results in vector
Res.1.sim <- matrix(NA, nrow=7, ncol=8)
dimnames(Res.1.sim)[[2]]<- c("DV","SG.CH","TP.ES","TP.SD",
"CI.TP", "CI.WD", "NO.TP", "TP.IN")
dimnames(Res.1.sim)[[1]]<- c("Q.TP","C.TP1","C.TP2","SR","SR-3",
"SR+3", "JP")

# first col. includes the deviance (DV)
Res.1.sim[1,1]<- dev2
Res.1.sim[2,1]<- dev3
Res.1.sim[4,1]<- devs
Res.1.sim[5,1]<- devs_3
Res.1.sim[6,1]<- devs3
Res.1.sim[7,1]<- devj

# second col. includes the (SG.CH%) of the change
```

```
Res.1.sim[1,2]<- sum(out1[,8] <= 0.05)/nrow(OUT1)*100
Res.1.sim[2,2]<- sum(out1[,23] <= 0.05)/nrow(OUT1)*100
Res.1.sim[4,2]<- sum(out1[,26] <= 0.05)/nrow(OUT1)*100 #slpoe
Res.1.sim[5,2]<- sum(out1[,28] <= 0.05)/nrow(OUT1)*100
Res.1.sim[6,2]<- sum(out1[,30] <= 0.05)/nrow(OUT1)*100
Res.1.sim[7,2]<- sum(out1[,36] <= 0.05)/nrow(OUT1)*100 #jp.p-value

# 3rd col. includes the mean of (TP.ES)
Res.1.sim[1,3]<- mean(out1[,1])
Res.1.sim[2,3]<- mean(out1[,10])
Res.1.sim[3,3]<- mean(out1[,15])
Res.1.sim[7,3]<- mean(out1[,32])

# 4th col. includes (TP.SD) of estimated change point
Res.1.sim[1,4]<- mean(out1[,2])
Res.1.sim[2,4]<- mean(out1[,11])
Res.1.sim[3,4]<- mean(out1[,16])
Res.1.sim[7,4]<- sd(out1[,32])

# 5th col. includes no. of CI has x0 (CI.TP1%)
Res.1.sim[1,5]<- sum(x0 >= out1[,3]& x0 <= out1[,4])/ nrow(OUT1)*100
Res.1.sim[2,5]<- sum(x0 >= out1[,12]& x0 <= out1[,13])/ nrow(OUT1)*100
Res.1.sim[3,5]<- sum(x0 >= out1[,17]& x0 <= out1[,18])/ nrow(OUT1)*100
Res.1.sim[7,5]<- sum(x0 >= out1[,33]& x0 <= out1[,34])/ nrow(OUT1)*100


# 6th col. - CI.WD- mean width of CI
Res.1.sim[1,6]<- mean (out1[,5])
Res.1.sim[2,6]<- mean (out1[,14])
Res.1.sim[3,6]<- mean (out1[,19])
Res.1.sim[10,6]<- mean (out1[,35])

# 7th col. - no. of times that the coefficient of change =0 (NO.TP%)
Res.1.sim[1,7]<- sum(out1[,9]<= 10^-6)/ nrow(OUT1)*100
Res.1.sim[2,7]<- sum(out1[,24]<= 10^-6)/ nrow(OUT1)*100
Res.1.sim[7,7]<- sum(out1[,37]<= 10^-6)/ nrow(OUT1)*100

# 8th col. - no. of times tp inside the range of data (TP.IN%)
Res.1.sim[1,8]<- sum(out1[,1] <= lenxqr & out1[,1] >= 1)/ nrow(OUT1)*100
Res.1.sim[2,8]<- sum(out1[,10] <= lenxqr & out1[,10] >= 1)/ nrow(OUT1)*100
Res.1.sim[3,8]<- sum(out1[,15] <= lenxqr & out1[,15] >= 1)/ nrow(OUT1)*100
Res.1.sim[7,8]<- sum(out1[,32] <= lenxqr & out1[,32] >= 1)/ nrow(OUT1)*100

Res.1.sim

QS.50.1.5 <- data.frame (cbind(Res.1.sim))
QS.50.3 <- data.frame (cbind(Res.1.sim))
QS.50.5 <- data.frame (cbind(Res.1.sim))

T.QS.50 <- rbind(QS.50.5, QS.50.3, QS.50.1.5)

# To save table into excel file
write.table(T.QS.50 , "T.QS.50.25.csv",
                sep=",", row.names=FALSE)

# To convert result of table from R to latex file
xtable(T.QS.50,
caption="Combined model, Data points=50, Change at 25",
label="QS.50.25", digits=1)
```

## D.2 R code for simulation study of comparing change points methods in case of two change points

See Section 6.4 for the results.

```
library(XLConnect)
library(boot)
library(xtable)

segm <- readWorksheetFromFile("joinpointDATA-uptoSep2014.xlsx",
sheet=5, header=T)
x <- segm$x

ttp1 <- 16
ttp2 <- 33
para <- function(ttp1,ttp2){
bb1=ttp1*ttp2;
bb2=-0.5*(ttp1+ttp2);
bb3=1/3;
values=c(bb1,bb2,bb3);
return(values);
}
m <- para(ttp1,ttp2)
m1<- m/10000
m2<- round(m1,6)
b0 <- 5  # can be 1.5 ,3 , 5
b1<- m2[1]
b2<- m2[2]
b3<- m2[3]
b4 <- -0.005
b5<- 0.01

# Set the original model (e.g. combined)
xq <- b0+ b1*x + b2*x^2 + b3*x^3 +
      b4*(ifelse(x<=ttp1,0,x-ttp1)) +
      b5*(ifelse(x<=ttp2,0,x-ttp2))

xqr<- exp(xq)
xd <- rpois(length(xq),xqr)
plot(x,log(xd), xlab="Time", ylab="Rate" , lwd=2)
abline(v=ttp1, lwd=2)
abline(v=ttp2, lwd=2)

#########################################################################
n.sim <- 200
out1 <- matrix(NA, nrow=n.sim, ncol=46)
d <- data.frame(cbind(segm))
d$x2 <- d$x^2
d$x3 <- d$x^3
lenxqr <- length(xqr)

########## Bootstrap function to
```

```
#construct CI for quadratic model##########
bs2 # as in code of one change point

########## Bootstrap function to
#construct CI for cubic model##########
bs3 # as in code of one change point

########## Bootstrap function to
#construct CI for two joinpoints ##########
bsjp <- function(data, indices) {
data$new <- round(data$fit +
            data$resid[indices]*sqrt(data$fit),0)
data$new[data$new<0] <- 0

z.int.s <- names(segm)[3:lenxqr]
if (exists("z1.out")) rm(z1.out)

# for first jp
for (z.index in 1: (length(z.int.s)-1)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]
rz1 <- glm(new ~ x + get(z.i)
           ,data=data , family=poisson)
z.dev <- rz1$deviance
z.df <- rz1$df.residual

# for second jp
for (z.index2 in (z.index+1): length(z.int.s)) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("ts","",z.j))
rz2 <- glm(new ~ x + get(z.i) + get(z.j),
            data=data , family=poisson)
z.dev2 <- rz2$deviance
z.df2 <- rz2$df.residual
z2r <- summary(rz2)
z.time <- z2r$coefficients["x",c( "Estimate",
                            "Std. Error","Pr(>|z|)")]
z.change1 <- z2r$coefficients["get(z.i)",c("Estimate",
                            "Std. Error","Pr(>|z|)")]
z.change2 <- z2r$coefficients["get(z.j)",c("Estimate",
                            "Std. Error","Pr(>|z|)")]
z.res <- matrix(c(z.row ,z.row2, z.time, z.change1, z.change2
,z.dev,z.df,z.dev2,z.df2),nrow=1)
dimnames(z.res) <- list(z.i ,c("jp1","jp2", "Slope",
                    "Slope.SE","Slope.P","Change.Slope",
                    "Change.Slope.SE","Change.Slope.P",
                    "Change2.Slope","Change2.Slope.SE",
                    "Change2.Slope.P","Dev.1","DF.1",
                    "Deviance","DF"))

if (exists("z1.out"))  z1.out <- rbind(z1.out,z.res) else z1.out <- z.res
}
}

MinDev <- min(z1.out[,"Deviance"])
two <- z1.out[z1.out[,"Deviance"] <= MinDev,]

two[1] # jp1
```

```
two[2] # jp2
two[14] # deviance

return(cbind(two[1],two[2],two[14]))
}

########## Simulation loop ##########
for(i in 1:n.sim) {
d$y <- rpois(lenxqr,xqr)

#Quadratic model#
out2<- #  as in code of one change point

#####################################################################
#Cubic model#
out3<- #  as in code of one change point

#####################################################################
#Segmented models#

zs1 <- glm(y ~ x +  ts16  +  ts33 ,data=d, family=poisson)
zzs1<- summary(zs1)
pvalues1s1<- zzs1$coefficients[3,4]
pvalues1s2<- zzs1$coefficients[4,4]

zs2 <- glm(y ~ x + ts13 + ts30 ,data=d, family=poisson)
zzs2<- summary(zs2)
pvalues2s1<- zzs2$coefficients[3,4]
pvalues2s2<- zzs2$coefficients[4,4]

zs3 <- glm(y ~ x +  ts19 + ts36 ,data=d, family=poisson)
zzs3<- summary(zs3)
pvalues3s1<- zzs3$coefficients[3,4]
pvalues3s2<- zzs3$coefficients[4,4]

outs<-  c(zs1$deviance,  pvalues1s1,  pvalues1s2,
          zs2$deviance, pvalues2s1,  pvalues2s2,
          zs3$deviance ,  pvalues3s1,  pvalues3s2)

#####################################################################
#Joinpoint model#
z.int.s <- names(segm)[3:lenxqr]
if (exists("z.out")) rm(z.out)

for (z.index in 1:(length(z.int.s)-1)) {
z.i <- z.int.s[z.index]
z.row <- as.numeric(gsub("ts","",z.i))
z.abline <- segm[z.row,"x"]
z <- glm(y ~ x + get(z.i) ,data=d, family=poisson)
z.dev <- z$deviance
z.df <- z$df.residual

for (z.index2 in (z.index+1): length(z.int.s)) {
z.j <- z.int.s[z.index2]
z.row2 <- as.numeric(gsub("ts","",z.j))
z.2 <- glm(y ~ x + get(z.i) + get(z.j), data=d, family=poisson)
z.dev2 <- z.2$deviance
z.df2 <- z.2$df.residual
```

```
z.res <- matrix(c(z.row ,z.row2,z.dev,z.df,z.dev2,z.df2),nrow=1)

dimnames(z.res) <- list(z.i ,c("jp1","jp2","Dev.1",
                    "DF.1","Deviance","DF"))

if (exists("z.out")) z.out <- rbind(z.out,z.res) else z.out <- z.res
}
}

MinDev <- min(z.out[,"Deviance"])
Two.CP.Dev <- z.out[z.out[,"Deviance"] <= MinDev,]

jp.dv <- Two.CP.Dev[5] # deviance
jp1 <- Two.CP.Dev[1] # first jp
jp2 <- Two.CP.Dev[2] # second jp

# Fit model with estimated jps
z12 <- glm( y ~ x + get(z.int.s[Two.CP.Dev[1]-1]) +
        get(z.int.s[Two.CP.Dev[2]-1])
        ,data=d, family=poisson)

zz1<- summary(z12)
pvalue2z1 <- zz1$coefficients[3,4]
pvalue2z2 <- zz1$coefficients[4,4]
esb2z1 <- zz1$coefficients[3,1]
ABS2z1 <- abs(esb2z1)
esb2z2 <- zz1$coefficients[4,1]
ABS2z2 <- abs(esb2z2)

z.res <- residuals(z12,type="pearson")
z.predict <- predict(z12,type="response")
data <- cbind(d, resid = z.res, fit = z.predict)

results1 <- boot(data=data, statistic=bsjp, R=500)

cljp1 <-  quantile(results1$t[,1],c(0.025, 0.975))
wdcljp1 <- cljp1[2]- cljp1[1]
cljp2 <-  quantile(results1$t[,2],c(0.025, 0.975))
wdcljp2 <- cljp2[2]- cljp2[1]

jpres <- c(jp.dv, jp1,  cljp1[1] ,cljp1[2], wdcljp1, pvalue2z1, ABS2z1,
           jp2 , cljp2[1] , cljp2[2], wdcljp2 , pvalue2z2 , ABS2z2 )
#########################################################################
result <- c(out2, out3, outs, jpres )
out1[i,] <- result
gc()
}

dimnames(out1)[[2]]<-c("Q.TP","Q.SD","Q.LCL","Q.UCL",
                    "Q.WDCL","Q.IM","Q.DV","Q.PV","Q.ES.b2",
                    "C.TP1","C.SD1","C.LCL1","C.UCL1","C.WDCL1",
                    "C.TP2","C.SD2","C.LCL2","C.UCL2","C.WDCL2",
                    "C.IM1","C.IM2","C.DV","C.PV", "C.ES.b3",
                    "zs1.deviance",  "pvalues1s1","pvalues1s2",
                    "zs2_3deviance", "pvalues2s1", "pvalues2s2",
                    "zs3+3deviance" , "pvalues3s1",  "pvalues3s2",
                    "JP.DV","JP1","Lcljp1","Ucljp1","WDjp1",
                    "pvalujp1", "ABS2z1","JP2","Lcljp2","Ucljp2",
                    "WDjp2", "pvalujp2", "ABS2z2")
```

```
# To remove all t1=t2 from cubic results in simulation study
OUT1 <- subset(out1, out1[,10]!=out1[,15])
nrow(OUT1)
out1 < OUT1

# deviances of models
dev2 <- mean(out1[,7])
dev3 <- mean(out1[,22])
devs <- mean(out1[,25])
devs_3 <- mean(out1[,28])
devs3 <- mean(out1[,31])
devj <- mean(out1[,34])

Res.1.sim <- matrix(NA, nrow=11, ncol=11)

dimnames(Res.1.sim)[[2]]<- c("DV","SG.CH","TP.ES","TP.SD","LCL","UCL",
                             "CI.WD","CI.TP1", "CI.TP2", "NO.TP", "TP.IN")
dimnames(Res.1.sim)[[1]]<- c("Q.TP","C.TP1","C.TP2",
                             "SR1","SR2","SR1 -3","SR2 -3",
                             "SR1 +3","SR2 +3", "JP1", "JP2")

# first col. includes the deviance (DV)
Res.1.sim[1,1]<- dev2
Res.1.sim[2,1]<- dev3
Res.1.sim[4,1]<- devs
Res.1.sim[6,1]<- devs_3
Res.1.sim[8,1]<- devs3
Res.1.sim[10,1]<- devj

# second col. includes the (SG.CH%) of the change
Res.1.sim[1,2]<- sum(out1[,8] <= 0.05)/nrow(OUT1)*100
Res.1.sim[2,2]<- sum(out1[,23] <= 0.05)/nrow(OUT1)*100
Res.1.sim[4,2]<- sum(out1[,26] <= 0.05)/nrow(OUT1)*100 #slope
Res.1.sim[5,2]<- sum(out1[,27] <= 0.05)/nrow(OUT1)*100 #slope
Res.1.sim[6,2]<- sum(out1[,29] <= 0.05)/nrow(OUT1)*100
Res.1.sim[7,2]<- sum(out1[,30] <= 0.05)/nrow(OUT1)*100 #slope
Res.1.sim[8,2]<- sum(out1[32] <= 0.05)/nrow(OUT1)*100 #slope
Res.1.sim[9,2]<- sum(out1[,33] <= 0.05)/nrow(OUT1)*100
Res.1.sim[10,2]<- sum(out1[,39] <= 0.05)/nrow(OUT1)*100 #slope
Res.1.sim[11,2]<- sum(out1[,45] <= 0.05)/nrow(OUT1)*100 #slope

# 3rd col. includes the mean of (TP.ES)
Res.1.sim[1,3]<- mean(out1[,1])
Res.1.sim[2,3]<- mean(out1[,10])
Res.1.sim[3,3]<- mean(out1[,15])
Res.1.sim[10,3]<- mean(out1[,35])
Res.1.sim[11,3]<- mean(out1[,41])

# 4th col. includes (TP.SD) of estimated change point
Res.1.sim[1,4]<- mean(out1[,2])
Res.1.sim[2,4]<- mean(out1[,11])
Res.1.sim[3,4]<- mean(out1[,16])
Res.1.sim[10,4]<- sd(out1[,35])
Res.1.sim[11,4]<- sd(out1[,41])

# 5th col. - the mean of lower CL
Res.1.sim[1,5] <- mean(out1[,3])
Res.1.sim[2,5] <- mean(out1[,12])
```

```
Res.1.sim[3,5] <- mean(out1[,17])
Res.1.sim[10,5] <- mean(out1[,36])
Res.1.sim[11,5] <- mean(out1[,42])

# 6th col. - the mean of upper CL
Res.1.sim[1,6] <- mean(out1[,4])
Res.1.sim[2,6] <- mean(out1[,13])
Res.1.sim[3,6] <- mean(out1[,18])
Res.1.sim[10,6] <- mean(out1[,37])
Res.1.sim[11,6] <- mean(out1[,43])

# 7th col. - CI.WD- mean width of CI
Res.1.sim[1,7]<- mean (out1[,5])
Res.1.sim[2,7]<- mean (out1[,14])
Res.1.sim[3,7]<- mean (out1[,19])
Res.1.sim[10,7]<- mean (out1[,38])
Res.1.sim[11,7]<- mean (out1[,44])

# 8th col. includes no. of CI has ttp1 (CI.TP1%)
Res.1.sim[1,8]<- sum(ttp1 >= out1[,3]& ttp1 <= out1[,4])/ nrow(OUT1)*100
Res.1.sim[2,8]<- sum(ttp1 >= out1[,12]& ttp1 <= out1[,13])/ nrow(OUT1)*100
Res.1.sim[10,8]<- sum(ttp1 >= out1[,36]& ttp1 <= out1[,37])/ nrow(OUT1)*100
Res.1.sim[11,8]<- sum(ttp1 >= out1[,42]& ttp1 <= out1[,43])/ nrow(OUT1)*100

# 9th col. includes no. of CI has ttp2 (CI.TP2%)
Res.1.sim[1,9]<- sum(ttp2 >= out1[,3]& ttp2 <= out1[,4])/ nrow(OUT1)*100
Res.1.sim[3,9]<- sum(ttp2 >= out1[,17]& ttp2 <= out1[,18])/ nrow(OUT1)*100
Res.1.sim[10,9]<- sum(ttp2 >= out1[,36]& ttp2 <= out1[,37])/ nrow(OUT1)*100
Res.1.sim[11,9]<- sum(ttp2 >= out1[,42]& ttp2 <= out1[,43])/ nrow(OUT1)*100

# 10th col. - no. of times that the coef of change =0 (NO.TP%)
Res.1.sim[1,10]<- sum(out1[,9]<= 10^-6)/ nrow(OUT1)*100
Res.1.sim[2,10]<- sum(out1[,24]<= 10^-6)/ nrow(OUT1)*100
Res.1.sim[10,10]<- sum(out1[,40]<= 10^-6)/ nrow(OUT1)*100
Res.1.sim[11,10]<- sum(out1[,46]<= 10^-6)/ nrow(OUT1)*100

# 11th co. - no. of times tp inside the range of data (TP.IN%)
Res.1.sim[1,11]<- sum(out1[,1] <= lenxqr & out1[,1] >= 1)/ nrow(OUT1)*100
Res.1.sim[2,11]<- sum(out1[,10] <= lenxqr & out1[,10] >= 1)/ nrow(OUT1)*100
Res.1.sim[3,11]<- sum(out1[,15] <= lenxqr & out1[,15] >= 1)/ nrow(OUT1)*100
Res.1.sim[10,11]<- sum(out1[,35] <= lenxqr & out1[,35] >= 1)/ nrow(OUT1)*100
Res.1.sim[11,11]<- sum(out1[,41] <= lenxqr & out1[,41] >= 1)/ nrow(OUT1)*100

Res.1.sim

CS2.50.1.5 <- data.frame (cbind(Res.1.sim))
CS2.50.3 <- data.frame (cbind(Res.1.sim))
CS2.50.5 <- data.frame (cbind(Res.1.sim))
T.CS2.50 <- rbind(CS2.50.5, CS2.50.3, CS2.50.1.5)

# To save results in Excel.csv file
write.table(T.CS2.50 , "T.CS2.50.16-33.csv",
                sep=",", row.names=FALSE)

# To convert R code to latex
xtable(T.CS2.50,
caption="Combined model, Data points=50, Change in 16 and 33",
label="CS2.50.16-33", digits=1)
```

## D.3 Simulation study with no change points

### D.3.1 Original model with no change and the slope $\beta_1 = 0.001$

Table D.1: Simulation study on linear model with slope $\beta_1 = 0.001$.

| Number of data points $n = 50$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 47.1 | 3.5 | 9.2 | 468.0 | 228.0 | 1.5 | 73.0 |
| | C.TP1 | 46.0 | 7.0 | <1 | 154.4 | 112.7 | 14.5 | 90.5 |
| | C.TP2 | | | >50 | 336.5 | 109.1 | | 82.5 |
| | JP | 44.7 | 25.5 | 25.7 | 18.9 | 46.0 | 0.0 | 100.0 |
| 3 | Q.TP | 47.0 | 2.5 | 36.8 | 947.2 | 182.7 | 0.5 | 83.5 |
| | C.TP1 | 45.8 | 8.5 | 7.5 | 151.2 | 104.5 | 4.0 | 87.5 |
| | C.TP2 | | | >50 | 185.1 | 134.0 | | 81.0 |
| | JP | 44.7 | 26.0 | 26.7 | 18.7 | 45.7 | 0.0 | 100.0 |
| 1.5 | Q.TP | 49.8 | 4.5 | 21.7 | 902.7 | 157.3 | 0.0 | 85.0 |
| | C.TP1 | 48.9 | 2.5 | 8.8 | 402.1 | 103.1 | 0.5 | 89.5 |
| | C.TP2 | | | >50 | 112.6 | 126.0 | | 88.5 |
| | JP | 47.5 | 23.0 | 26.0 | 18.2 | 45.5 | 0.0 | 100.0 |
| Number of data points $n = 35$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 31.7 | 6.0 | 24.2 | 375.3 | 120.5 | 0.5 | 77.5 |
| | C.TP1 | 30.6 | 3.5 | 4.2 | 93.4 | 78.2 | 6.0 | 83.5 |
| | C.TP2 | | | 32.2 | 545.3 | 76.7 | | 96.0 |
| | JP | 29.8 | 24.5 | 17.5 | 12.0 | 31.1 | 0.0 | 100.0 |
| 3 | Q.TP | 31.4 | 3.5 | 20.7 | 684.3 | 110.3 | 0.0 | 84.0 |
| | C.TP1 | 30.5 | 4.0 | <1 | 57.3 | 79.5 | 3.0 | 85.5 |
| | C.TP2 | | | 30.1 | 339.4 | 74.1 | | 93.0 |
| | JP | 29.5 | 22.5 | 18.6 | 12.1 | 31.3 | 0.0 | 100.0 |
| 1.5 | Q.TP | 34.2 | 6.0 | 18.9 | 651.6 | 107.9 | 0.0 | 84.0 |
| | C.TP1 | 33.2 | 6.0 | 6.6 | 649.1 | 80.7 | 0.0 | 85.5 |
| | C.TP2 | | | >35 | 51.7 | 73.0 | | 94.5 |
| | JP | 31.9 | 24.0 | 18.9 | 12.6 | 30.3 | 0.0 | 100.0 |
| Number of data points $n = 20$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 17.0 | 2.5 | 14.8 | 204.8 | 69.0 | 0.0 | 86.0 |
| | C.TP1 | 16.1 | 5.0 | 2.4 | 40.3 | 45.7 | 0.0 | 86.5 |
| | C.TP2 | | | 18.0 | 34.9 | 45.8 | | 98.0 |
| | JP | 15.4 | 21.5 | 10.9 | 6.5 | 16.5 | 0.0 | 100.0 |
| 3 | Q.TP | 17.7 | 4.5 | 19.8 | >1000 | 70.2 | 0.0 | 76.5 |
| | C.TP1 | 16.6 | 6.0 | <1 | 53.5 | 43.4 | 0.0 | 82.0 |
| | C.TP2 | | | 17.3 | 77.5 | 49.8 | | 97.5 |
| | JP | 16.0 | 19.0 | 11.0 | 6.7 | 16.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 18.1 | 3.0 | 5.6 | 241.9 | 74.4 | 0.0 | 76.5 |
| | C.TP1 | 17.0 | 6.0 | 4.1 | 36.5 | 44.5 | 0.0 | 90.0 |
| | C.TP2 | | | 16.7 | 72.3 | 46.8 | | 99.0 |
| | JP | 16.3 | 22.0 | 11.1 | 6.5 | 16.4 | 0.0 | 100.0 |

See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, CI.WD, NO.TP%, TP.IN%**.

### D.3.2 Original model with no change and the slope $\beta_1 = 0.008$

Table D.2: Simulation study on linear model with slope $\beta_1 = 0.008$.

| Number of data point $n = 50$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 46.0 | 4.0 | >50 | >1000 | >1000 | 0.5 | 0.0 |
| | C.TP1 | 44.9 | 5.5 | 3.7 | 134.1 | 139.4 | 17.5 | 53.0 |
| | C.TP2 | | | 47.1 | 147.5 | 149.5 | | 52.0 |
| | JP | 43.8 | 20.5 | 24.9 | 18.8 | 45.5 | 0.0 | 100.0 |
| 3 | Q.TP | 48.3 | 5.5 | >50 | >1000 | 596.6 | 0.5 | 31.5 |
| | C.TP1 | 47.2 | 5.5 | <1 | 127.7 | 163.0 | 4.5 | 73.0 |
| | C.TP2 | | | >50 | 100.8 | 119.5 | | 74.5 |
| | JP | 46.0 | 23.5 | 25.5 | 18.2 | 45.7 | 0.0 | 100.0 |
| 1.5 | Q.TP | 50.0 | 6.0 | >50 | 770.2 | 284.6 | 0.5 | 65.0 |
| | C.TP1 | 48.8 | 5.5 | 11.3 | 283.6 | 151.4 | 2.5 | 87.0 |
| | C.TP2 | | | 39.9 | 133.1 | 135.6 | | 83.0 |
| | JP | 47.4 | 25.5 | 28.9 | 18.2 | 45.4 | 0.0 | 100.0 |
| Number of data point $n = 35$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 32.5 | 6.0 | >35 | >1000 | 638.7 | 1.0 | 11.5 |
| | C.TP1 | 31.5 | 6.0 | <1 | 227.1 | 87.0 | 6.0 | 77.5 |
| | C.TP2 | | | 28.9 | 137.1 | 94.7 | | 93.5 |
| | JP | 30.6 | 28.0 | 18.3 | 12.2 | 31.4 | 0.0 | 100.0 |
| 3 | Q.TP | 32.9 | 4.0 | 28.2 | >1000 | 239.1 | 0.0 | 57.5 |
| | C.TP1 | 31.7 | 7.0 | 8.9 | 72.1 | 78.8 | 1.5 | 84.0 |
| | C.TP2 | | | 30.3 | 86.7 | 84.2 | | 91.0 |
| | JP | 30.9 | 28.0 | 17.8 | 11.9 | 31.1 | 0.0 | 100.0 |
| 1.5 | Q.TP | 32.8 | 3.0 | 37.4 | 564.9 | 145.1 | 0.0 | 77.5 |
| | C.TP1 | 31.8 | 3.5 | <1 | 178.5 | 76.4 | 2.0 | 88.5 |
| | C.TP2 | | | >35 | 185.7 | 78.1 | | 94.5 |
| | JP | 30.9 | 19.5 | 19.3 | 12.2 | 31.1 | 0.0 | 100.0 |
| Number of data point $n = 20$ | | | | | | | | |
| $\beta_0$ | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.WD | NO.TP% | TP.IN% |
| 5 | Q.TP | 17.3 | 6.5 | <1 | 514.8 | 158.5 | 0.5 | 49.5 |
| | C.TP1 | 16.4 | 3.0 | 2.9 | 194.1 | 53.8 | 1.5 | 82.0 |
| | C.TP2 | | | 14.4 | 197.0 | 50.8 | | 99.0 |
| | JP | 15.7 | 18.5 | 10.7 | 6.4 | 16.4 | 0.0 | 100.0 |
| 3 | Q.TP | 17.2 | 7.5 | <1 | 214.6 | 77.7 | 0.0 | 72.5 |
| | C.TP1 | 16.1 | 6.0 | 5.4 | >1000 | 48.8 | 0.0 | 90.0 |
| | C.TP2 | | | >20 | 88.2 | 49.7 | | 96.5 |
| | JP | 15.4 | 17.0 | 10.6 | 6.5 | 16.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 17.8 | 6.0 | 15.2 | 342.0 | 71.7 | 0.0 | 77.5 |
| | C.TP1 | 16.7 | 5.5 | 3.1 | 66.9 | 45.8 | 0.5 | 87.5 |
| | C.TP2 | | | >20 | 57.9 | 44.6 | | 97.5 |
| | JP | 16.0 | 16.5 | 10.3 | 6.6 | 16.0 | 0.0 | 100.0 |

See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, CI.WD, NO.TP%, TP.IN%**.

## D.4 Simulation study with one change point

### D.4.1 Change occurs in the middle of dataset

#### Number of data points=35 and change in middle at 18

Table D.3: Number of data points=35 and the true turning point in the middle at 18.

| | | | | | $\beta_0 = 5$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 32.7 | 100.0 | 18.2 | 0.3 | 86.0 | 1.1 | 0.0 | 100.0 |
| ratic | C.TP1 | 31.7 | 4.0 | <1 | >1000 | 70.0 | 964.9 | 4.0 | 55.0 |
| | C.TP2 | | | >35 | >1000 | 39.0 | >1000 | | 45.0 |
| | SR | 48.3 | 100.0 | | | | | | |
| | SR-3 | 52.8 | 100.0 | | | | | | |
| | SR+3 | 55.2 | 100.0 | | | | | | |
| | JP | 46.8 | 100.0 | 18.0 | 1.7 | 74.0 | 3.0 | 0.0 | 100.0 |
| Segm- | Q.TP | 34.0 | 100.0 | 13.9 | 10.4 | 1.0 | 5.0 | 0.0 | 100.0 |
| ented | C.TP1 | 32.9 | 8.0 | <1 | 574.9 | 0.0 | 406.2 | 6.0 | 57.0 |
| with | C.TP2 | | | >35 | 864.3 | 98.0 | 412.0 | | 64.0 |
| one | SR | 31.1 | 100.0 | | | | | | |
| change | SR-3 | 31.5 | 100.0 | | | | | | |
| point | SR+3 | 32.6 | 100.0 | | | | | | |
| | JP | 30.8 | 100.0 | 17.4 | 3.0 | 96.0 | 9.4 | 0.0 | 100.0 |
| Comb- | Q.TP | 36.5 | 100.0 | 15.3 | 0.3 | 0.0 | 1.0 | 0.0 | 100.0 |
| ined | C.TP1 | 35.5 | 5.0 | <1 | >1000 | 0.0 | >1000 | 3.0 | 43.0 |
| | C.TP2 | | | >35 | >1000 | 95.0 | >1000 | | 57.0 |
| | SR | 46.2 | 100.0 | | | | | | |
| | SR-3 | 53.1 | 100.0 | | | | | | |
| | SR+3 | 69.0 | 100.0 | | | | | | |
| | JP | 45.1 | 100.0 | 17.6 | 0.9 | 87.0 | 2.0 | 0.0 | 100.0 |
| | | | | | $\beta_0 = 3$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 32.0 | 100.0 | 18.3 | 0.8 | 92.5 | 3.1 | 0.0 | 100.0 |
| ratic | C.TP1 | 31.0 | 7.5 | <1 | 298.8 | 65.0 | 331.2 | 0.5 | 53.0 |
| | C.TP2 | | | >35 | 729.3 | 53.0 | 362.0 | | 66.0 |
| | SR | 33.9 | 100.0 | | | | | | |
| | SR-3 | 34.5 | 100.0 | | | | | | |
| | SR+3 | 34.7 | 100.0 | | | | | | |
| | JP | 32.1 | 100.0 | 18.1 | 4.4 | 67.5 | 8.2 | 0.0 | 100.0 |
| Segm- | Q.TP | 33.2 | 66.0 | 14.0 | 209.0 | 46.5 | 62.2 | 0.0 | 89.5 |
| ented | C.TP1 | 32.2 | 4.0 | <1 | 466.6 | 9.0 | 185.7 | 1.5 | 71.0 |
| with | C.TP2 | | | >35 | 188.8 | 94.0 | 182.9 | | 83.0 |
| one | SR | 31.7 | 69.5 | | | | | | |
| change | SR-3 | 31.9 | 64.0 | | | | | | |
| point | SR+3 | 31.9 | 49.0 | | | | | | |
| | JP | 31.1 | 85.5 | 17.8 | 7.5 | 96.5 | 24.5 | 0.0 | 100.0 |
| Comb- | Q.TP | 34.3 | 100.0 | 15.7 | 2.5 | 38.5 | 8.7 | 0.0 | 97.0 |
| ined | C.TP1 | 32.5 | 13.0 | <1 | 832.7 | 0.0 | 498.0 | 1.0 | 50.0 |
| | C.TP2 | | | >35 | 638.3 | 98.0 | 513.1 | | 54.0 |
| | SR | 35.0 | 98.0 | | | | | | |

*Continued on next page*

Table D.3 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| | SR-3 | 38.9 | 66.5 | | | | | | |
| | SR+3 | 34.4 | 100.0 | | | | | | |
| | JP | 32.9 | 100.0 | 17.4 | 3.9 | 62.0 | 8.1 | 0.0 | 100.0 |
| | | | | | $\beta_0 = 1.5$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 34.9 | 87.5 | 19.7 | 20.9 | 96.0 | 14.0 | 0.0 | 99.5 |
| ratic | C.TP1 | 33.9 | 5.0 | <1 | 176.9 | 64.0 | 204.4 | 0.5 | 61.0 |
| | C.TP2 | | | >35 | 458.2 | 52.0 | 222.1 | | 82.0 |
| | SR | 34.7 | 86.5 | | | | | | |
| | SR-3 | 34.8 | 78.0 | | | | | | |
| | SR+3 | 35.3 | 74.5 | | | | | | |
| | JP | 32.9 | 89.5 | 17.6 | 7.7 | 85.5 | 18.8 | 0.0 | 100.0 |
| Segm- | Q.TP | 35.1 | 23.0 | 16.9 | 310.7 | 91.0 | 98.6 | 0.0 | 87.0 |
| ented | C.TP1 | 34.0 | 4.0 | 3.8 | 326.0 | 46.0 | 112.8 | 1.0 | 83.0 |
| with | C.TP2 | | | >35 | 110.9 | 86.0 | 124.3 | | 86.0 |
| one | SR | 33.9 | 25.5 | | | | | | |
| change | SR-3 | 34.0 | 20.5 | | | | | | |
| point | SR+3 | 34.0 | 18.5 | | | | | | |
| | JP | 33.0 | 33.0 | 19.2 | 10.3 | 96.5 | 29.4 | 0.0 | 100.0 |
| Comb- | Q.TP | 36.0 | 96.5 | 15.1 | 10.4 | 45.0 | 9.6 | 0.0 | 100.0 |
| ined | C.TP1 | 35.1 | 3.5 | <1 | 292.1 | 7.0 | 243.8 | 0.0 | 55.0 |
| | C.TP2 | | | >35 | 216.9 | 91.0 | 196.5 | | 82.0 |
| | SR | 35.2 | 96.5 | | | | | | |
| | SR-3 | 35.4 | 90.5 | | | | | | |
| | SR+3 | 36.1 | 84.0 | | | | | | |
| | JP | 34.1 | 96.0 | 17.5 | 5.3 | 91.5 | 17.0 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

## Number of data points=20 and change in middle at 10

Table D.4: Number of data points=20 and the true turning point in the middle at 10.

| | | | | | $\beta_0 = 5$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 17.3 | 100.0 | 10.1 | 0.4 | 93.0 | 1.6 | 0.0 | 100.0 |
| ratic | C.TP1 | 16.3 | 4.0 | <1 | 269.7 | 64.0 | 314.9 | 1.5 | 50.0 |
| | C.TP2 | | | >20 | 401.1 | 48.0 | 235.4 | | 77.0 |
| | SR | 18.7 | 100.0 | | | | | | |
| | SR-3 | 21.5 | 99.5 | | | | | | |
| | SR+3 | 21.5 | 99.0 | | | | | | |
| | JP | 17.5 | 100.0 | 10.5 | 2.2 | 78.0 | 4.2 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.1 | 75.0 | 4.0 | 37.5 | 52.0 | 26.5 | 0.0 | 96.5 |
| ented | C.TP1 | 15.9 | 6.5 | <1 | 181.0 | 13.0 | 106.6 | 0.0 | 79.0 |
| with | C.TP2 | | | >20 | 280.4 | 97.0 | 139.5 | | 85.0 |
| one | SR | 15.7 | 76.0 | | | | | | |
| change | SR-3 | 16.1 | 44.5 | | | | | | |
| point | SR+3 | 16.4 | 40.0 | | | | | | |

*Continued on next page*

Table D.4 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| | JP | 15.0 | 84.5 | 10.6 | 3.0 | 93.0 | 11.7 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.1 | 100.0 | 8.1 | 0.4 | 0.0 | 1.5 | 0.0 | 100.0 |
| ined | C.TP1 | 17.1 | 4.0 | <1 | 563.2 | 0.0 | 367.6 | 0.5 | 48.0 |
| | C.TP2 | | | >20 | >1000 | 96.0 | 665.1 | | 3.0 |
| | SR | 17.7 | 100.0 | | | | | | |
| | SR-3 | 22.8 | 100.0 | | | | | | |
| | SR+3 | 26.3 | 100.0 | | | | | | |
| | JP | 17.2 | 100.0 | 10.1 | 1.2 | 86.5 | 2.8 | 0.0 | 100 |
| | | | | | $\beta_0 = 3$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 17.1 | 98.0 | 10.0 | 2.1 | 93.5 | 3.5 | 0.0 | 100.0 |
| ratic | C.TP1 | 16.2 | 3.5 | <1 | 219.1 | 58.0 | 167.3 | 0.5 | 53.0 |
| | C.TP2 | | | >20 | 298.7 | 50.0 | 138.5 | | 89.0 |
| | SR | 17.6 | 98.0 | | | | | | |
| | SR-3 | 18.5 | 84.0 | | | | | | |
| | SR+3 | 18.4 | 85.0 | | | | | | |
| | JP | 16.0 | 99.5 | 10.5 | 3.3 | 77.0 | 7.4 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.4 | 34.0 | 4.5 | 234.8 | 81.0 | 73.0 | 0.0 | 81.5 |
| ented | C.TP1 | 16.4 | 6.0 | <1 | 106.0 | 34.0 | 72.6 | 0.0 | 77.0 |
| with | C.TP2 | | | >20 | 106.7 | 87.0 | 78.9 | | 90.0 |
| one | SR | 16.4 | 35.0 | | | | | | |
| change | SR-3 | 16.6 | 20.5 | | | | | | |
| point | SR+3 | 16.5 | 18.0 | | | | | | |
| | JP | 15.5 | 55.5 | 10.3 | 5.4 | 93.5 | 14.5 | 0.0 | 100.0 |
| Comb- | Q.TP | 17.8 | 89.0 | 7.9 | 18.3 | 40.0 | 10.5 | 0.0 | 99.5 |
| ined | C.TP1 | 16.7 | 6.0 | <1 | 133.3 | 4.0 | 122.0 | 0.5 | 63.05 |
| | C.TP2 | | | >20 | 139.7 | 88.0 | 119.3 | | 88.0 |
| | SR | 17.2 | 90.0 | | | | | | |
| | SR-3 | 17.9 | 70.0 | | | | | | |
| | SR+3 | 18.2 | 50.5 | | | | | | |
| | JP | 15.9 | 94.5 | 10.1 | 3.6 | 86.0 | 9.5 | 0.0 | 100.0 |
| | | | | | $\beta_0 = 1.5$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 18.2 | 54.0 | 10.1 | 28.0 | 92.0 | 18.5 | 0.0 | 99.5 |
| ratic | C.TP1 | 17.3 | 3.0 | <1 | 113.7 | 67.0 | 76.7 | 0.0 | 73.0 |
| | C.TP2 | | | >20 | 234.0 | 49.0 | 104.4 | | 95.0 |
| | SR | 17.5 | 54.0 | | | | | | |
| | SR-3 | 17.6 | 33.0 | | | | | | |
| | SR+3 | 18.1 | 25.5 | | | | | | |
| | JP | 16.2 | 64.5 | 10.7 | 4.9 | 89.5 | 19.0 | 0.0 | 100.0 |
| Segm- | Q.TP | 18.2 | 13.0 | 7.1 | 298.9 | 90.0 | 62.0 | 0.0 | 82.0 |
| ented | C.TP1 | 17.1 | 7.0 | 2.9 | 63.4 | 60.0 | 57.0 | 0.0 | 88.0 |
| with | C.TP2 | | | >20 | 146.6 | 75.0 | 54.3 | | 95.0 |
| one | SR | 17.0 | 12.0 | | | | | | |
| change | SR-3 | 17.2 | 10.0 | | | | | | |
| point | SR+3 | 17.2 | 7.5 | | | | | | |
| | JP | 16.3 | 25.5 | 10.8 | 5.7 | 97.0 | 15.6 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.3 | 33.0 | 8.0 | 231.4 | 91.0 | 44.5 | 0.0 | 88.5 |
| ined | C.TP1 | 17.1 | 5.0 | <1 | 100.6 | 35.0 | 75.5 | 0.0 | 79.0 |
| | C.TP2 | | | 17.9 | 63.1 | 90.0 | 69.8 | | 96.0 |
| | SR | 17.3 | 30.5 | | | | | | |
| | SR-3 | 17.5 | 15.5 | | | | | | |
| | SR+3 | 17.7 | 17.0 | | | | | | |

*Continued on next page*

Table D.4 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| | JP | 16.3 | 41.5 | 10.3 | 5.5 | 89.5 | 14.0 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

## D.4.2 Change occurs in the beginning of dataset

### Number of data points=35 and change in beginning at 9

Table D.5: Number of data points=35 and the true turning point in the beginning at 9.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 32.6 | 100.0 | 9.1 | 0.6 | 90.0 | 2.3 | 0.0 | 100.0 |
| ratic | C.TP1 | 31.4 | 4.0 | <1 | >1000 | 84.0 | 960.3 | 2.0 | 46.0 |
| | C.TP2 | | | >35 | >1000 | 81.0 | >1000 | | 54.0 |
| | SR | 73.5 | 100.0 | | | | | | |
| | SR-3 | 106.2 | 100.0 | | | | | | |
| | SR+3 | 54.1 | 100.0 | | | | | | |
| | JP | 44.0 | 100.0 | 16.8 | 2.0 | 1.0 | 3.6 | 0.0 | 100.0 |
| Segm- | Q.TP | 33.2 | 46.0 | <1 | >1000 | 94.0 | 355.4 | 1.0 | 45.0 |
| ented | C.TP1 | 30.8 | 21.0 | 4.9 | 260.1 | 91.0 | 77.9 | 0.0 | 87.0 |
| with | C.TP2 | | | >35 | 596.5 | 59.0 | 124.3 | | 79.0 |
| one | SR | 30.6 | 45.0 | | | | | | |
| change | SR-3 | 30.9 | 17.0 | | | | | | |
| point | SR+3 | 31.0 | 54.0 | | | | | | |
| | JP | 30.1 | 83.0 | 11.8 | 8.2 | 98.0 | 25.6 | 0.0 | 100.0 |
| Comb- | Q.TP | 33.7 | 100.0 | 7.5 | 0.6 | 28.0 | 2.5 | 0.0 | 100.0 |
| ined | C.TP1 | 31.8 | 18.0 | <1 | >1000 | 13.0 | 808.6 | 0.0 | 81.0 |
| | C.TP2 | | | >35 | >1000 | 77.0 | 852.3 | | 19.0 |
| | SR | 63.9 | 100.0 | | | | | | |
| | SR-3 | 93.9 | 100.0 | | | | | | |
| | SR+3 | 50.4 | 100.0 | | | | | | |
| | JP | 46.6 | 100.0 | 14.9 | 1.9 | 0.0 | 3.3 | 0.0 | 100.0 |
| $\beta_0 = 3$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 32.1 | 100.0 | 8.7 | 3.5 | 93.5 | 8.2 | 0.0 | 99.5 |
| ratic | C.TP1 | 31.1 | 5.0 | <1 | 570.1 | 98.0 | 392.7 | 1.0 | 59.0 |
| | C.TP2 | | | >35 | 506.0 | 86.0 | 410.8 | | 60.0 |
| | SR | 37.3 | 86.0 | | | | | | |
| | SR-3 | 41.0 | 36.5 | | | | | | |
| | SR+3 | 34.6 | 99.0 | | | | | | |
| | JP | 31.6 | 100.0 | 17.0 | 5.1 | 33.5 | 10.3 | 0.0 | 100.0 |
| Segm- | Q.TP | 34.4 | 26.0 | 19.2 | >1000 | 100.0 | 406.6 | 0.0 | 47.0 |
| ented | C.TP1 | 32.8 | 12.5 | <1 | 63.2 | 97.0 | 84.7 | 0.5 | 86.0 |
| with | C.TP2 | | | 33.8 | 80.6 | 69.0 | 108.5 | | 91.0 |
| one | SR | 32.5 | 23.0 | | | | | | |
| change | SR-3 | 32.8 | 9.5 | | | | | | |

*Continued on next page*

Table D.5 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| point | SR+3 | 32.9 | 27.0 | | | | | | |
| | JP | 31.8 | 58.5 | 14.9 | 10.7 | 96.5 | 28.7 | 0.0 | 100.0 |
| Comb- | Q.TP | 32.6 | 100.0 | 7.4 | 2.2 | 81.5 | 7.8 | 0.0 | 99.5 |
| ined | C.TP1 | 31.5 | 4.0 | <1 | 582.3 | 84.0 | 367.9 | 1.0 | 70.0 |
| | C.TP2 | | | >35 | >1000 | 88.0 | 497.9 | | 49.0 |
| | SR | 35.8 | 97.5 | | | | | | |
| | SR-3 | 40.0 | 51.0 | | | | | | |
| | SR+3 | 34.0 | 100.0 | | | | | | |
| | JP | 31.9 | 100.0 | 14.9 | 4.2 | 40.0 | 9.2 | 0.0 | 100.0 |
| | | | | $\beta_0$ = 1.5 | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 34.5 | 70.5 | 8.1 | 252.7 | 96.5 | 69.1 | 0.0 | 93.0 |
| ratic | C.TP1 | 33.5 | 5.0 | <1 | 222.3 | 100.0 | 168.1 | 1.0 | 64.0 |
| | C.TP2 | | | >35 | 262.0 | 66.0 | 230.3 | | 80.0 |
| | SR | 35.0 | 32.0 | | | | | | |
| | SR-3 | 35.9 | 10.5 | | | | | | |
| | SR+3 | 34.2 | 55.0 | | | | | | |
| | JP | 32.3 | 75.5 | 17.4 | 8.4 | 76.5 | 22.5 | 0.0 | 100.0 |
| Segm- | Q.TP | 36.7 | 8.5 | <1 | 691.8 | 99.0 | 219.5 | 0.5 | 54.0 |
| ented | C.TP1 | 35.3 | 11.0 | 3.8 | 60.0 | 99.0 | 91.6 | 0.5 | 90.0 |
| with | C.TP2 | | | >35 | 90.4 | 41.0 | 90.0 | | 93.0 |
| one | SR | 35.4 | 12.0 | | | | | | |
| change | SR-3 | 35.3 | 8.0 | | | | | | |
| point | SR+3 | 35.5 | 11.0 | | | | | | |
| | JP | 34.2 | 26.0 | 18.0 | 12.5 | 95.5 | 29.9 | 0.0 | 100.0 |
| Comb- | Q.TP | 34.8 | 73.5 | 5.9 | 307.9 | 96.5 | 83.6 | 0.0 | 88.0 |
| ined | C.TP1 | 33.7 | 5.0 | <1 | 341.8 | 99.0 | 175.2 | 1.0 | 78.0 |
| | C.TP2 | | | >35 | 192.7 | 78.0 | 200.1 | | 79.0 |
| | SR | 34.9 | 32.0 | | | | | | |
| | SR-3 | 35.9 | 11.5 | | | | | | |
| | SR+3 | 34.6 | 58.5 | | | | | | |
| | JP | 32.6 | 78.5 | 16.8 | 8.6 | 73.0 | 21.8 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

### Number of data points=20 and change in beginning at 5

Table D.6: Number of data points=20 and the true turning point in the beginning at 5.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 17.0 | 100.0 | 4.9 | 1.0 | 89.5 | 3.9 | 0.0 | 100.0 |
| ratic | C.TP1 | 16.0 | 7.0 | <1 | 514.7 | 99.0 | 233.3 | 0.0 | 56.0 |
| | C.TP2 | | | >20 | 333.1 | 80.0 | 220.5 | | 80.0 |
| | SR | 23.6 | 91.0 | | | | | | |
| | SR-3 | 37.7 | 8.0 | | | | | | |
| | SR+3 | 18.9 | 100.0 | | | | | | |
| | JP | 16.6 | 100.0 | 9.9 | 2.2 | 16.0 | 4.4 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.5 | 6.5 | 2.5 | 711.2 | 99.5 | 193.9 | 0.0 | 43.5 |
| ented | C.TP1 | 16.3 | 9.5 | 3.2 | 52.9 | 98.0 | 54.1 | 0.5 | 85.0 |
| with | C.TP2 | | | >20 | 88.1 | 48.0 | 64.6 | | 93.0 |
| one | SR | 16.1 | 11.0 | | | | | | |
| change | SR-3 | 16.3 | 6.0 | | | | | | |
| point | SR+3 | 16.2 | 10.0 | | | | | | |
| | JP | 15.5 | 35.0 | 8.6 | 6.1 | 97.5 | 16.1 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.9 | 100.0 | 4.3 | 1.4 | 30.0 | 5.1 | 0.0 | 82.0 |
| ined | C.TP1 | 16.4 | 23.0 | <1 | 188.0 | 14.0 | 164.5 | 0.0 | 85.0 |
| | C.TP2 | | | >20 | 164.5 | 70.0 | 214.8 | | 73.0 |
| | SR | 20.4 | 100.0 | | | | | | |
| | SR-3 | 35.5 | 18.5 | | | | | | |
| | SR+3 | 19.9 | 100.0 | | | | | | |
| | JP | 17.9 | 100.0 | 8.5 | 1.9 | 44.0 | 3.7 | 0.0 | 100.0 |
| $\beta_0 = 3$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 17.9 | 93.5 | 4.5 | 25.3 | 93.0 | 13.9 | 0.0 | 96.0 |
| ratic | C.TP1 | 16.8 | 8.0 | <1 | 204.9 | 100.0 | 180.7 | 0.0 | 60.0 |
| | C.TP2 | | | >20 | 172.5 | 75.0 | 186.6 | | 85.5 |
| | SR | 19.2 | 43.0 | | | | | | |
| | SR-3 | 23.7 | 4.5 | | | | | | |
| | SR+3 | 17.5 | 89.5 | | | | | | |
| | JP | 16.1 | 98.0 | 9.7 | 3.5 | 53.5 | 8.7 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.0 | 9.5 | <1 | 576.2 | 96.5 | 128.9 | 0.0 | 60.0 |
| ented | C.TP1 | 15.9 | 9.0 | <1 | 83.2 | 99.0 | 43.8 | 0.0 | 85.0 |
| with | C.TP2 | | | 16.1 | 59.9 | 26.0 | 54.0 | | 97.0 |
| one | SR | 16.1 | 8.5 | | | | | | |
| change | SR-3 | 16.0 | 4.0 | | | | | | |
| point | SR+3 | 16.2 | 10.0 | | | | | | |
| | JP | 15.5 | 25.0 | 10.1 | 6.3 | 96.5 | 15.9 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.8 | 68.5 | 5.8 | 191.0 | 90.5 | 73.9 | 0.0 | 68.5 |
| ined | C.TP1 | 17.3 | 12.0 | <1 | 197.9 | 94.0 | 85.5 | 0.0 | 76.0 |
| | C.TP2 | | | >20 | 106.8 | 87.0 | 126.2 | | 81.0 |
| | SR | 18.0 | 43.5 | | | | | | |
| | SR-3 | 20.0 | 5.5 | | | | | | |
| | SR+3 | 18.2 | 67.0 | | | | | | |
| | JP | 16.5 | 83.5 | 8.7 | 4.9 | 77.5 | 11.8 | 0.0 | 100.0 |
| $\beta_0 = 1.5$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 19.2 | 37.0 | 5.3 | 191.5 | 94.0 | 72.0 | 0.0 | 80.5 |

*Continued on next page*

Table D.6 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| ratic | C.TP1 | 18.1 | 6.5 | 2.1 | 92.6 | 98.0 | 61.2 | 0.0 | 83.5 |
| | C.TP2 | | | >20 | 48.5 | 52.0 | 68.9 | | 95.0 |
| | SR | 18.7 | 16.0 | | | | | | |
| | SR-3 | 19.4 | 2.0 | | | | | | |
| | SR+3 | 18.4 | 38.0 | | | | | | |
| | JP | 17.0 | 44.5 | 10.0 | 5.4 | 87.0 | 13.4 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.7 | 4.5 | 14.2 | 433.7 | 93.0 | 81.1 | 0.0 | 76.0 |
| ented | C.TP1 | 16.7 | 6.5 | 1.3 | 364.7 | 98.5 | 48.5 | 0.0 | 80.0 |
| with | C.TP2 | | | >20 | >1000 | 10.0 | 60.1 | | 93.5 |
| one | SR | 16.6 | 4.0 | | | | | | |
| change | SR-3 | 16.7 | 0.5 | | | | | | |
| point | SR+3 | 16.7 | 7.5 | | | | | | |
| | JP | 16.0 | 15.0 | 10.3 | 6.5 | 97.5 | 16.3 | 0.0 | 100.0 |
| Comb- | Q.TP | 19.4 | 20.5 | 8.4 | 467.0 | 97.0 | 117.9 | 0.0 | 62.0 |
| ined | C.TP1 | 18.4 | 5.5 | 1.1 | 69.5 | 99.0 | 53.9 | 0.0 | 83.0 |
| | C.TP2 | | | >20 | 368.8 | 54.0 | 65.5 | | 93.5 |
| | SR | 18.2 | 14.0 | | | | | | |
| | SR-3 | 18.8 | 0.5 | | | | | | |
| | SR+3 | 18.6 | 19.5 | | | | | | |
| | JP | 17.3 | 30.5 | 10.1 | 9.0 | 89.0 | 22.0 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

### D.4.3 Change occurs at the end of dataset

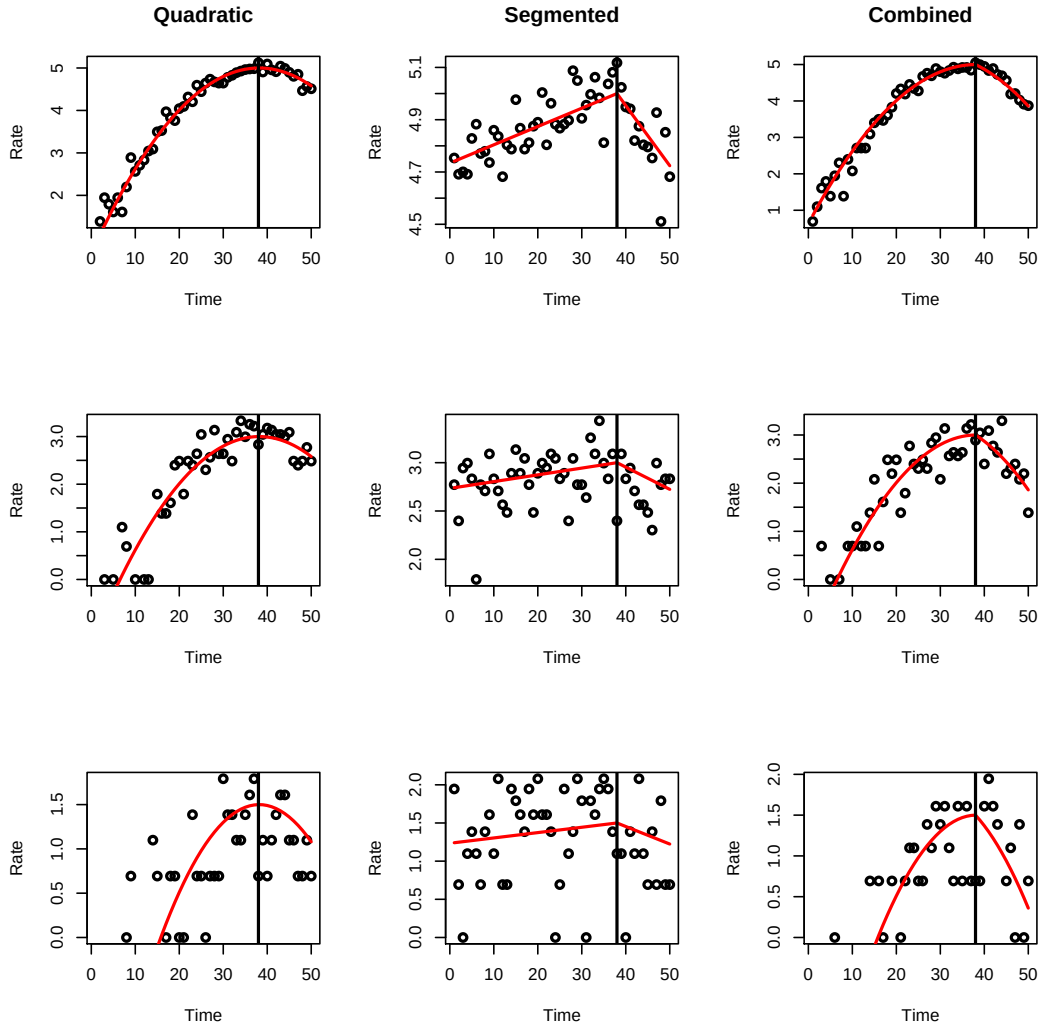



Figure D.1: Original model (red line) and simulated data (black circles) from quadratic, segmented and combined models with one true turning point in the end (black vertical line). Plots in the first row associated with $\beta_0 = 5$, the second row with $\beta_0 = 3$ and the third row with $\beta_0 = 1.5$.

### Number of data points=50 and change in end at 38

Table D.7: Number of data points=50 and the true turning point in the end at 38.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 47.0 | 100.0 | 38.2 | 0.4 | 90.0 | 1.5 | 0.0 | 100.0 |
| ratic | C.TP1 | 46.0 | 3.0 | <1 | >1000 | 89.0 | >1000 | 9.0 | 55.0 |
| | C.TP2 | | | >50 | >1000 | 69.0 | >1000 | | 45.0 |
| | SR | 178.6 | 100.0 | | | | | | |
| | SR-3 | 128.3 | 100.0 | | | | | | |
| | SR+3 | 258.0 | 100.0 | | | | | | |
| | JP | 92.6 | 100.0 | 29.9 | 1.4 | 0.0 | 2.3 | 0.0 | 100.0 |
| Segm- | Q.TP | 50.7 | 90.5 | 20.0 | 154.0 | 8.0 | 28.2 | 0.0 | 100.0 |
| ented | C.TP1 | 45.9 | 50.0 | <1 | 106.6 | 4.0 | 127.2 | 4.5 | 70.0 |
| with | C.TP2 | | | 48.9 | 184.2 | 52.0 | 90.3 | | 97.0 |
| one | SR | 45.3 | 79.0 | | | | | | |
| change | SR-3 | 45.3 | 96.0 | | | | | | |
| point | SR+3 | 46.1 | 37.0 | | | | | | |
| | JP | 45.0 | 98.0 | 36.5 | 5.5 | 95.0 | 23.0 | 0.0 | 100.0 |
| Comb- | Q.TP | 55.4 | 100.0 | 35.8 | 0.3 | 0.0 | 1.2 | 0.0 | 100.0 |
| ined | C.TP1 | 49.5 | 11.5 | <1 | >1000 | 0.0 | >1000 | 1.0 | 20.0 |
| | C.TP2 | | | >50 | >1000 | 76.0 | >1000 | | 80.0 |
| | SR | 177.4 | 100.0 | | | | | | |
| | SR-3 | 131.3 | 100.0 | | | | | | |
| | SR+3 | 278.1 | 100.0 | | | | | | |
| | JP | 112.8 | 100.0 | 32.0 | 1.4 | 0.0 | 2.1 | 0.0 | 100.0 |
| $\beta_0 = 3$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 48.4 | 100.0 | 38.5 | 1.6 | 94.0 | 6.2 | 0.0 | 100.0 |
| ratic | C.TP1 | 47.4 | 5.5 | <1 | 911.2 | 86.0 | 785.4 | 3.5 | 49.5 |
| | C.TP2 | | | >50 | >1000 | 90.0 | 773.0 | | 54.5 |
| | SR | 60.0 | 99.5 | | | | | | |
| | SR-3 | 55.5 | 100.0 | | | | | | |
| | SR+3 | 66.6 | 81.0 | | | | | | |
| | JP | 49.8 | 100.0 | 28.8 | 4.4 | 12.5 | 8.6 | 0.0 | 100.0 |
| Segm- | Q.TP | 51.0 | 35.5 | 28.3 | 488.8 | 79.5 | 98.4 | 0.0 | 95.5 |
| ented | C.TP1 | 49.0 | 14.5 | <1 | 212.8 | 7.0 | 146.0 | 3.5 | 74.5 |
| with | C.TP2 | | | 45.2 | 167.1 | 90.5 | 99.1 | | 94.0 |
| one | SR | 48.9 | 26.0 | | | | | | |
| change | SR-3 | 48.9 | 48.0 | | | | | | |
| point | SR+3 | 49.1 | 14.0 | | | | | | |
| | JP | 48.2 | 72.5 | 34.7 | 11.6 | 99.0 | 41.6 | 0.0 | 100.0 |
| Comb- | Q.TP | 48.2 | 100.0 | 36.8 | 0.9 | 74.0 | 3.6 | 0.0 | 100.0 |
| ined | C.TP1 | 47.0 | 7.0 | 3.6 | 91.1 | 8.0 | 114.7 | 1.5 | 83.0 |
| | C.TP2 | | | 41.3 | 266.9 | 100.0 | 124.1 | | 89.0 |
| | SR | 64.6 | 100.0 | | | | | | |
| | SR-3 | 58.5 | 100.0 | | | | | | |
| | SR+3 | 76.9 | 99.0 | | | | | | |
| | JP | 53.5 | 100.0 | 31.0 | 3.8 | 13.5 | 5.5 | 0.0 | 100.0 |
| $\beta_0 = 1.5$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 48.3 | 98.0 | 38.6 | 42.7 | 96.0 | 24.8 | 0.0 | 98.0 |

*Continued on next page*

Table D.7 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| ratic | C.TP1 | 47.3 | 3.0 | <1 | 459.3 | 81.5 | 804.8 | 1.0 | 87.5 |
| | C.TP2 | | | >50 | 278.1 | 99.5 | 382.7 | | 61.5 |
| | SR | 51.4 | 59.0 | | | | | | |
| | SR-3 | 50.0 | 87.5 | | | | | | |
| | SR+3 | 54.1 | 34.5 | | | | | | |
| | JP | 46.7 | 95.0 | 28.0 | 8.3 | 53.5 | 19.0 | 0.0 | 100.0 |
| Segm- | Q.TP | 52.4 | 14.5 | 26.3 | 811.2 | 86.0 | 130.8 | 0.0 | 88.0 |
| ented | C.TP1 | 50.9 | 10.5 | <1 | 151.3 | 3.0 | 139.5 | 1.0 | 75.5 |
| with | C.TP2 | | | >50 | 304.5 | 96.5 | 117.7 | | 87.5 |
| one | SR | 50.8 | 12.0 | | | | | | |
| change | SR-3 | 50.8 | 17.5 | | | | | | |
| point | SR+3 | 51.1 | 7.0 | | | | | | |
| | JP | 49.6 | 40.5 | 27.7 | 17.2 | 96.5 | 43.2 | 0.0 | 100.0 |
| Comb- | Q.TP | 47.2 | 100.0 | 37.6 | 6.7 | 91.0 | 11.0 | 0.0 | 100.0 |
| ined | C.TP1 | 46.3 | 3.0 | <1 | >1000 | 66.0 | 487.3 | 0.5 | 46.5 |
| | C.TP2 | | | >50 | 914.4 | 94.5 | 425.2 | | 63.0 |
| | SR | 50.7 | 82.5 | | | | | | |
| | SR-3 | 49.3 | 95.5 | | | | | | |
| | SR+3 | 53.0 | 48.0 | | | | | | |
| | JP | 46.5 | 96.0 | 29.5 | 6.8 | 47.5 | 16.4 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

## Number of data points=35 and change in end at 26

Table D.8: Number of data points=35 and the true turning point in the end at 26.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 32.5 | 100.0 | 26.2 | 0.5 | 91.0 | 2.1 | 0.0 | 100.0 |
| ratic | C.TP1 | 31.5 | 3.5 | <1 | >1000 | 89.5 | >1000 | 3.0 | 54.0 |
| | C.TP2 | | | >35 | >1000 | 72.0 | >1000 | | 46.0 |
| | SR | 74.5 | 100.0 | | | | | | |
| | SR-3 | 54.7 | 100.0 | | | | | | |
| | SR+3 | 110.1 | 99.5 | | | | | | |
| | JP | 44.5 | 100.0 | 19.6 | 2.0 | 2.0 | 3.4 | 0.0 | 100.0 |
| Segm- | Q.TP | 33.2 | 60.5 | 20.7 | 99.0 | 66.0 | 41.7 | 0.0 | 99.5 |
| ented | C.TP1 | 31.0 | 19.5 | <1 | 170.6 | 3.5 | 115.3 | 2.0 | 78.5 |
| with | C.TP2 | | | >35 | 94.2 | 89.5 | 102.6 | | 91.5 |
| one | SR | 30.5 | 43.0 | | | | | | |
| change | SR-3 | 30.6 | 68.0 | | | | | | |
| point | SR+3 | 31.2 | 9.0 | | | | | | |
| | JP | 30.1 | 82.5 | 24.3 | 7.1 | 97.0 | 24.4 | 0.0 | 100.0 |
| Comb- | Q.TP | 32.6 | 100.0 | 24.8 | 0.4 | 2.0 | 1.6 | 0.0 | 100.0 |
| ined | C.TP1 | 30.7 | 14.0 | <1 | >1000 | 6.5 | 843.9 | 2.5 | 18.0 |
| | C.TP2 | | | >35 | >1000 | 88.5 | 877.7 | | 82.0 |
| | SR | 74.6 | 100.0 | | | | | | |
| | SR-3 | 54.2 | 100.0 | | | | | | |
| | SR+3 | 118.1 | 99.5 | | | | | | |

*Continued on next page*

Table D.8 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| | JP | 47.8 | 100.0 | 20.8 | 1.7 | 2.0 | 2.8 | 0.0 | 100.0 |
| | | | | | $\beta_0 = 3$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 33.7 | 100.0 | 26.3 | 8.0 | 95.0 | 6.8 | 0.0 | 100.0 |
| ratic | C.TP1 | 32.5 | 7.0 | <1 | 363.4 | 97.5 | 334.1 | 0.5 | 51.0 |
| | C.TP2 | | | >35 | 337.5 | 97.0 | 334.1 | | 78.5 |
| | SR | 38.9 | 81.5 | | | | | | |
| | SR-3 | 36.2 | 99.5 | | | | | | |
| | SR+3 | 44.1 | 35.5 | | | | | | |
| | JP | 32.7 | 99.5 | 18.8 | 4.4 | 32.5 | 9.5 | 0.0 | 100.0 |
| Segm- | Q.TP | 33.2 | 35.5 | 19.2 | 118.7 | 71.0 | 55.4 | 0.0 | 95.5 |
| ented | C.TP1 | 31.2 | 19.0 | 3.1 | 123.0 | 2.5 | 93.1 | 2.0 | 77.0 |
| with | C.TP2 | | | >35 | 248.9 | 96.0 | 98.3 | | 92.0 |
| one | SR | 31.3 | 23.5 | | | | | | |
| change | SR-3 | 31.6 | 37.5 | | | | | | |
| point | SR+3 | 31.7 | 8.0 | | | | | | |
| | JP | 30.5 | 57.5 | 22.3 | 9.2 | 95.5 | 27.5 | 0.0 | 100.0 |
| Comb- | Q.TP | 32.8 | 100.0 | 24.7 | 1.4 | 75.5 | 4.8 | 0.0 | 100.0 |
| ined | C.TP1 | 31.6 | 7.0 | <1 | 373.6 | 43.5 | 348.4 | 3.5 | 40.0 |
| | C.TP2 | | | >35 | 849.6 | 95.5 | 401.9 | | 72.0 |
| | SR | 37.8 | 95.5 | | | | | | |
| | SR-3 | 35.3 | 99.5 | | | | | | |
| | SR+3 | 43.9 | 44.5 | | | | | | |
| | JP | 32.5 | 100.0 | 20.4 | 4.4 | 39.5 | 7.8 | 0.0 | 100.0 |
| | | | | | $\beta_0 = 1.5$ | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 35.0 | 71.0 | 29.6 | 955.2 | 96.0 | 69.7 | 0.0 | 91.5 |
| ratic | C.TP1 | 33.9 | 5.5 | <1 | 124.3 | 75.5 | 154.1 | 0.5 | 69.0 |
| | C.TP2 | | | >35 | 458.0 | 96.0 | 157.2 | | 91.5 |
| | SR | 35.4 | 32.5 | | | | | | |
| | SR-3 | 34.7 | 58.5 | | | | | | |
| | SR+3 | 36.4 | 11.5 | | | | | | |
| | JP | 32.9 | 80.5 | 18.9 | 8.2 | 78.0 | 22.2 | 0.0 | 100.0 |
| Segm- | Q.TP | 34.7 | 10.0 | 23.7 | 898.9 | 91.5 | 96.4 | 0.0 | 87.0 |
| ented | C.TP1 | 33.5 | 8.0 | 5.1 | 90.9 | 2.0 | 85.0 | 0.0 | 81.5 |
| with | C.TP2 | | | >35 | 87.7 | 98.0 | 82.1 | | 92.5 |
| one | SR | 33.5 | 10.0 | | | | | | |
| change | SR-3 | 33.5 | 12.5 | | | | | | |
| point | SR+3 | 33.4 | 8.0 | | | | | | |
| | JP | 32.2 | 30.0 | 19.4 | 11.7 | 96.5 | 30.0 | 0.0 | 100.0 |
| Comb- | Q.TP | 34.7 | 81.5 | 26.7 | 96.6 | 93.5 | 52.8 | 0.0 | 94.0 |
| ined | C.TP1 | 33.8 | 4.0 | <1 | 554.2 | 63.5 | 199.1 | 0.5 | 54.0 |
| | C.TP2 | | | 34.1 | 145.4 | 97.0 | 184.4 | | 88.0 |
| | SR | 35.2 | 35.0 | | | | | | |
| | SR-3 | 34.8 | 63.5 | | | | | | |
| | SR+3 | 36.4 | 12.0 | | | | | | |
| | JP | 33.0 | 78.0 | 19.0 | 7.5 | 81.0 | 22.3 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

### Number of data points=20 and change in end at 15

Table D.9: Number of data points=20 and the true turning point in the end at 15.

| $\beta_0 = 5$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 16.5 | 100.0 | 15.2 | 0.8 | 92.0 | 3.2 | 0.0 | 100.0 |
| ratic | C.TP1 | 15.4 | 6.5 | <1 | 585.8 | 80.0 | 230.7 | 0.0 | 55.5 |
| | C.TP2 | | | >20 | 355.3 | 89.0 | 237.7 | | 78.5 |
| | SR | 24.7 | 91.5 | | | | | | |
| | SR-3 | 18.8 | 100.0 | | | | | | |
| | SR+3 | 38.6 | 10.5 | | | | | | |
| | JP | 16.7 | 100.0 | 10.9 | 2.3 | 27.0 | 4.2 | 0.0 | 100.0 |
| Segm- | Q.TP | 16.7 | 13.5 | >20 | 297.8 | 87.5 | 54.3 | 0.5 | 88.0 |
| ented | C.TP1 | 15.4 | 8.5 | <1 | 70.7 | 10.0 | 52.1 | 1.0 | 79.5 |
| with | C.TP2 | | | 15.6 | 36.9 | 97.5 | 49.2 | | 100.0 |
| one | SR | 15.5 | 12.5 | | | | | | |
| change | SR-3 | 15.4 | 17.0 | | | | | | |
| point | SR+3 | 15.9 | 3.5 | | | | | | |
| | JP | 14.8 | 40.0 | 11.2 | 6.0 | 95.5 | 15.5 | 0.0 | 100.0 |
| Comb- | Q.TP | 19.0 | 100.0 | 12.8 | 0.4 | 3.5 | 1.6 | 0.0 | 100.0 |
| ined | C.TP1 | 16.3 | 28.0 | <1 | 275.7 | 0.0 | 186.3 | 0.0 | 9.0 |
| | C.TP2 | | | >20 | 239.0 | 61.5 | 152.4 | | 94.5 |
| | SR | 22.8 | 100.0 | | | | | | |
| | SR-3 | 20.5 | 100.0 | | | | | | |
| | SR+3 | 49.1 | 15.0 | | | | | | |
| | JP | 19.2 | 100.0 | 12.9 | 1.6 | 49.0 | 3.1 | 0.0 | 100.0 |
| $\beta_0 = 3$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 16.4 | 96.0 | 15.6 | 62.0 | 96.0 | 11.4 | 0.0 | 97.5 |
| ratic | C.TP1 | 15.5 | 3.0 | <1 | 210.5 | 65.0 | 177.1 | 0.0 | 48.5 |
| | C.TP2 | | | >20 | 171.5 | 95.0 | 122.8 | | 95.5 |
| | SR | 18.3 | 43.5 | | | | | | |
| | SR-3 | 16.3 | 92.5 | | | | | | |
| | SR+3 | 23.5 | 5.5 | | | | | | |
| | JP | 15.1 | 99.0 | 10.9 | 3.3 | 58.0 | 7.9 | 0.0 | 100.0 |
| Segm- | Q.TP | 17.2 | 7.0 | 7.9 | >1000 | 93.5 | 53.4 | 0.0 | 87.5 |
| ented | C.TP1 | 16.2 | 5.5 | <1 | 122.5 | 5.5 | 53.4 | 0.0 | 78.5 |
| with | C.TP2 | | | 15.1 | 68.3 | 95.0 | 50.1 | | 99.5 |
| one | SR | 15.9 | 5.5 | | | | | | |
| change | SR-3 | 16.1 | 7.5 | | | | | | |
| point | SR+3 | 16.2 | 5.5 | | | | | | |
| | JP | 15.4 | 32.0 | 11.3 | 6.2 | 96.0 | 16.0 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.4 | 83.0 | 13.2 | 46.2 | 69.5 | 15.6 | 0.0 | 99.5 |
| ined | C.TP1 | 17.0 | 10.0 | <1 | 235.3 | 21.0 | 138.9 | 0.0 | 47.0 |
| | C.TP2 | | | >20 | 181.1 | 96.0 | 116.6 | | 92.0 |
| | SR | 18.2 | 40.0 | | | | | | |
| | SR-3 | 17.7 | 79.5 | | | | | | |
| | SR+3 | 21.8 | 4.5 | | | | | | |
| | JP | 16.2 | 92.5 | 11.4 | 4.2 | 81.0 | 10.3 | 0.0 | 100.0 |
| $\beta_0 = 1.5$ | | | | | | | | | |
| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
| Quad- | Q.TP | 17.9 | 44.0 | 14.2 | 273.4 | 96.5 | 56.0 | 0.0 | 86.5 |

*Continued on next page*

Table D.9 – *Continued from previous page*

| O.M | S.M | DV | SG.CH% | TP.ES | TP.SD | CI.TP% | CI.WD | NO.TP% | TP.IN% |
|---|---|---|---|---|---|---|---|---|---|
| ratic | C.TP1 | 17.1 | 3.5 | 2.1 | 271.2 | 52.0 | 77.1 | 0.0 | 74.0 |
| | C.TP2 | | | >20 | 107.2 | 98.5 | 86.9 | | 96.5 |
| | SR | 17.7 | 15.0 | | | | | | |
| | SR-3 | 17.3 | 37.5 | | | | | | |
| | SR+3 | 19.0 | 0.0 | | | | | | |
| | JP | 16.1 | 47.5 | 10.6 | 5.1 | 83.0 | 13.4 | 0.0 | 100.0 |
| Segm- | Q.TP | 18.4 | 3.5 | 17.1 | 213.7 | 94.5 | 60.8 | 0.0 | 86.5 |
| ented | C.TP1 | 17.4 | 5.0 | 4.7 | 58.9 | 6.0 | 49.3 | 0.0 | 93.0 |
| with | C.TP2 | | | >20 | 64.7 | 98.0 | 48.9 | | 97.0 |
| one | SR | 17.4 | 3.5 | | | | | | |
| change | SR-3 | 17.3 | 2.5 | | | | | | |
| point | SR+3 | 17.3 | 1.0 | | | | | | |
| | JP | 16.6 | 21.5 | 10.7 | 6.4 | 98.0 | 16.5 | 0.0 | 100.0 |
| Comb- | Q.TP | 18.1 | 25.5 | 15.6 | 186.2 | 84.5 | 48.1 | 0.0 | 86.5 |
| ined | C.TP1 | 17.1 | 4.5 | <1 | 191.3 | 17.5 | 71.0 | 0.0 | 64.5 |
| | C.TP2 | | | >20 | >1000 | 95.0 | 61.1 | | 98.5 |
| | SR | 17.4 | 10.0 | | | | | | |
| | SR-3 | 17.4 | 22.0 | | | | | | |
| | SR+3 | 18.1 | 1.0 | | | | | | |
| | JP | 16.2 | 38.0 | 11.8 | 5.6 | 88.5 | 14.5 | 0.0 | 100.0 |

See Table 6.2 for the definition of **O.M, SR, CI.TP%**. See Table 6.1 for the definition of $\beta_1$, $\beta_0$, **S.M, Q.TP, C.TP, JP, DV, SG.CH%, TP.ES, TP.SD, CI.WD, NO.TP%, TP.IN%**.

# D.5 Simulation study with two change points

## D.5.1 Two change points occur about in the middle of data

### Number of data points=35 and change at 11 and 24

Table D.10: Number of data points=35 and the true turning points at 11 and 24.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 35.9 | 8.5 | 14.7 | 508.5 | -53.1 | 86.2 | 139.3 | 99.0 | 94.0 | 0.0 | 79.5 |
| | C.TP1 | 31.0 | 58.0 | 11.4 | 111.1 | -9.2 | 17.1 | 26.3 | 94.5 | | 0.5 | 100.0 |
| | C.TP2 | | | 24.9 | 49.3 | 18.4 | 45.6 | 27.2 | | 94.0 | | 97.5 |
| | SR1 | 29.4 | 28.0 | | | | | | | | | |
| | SR2 | | 35.5 | | | | | | | | | |
| | SR1 -3 | 29.7 | 18.5 | | | | | | | | | |
| | SR2 -3 | | 32.0 | | | | | | | | | |
| | SR1 +3 | 29.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 22.5 | | | | | | | | | |
| | JP1 | 26.9 | 79.0 | 12.8 | 8.3 | 2.9 | 26.6 | 23.7 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 83.5 | 22.3 | 8.5 | 10.2 | 32.9 | 22.7 | | 94.0 | 0.0 | 100.0 |

*Continued on next page*

Table D.10 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Q.TP | 32.1 | 6.5 | >35 | >1000 | -41.9 | 75.7 | 117.6 | 96.5 | 95.5 | 0.5 | 84.5 |
| | C.TP1 | 30.5 | 13.0 | 11.6 | 359.9 | -44.9 | 19.5 | 64.4 | 99.5 | | 1.5 | 98.5 |
| | C.TP2 | | | >35 | >1000 | 16.4 | 83.7 | 67.3 | | 96.0 | | 86.0 |
| | SR1 | 28.6 | 6.5 | | | | | | | | | |
| | SR2 | | 7.0 | | | | | | | | | |
| | SR1 -3 | 28.7 | 6.0 | | | | | | | | | |
| | SR2 -3 | | 6.5 | | | | | | | | | |
| | SR1 +3 | 28.9 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 5.0 | | | | | | | | | |
| | JP1 | 26.5 | 60.5 | 14.1 | 10.2 | 2.3 | 29.3 | 27.0 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 63.5 | 19.8 | 10.3 | 6.2 | 33.6 | 27.4 | | 98.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.3 | 5.0 | 20.9 | 428.4 | -37.6 | 77.1 | 114.6 | 96.0 | 95.0 | 0.0 | 82.5 |
| | C.TP1 | 32.0 | 5.5 | 11.4 | 166.6 | -53.4 | 19.9 | 73.3 | 98.0 | | 0.5 | 98.0 |
| | C.TP2 | | | >35 | 239.0 | 15.5 | 87.8 | 72.3 | | 100.0 | | 86.0 |
| | SR1 | 30.3 | 8.0 | | | | | | | | | |
| | SR2 | | 6.0 | | | | | | | | | |
| | SR1 -3 | 30.2 | 5.5 | | | | | | | | | |
| | SR2 -3 | | 6.0 | | | | | | | | | |
| | SR1 +3 | 30.1 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 6.5 | | | | | | | | | |
| | JP1 | 27.6 | 55.0 | 15.5 | 10.3 | 2.4 | 29.1 | 26.7 | 97.0 | | 0.0 | 100.0 |
| | JP2 | | 54.0 | 20.7 | 10.6 | 6.6 | 33.6 | 27.0 | | 98.0 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 33.2 | 8.0 | 2.4 | 291.9 | -47.1 | 81.1 | 128.2 | 96.0 | 92.5 | 2.5 | 79.0 |
| | C.TP1 | 31.2 | 14.0 | 11.0 | 221 | -36.6 | 19.3 | 55.9 | 97.5 | | 0.5 | 99.0 |
| | C.TP2 | | | 28.8 | 295.7 | 15.4 | 74.7 | 59.3 | | 98.0 | | 93.0 |
| | SR1 | 28.8 | 11.5 | | | | | | | | | |
| | SR2 | | 17.0 | | | | | | | | | |
| | SR1 -3 | 29.0 | 9.0 | | | | | | | | | |
| | SR2 -3 | | 20.0 | | | | | | | | | |
| | SR1 +3 | 29.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 7.5 | | | | | | | | | |
| | JP1 | 26.4 | 76.0 | 15.4 | 9.9 | 2.9 | 28.2 | 25.3 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 75.5 | 20.9 | 9.6 | 7.7 | 33.2 | 25.5 | | 97.0 | 0.0 | 100.0 |
| 3 | Q.TP | 33.0 | 8.5 | 14.4 | 272.5 | -37.8 | 73.5 | 111.3 | 95.0 | 91.0 | 0.0 | 88.5 |
| | C.TP1 | 31.7 | 7.0 | 11.0 | >1000 | -51.2 | 20.3 | 71.5 | 99.0 | | 0.0 | 99.0 |
| | C.TP2 | | | 29.0 | 200.0 | 15.7 | 87.2 | 71.5 | | 98.0 | | 89.0 |
| | SR1 | 29.4 | 8.0 | | | | | | | | | |
| | SR2 | | 9.5 | | | | | | | | | |
| | SR1 -3 | 29.9 | 5.0 | | | | | | | | | |
| | SR2 -3 | | 12.0 | | | | | | | | | |
| | SR1 +3 | 29.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 10.5 | | | | | | | | | |
| | JP1 | 27.3 | 60.0 | 18.0 | 10.0 | 2.5 | 29.6 | 27.1 | 99.0 | | 0.0 | 100.0 |
| | JP2 | | 61.5 | 23.0 | 9.8 | 7.3 | 33.5 | 26.2 | | 97.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.4 | 6.5 | <1 | >1000 | -44.4 | 78.1 | 122.4 | 94.5 | 96.5 | 0.5 | 81.5 |
| | C.TP1 | 32.3 | 5.0 | 11.0 | 420.0 | -52.4 | 20.2 | 72.6 | 99.5 | | 0.0 | 96.5 |
| | C.TP2 | | | 29.6 | >1000 | 15.7 | 89.2 | 73.5 | | 97.0 | | 84.0 |
| | SR1 | 30.3 | 6.0 | | | | | | | | | |
| | SR2 | | 6.5 | | | | | | | | | |
| | SR1 -3 | 30.5 | 4.0 | | | | | | | | | |

*Continued on next page*

Table D.10 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR2 -3 | | 5.0 | | | | | | | | | |
| | SR1 +3 | 30.2 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 8.0 | | | | | | | | | |
| | JP1 | 27.8 | 55.0 | 14.7 | 10.6 | 2.5 | 29.6 | 27.1 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 53.0 | 20.1 | 11.1 | 6.5 | 33.7 | 27.1 | | 96.5 | 0.0 | 100.0 |

Original model is combined model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 41.5 | 18.0 | 18.8 | >1000 | -28.8 | 64.1 | 92.9 | 94.0 | 89.0 | 0.0 | 87.0 |
| | C.TP1 | 30.3 | 85.5 | 8.9 | 86.0 | -1.2 | 12.3 | 13.5 | 67.0 | | 0.0 | 100.0 |
| | C.TP2 | | | 24.8 | 17.4 | 21.2 | 30.2 | 9.0 | | 89.5 | | 99.5 |
| | SR1 | 29.7 | 52.0 | | | | | | | | | |
| | SR2 | | 75.5 | | | | | | | | | |
| | SR1 -3 | 30.6 | 38.5 | | | | | | | | | |
| | SR2 -3 | | 74.5 | | | | | | | | | |
| | SR1 +3 | 30.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 46.0 | | | | | | | | | |
| | JP1 | 27.7 | 87.5 | 13.5 | 8.4 | 4.0 | 24.9 | 20.9 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 91.5 | 24.7 | 5.8 | 14.6 | 31.7 | 17.1 | | 94.5 | 0.0 | 100.0 |
| 3 | Q.TP | 33.5 | 6.5 | 16.0 | 302.8 | -39.0 | 75.1 | 114.1 | 96.0 | 95.0 | 0.0 | 80.5 |
| | C.TP1 | 31.4 | 21.0 | 10.4 | 157.7 | -31.0 | 17.6 | 48.6 | 93.5 | | 0.5 | 99.0 |
| | C.TP2 | | | >35 | 127.8 | 18.0 | 66.1 | 48.1 | | 94.5 | | 90.5 |
| | SR1 | 29.2 | 12.0 | | | | | | | | | |
| | SR2 | | 15.5 | | | | | | | | | |
| | SR1 -3 | 29.4 | 8.5 | | | | | | | | | |
| | SR2 -3 | | 16.0 | | | | | | | | | |
| | SR1 +3 | 29.7 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 7.5 | | | | | | | | | |
| | JP1 | 27.0 | 62.5 | 14.4 | 9.7 | 2.6 | 28.8 | 26.3 | 99.5 | | 0.0 | 100.0 |
| | JP2 | | 67.0 | 20.9 | 9.6 | 7.6 | 33.4 | 25.8 | | 97.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 32.9 | 4.0 | 4.4 | 383.9 | -46.2 | 80.2 | 126.4 | 95.5 | 96.0 | 0.0 | 80.0 |
| | C.TP1 | 32.9 | 13.5 | 10.9 | >1000 | -45.9 | 19.4 | 65.3 | 98.5 | | 0.5 | 99.5 |
| | C.TP2 | | | >35 | 152.6 | 16.5 | 80.3 | 63.8 | | 96.5 | | 87.5 |
| | SR1 | 29.6 | 6.0 | | | | | | | | | |
| | SR2 | | 7.5 | | | | | | | | | |
| | SR1 -3 | 29.6 | 4.0 | | | | | | | | | |
| | SR2 -3 | | 4.0 | | | | | | | | | |
| | SR1 +3 | 29.8 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 8.0 | | | | | | | | | |
| | JP1 | 27.2 | 56.0 | 16.4 | 10.1 | 2.5 | 29.3 | 26.8 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 57.0 | 21.2 | 9.9 | 6.9 | 33.5 | 26.6 | | 98.0 | 0.0 | 100.0 |

See Table 6.4 for the definition of $\beta_0$**, S.M, Q.TP, C.TP, SR, JP, DV, SG%, ETP, SD, LCL, UCL, WD, CI1%, CI2%, NO%, IN%**.

### Number of data points=20 and changes at 7 and 13

Table D.11: Number of data points=20 and the true turning points at 7 and 13.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 17.4 | 6.0 | <1 | >1000 | -22.3 | 43.6 | 65.9 | 95.5 | 97.5 | 1.0 | 84.5 |
| | C.TP1 | 16.5 | 6.5 | 6.4 | 154.0 | -24.7 | 11.4 | 36.1 | 95.5 | | 0.5 | 96.5 |
| | C.TP2 | | | >20 | 270.0 | 9.3 | 45.6 | 36.3 | | 94.0 | | 87.5 |
| | SR1 | 14.4 | 5.5 | | | | | | | | | |
| | SR2 | | 5.0 | | | | | | | | | |
| | SR1 -3 | 14.4 | 4.5 | | | | | | | | | |
| | SR2 -3 | | 5.0 | | | | | | | | | |
| | SR1 +3 | 14.4 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 4.0 | | | | | | | | | |
| | JP1 | 12.7 | 49.5 | 8.7 | 5.1 | 2.2 | 15.6 | 13.3 | 96.0 | | 0.0 | 100.0 |
| | JP2 | | 53.0 | 12.0 | 5.2 | 5.4 | 18.5 | 13.1 | | 95.0 | 0.0 | 100.0 |
| 3 | Q.TP | 17.1 | 3.5 | >20 | 180.3 | -22.4 | 45.0 | 67.4 | 93.0 | 95.0 | 0.0 | 81.0 |
| | C.TP1 | 15.9 | 4.5 | 6.3 | 173.5 | -28.1 | 11.5 | 39.6 | 99.0 | | 0.0 | 96.0 |
| | C.TP2 | | | 17.7 | 104.5 | 9.4 | 48.5 | 39.1 | | 95.5 | | 85.5 |
| | SR1 | 14.1 | 8.5 | | | | | | | | | |
| | SR2 | | 5.0 | | | | | | | | | |
| | SR1 -3 | 13.9 | 3.5 | | | | | | | | | |
| | SR2 -3 | | 5.0 | | | | | | | | | |
| | SR1 +3 | 14.1 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 6.0 | | | | | | | | | |
| | JP1 | 12.0 | 48.5 | 9.2 | 5.5 | 2.3 | 15.3 | 13.1 | 97.0 | | 0.0 | 100.0 |
| | JP2 | | 51.0 | 12.7 | 5.4 | 5.7 | 18.6 | 12.9 | | 94.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 17.9 | 5.0 | 9.7 | 178.2 | -22.0 | 42.4 | 64.4 | 94.5 | 96.5 | 0.0 | 83.5 |
| | C.TP1 | 16.9 | 8.0 | 6.4 | 70.9 | -24.2 | 11.2 | 35.4 | 97.5 | | 0.0 | 96.0 |
| | C.TP2 | | | >20 | 87.9 | 10.0 | 47.5 | 37.5 | | 90.5 | | 85.0 |
| | SR1 | 14.8 | 2.5 | | | | | | | | | |
| | SR2 | | 4.5 | | | | | | | | | |
| | SR1 -3 | 14.9 | 5.5 | | | | | | | | | |
| | SR2 -3 | | 5.0 | | | | | | | | | |
| | SR1 +3 | 14.9 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 4.5 | | | | | | | | | |
| | JP1 | 13.3 | 39.5 | 9.0 | 4.8 | 2.2 | 15.8 | 13.5 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 41.0 | 12.4 | 4.9 | 5.2 | 18.8 | 13.6 | | 97.0 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 17.3 | 6.0 | 9.4 | 998.9 | -25.5 | 48.9 | 74.5 | 96.0 | 95.5 | 0.0 | 79.5 |
| | C.TP1 | 16.2 | 6.5 | 6.6 | 120.7 | -25.7 | 11.5 | 37.2 | 99.5 | | 0.5 | 99.0 |
| | C.TP2 | | | 19.9 | 97.4 | 9.3 | 50.3 | 41.0 | | 92.0 | | 82.5 |
| | SR1 | 14.0 | 5.0 | | | | | | | | | |
| | SR2 | | 7.5 | | | | | | | | | |
| | SR1 -3 | 14.0 | 7.5 | | | | | | | | | |
| | SR2 -3 | | 9.5 | | | | | | | | | |
| | SR1 +3 | 14.3 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 3.5 | | | | | | | | | |
| | JP1 | 12.4 | 49.0 | 8.5 | 5.2 | 2.4 | 15.3 | 13.0 | 95.5 | | 0.0 | 100.0 |
| | JP2 | | 49.5 | 11.9 | 5.2 | 5.5 | 18.6 | 13.0 | | 94.0 | 0.0 | 100.0 |

*Continued on next page*

Table D.11 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | Q.TP | 17.7 | 2.0 | 9.6 | 418.4 | -20.8 | 42.4 | 63.2 | 95.0 | 98.5 | 0.0 | 86.0 |
| | C.TP1 | 16.8 | 6.0 | 6.6 | 125.5 | -28.1 | 11.6 | 39.7 | 98.5 | | 0.0 | 96.0 |
| | C.TP2 | | | 17.9 | 134.7 | 9.5 | 49.4 | 39.9 | | 97.0 | | 85.5 |
| | SR1 | 14.8 | 5.0 | | | | | | | | | |
| | SR2 | | 5.5 | | | | | | | | | |
| | SR1 -3 | 14.7 | 5.5 | | | | | | | | | |
| | SR2 -3 | | 3.5 | | | | | | | | | |
| | SR1 +3 | 14.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 6.0 | | | | | | | | | |
| | JP1 | 12.9 | 50.5 | 9.1 | 5.1 | 2.3 | 15.4 | 13.2 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 52.0 | 11.9 | 5.4 | 5.2 | 18.6 | 13.3 | | 96.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 18.3 | 3.0 | 7.5 | 137.9 | -21.1 | 42.7 | 63.8 | 95.5 | 98.5 | 0.0 | 82.5 |
| | C.TP1 | 17.3 | 8.5 | 6.7 | >1000 | -26.0 | 11.6 | 37.6 | 96.5 | | 0.0 | 97.0 |
| | C.TP2 | | | 17.7 | 119.0 | 9.6 | 48.5 | 38.9 | | 90.5 | | 86.5 |
| | SR1 | 15.1 | 6.5 | | | | | | | | | |
| | SR2 | | 6.0 | | | | | | | | | |
| | SR1 -3 | 15.1 | 6.0 | | | | | | | | | |
| | SR2 -3 | | 5.0 | | | | | | | | | |
| | SR1 +3 | 15.2 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.0 | | | | | | | | | |
| | JP1 | 13.2 | 45.0 | 8.8 | 5.1 | 2.4 | 15.3 | 12.9 | 95.0 | | 0.0 | 100.0 |
| | JP2 | | 44.0 | 11.8 | 5.1 | 5.5 | 18.5 | 12.9 | | 93.5 | 0.0 | 100.0 |

Original model is combined model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 16.8 | 6.5 | 13.9 | 321.1 | -20.1 | 40.7 | 60.8 | 94.5 | 93.0 | 0.0 | 85.5 |
| | C.TP1 | 15.7 | 13.0 | 6.1 | 147.0 | -29.0 | 10.9 | 39.9 | 91.0 | | 0.0 | 96.5 |
| | C.TP2 | | | 17.0 | 799.0 | 9.6 | 46.7 | 37.1 | | 92.0 | | 84.0 |
| | SR1 | 13.7 | 2.5 | | | | | | | | | |
| | SR2 | | 10.0 | | | | | | | | | |
| | SR1 -3 | 13.6 | 4.0 | | | | | | | | | |
| | SR2 -3 | | 7.5 | | | | | | | | | |
| | SR1 +3 | 13.6 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 7.5 | | | | | | | | | |
| | JP1 | 12.1 | 48.5 | 9.5 | 5.0 | 2.4 | 15.6 | 13.2 | 96.0 | | 0.0 | 100.0 |
| | JP2 | | 52.0 | 12.7 | 4.8 | 6.0 | 18.7 | 12.7 | | 95.5 | 0.0 | 100.0 |
| 3 | Q.TP | 18.0 | 8.5 | 10.5 | 369.4 | -22.2 | 44.0 | 66.2 | 95.0 | 96.0 | 0.0 | 79.5 |
| | C.TP1 | 16.8 | 9.0 | 6.9 | 108.0 | -26.2 | 11.7 | 37.9 | 99.0 | | 0.0 | 95.0 |
| | C.TP2 | | | 19.2 | 85.7 | 9.2 | 47.6 | 38.4 | | 96.0 | | 84.5 |
| | SR1 | 14.8 | 5.0 | | | | | | | | | |
| | SR2 | | 10.5 | | | | | | | | | |
| | SR1 -3 | 14.9 | 5.0 | | | | | | | | | |
| | SR2 -3 | | 4.0 | | | | | | | | | |
| | SR1 +3 | 14.9 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 6.5 | | | | | | | | | |
| | JP1 | 12.8 | 54.5 | 9.1 | 5.5 | 2.3 | 15.1 | 12.8 | 95.0 | | 0.0 | 100.0 |
| | JP2 | | 55.0 | 12.2 | 5.4 | 5.5 | 18.6 | 13.1 | | 95.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 17.4 | 5.5 | 19.2 | 386.4 | -21.5 | 43.1 | 64.6 | 97.5 | 95.5 | 0.0 | 82.0 |
| | C.TP1 | 16.5 | 5.5 | 6.5 | 202.0 | -28.7 | 11.5 | 40.2 | 98.0 | | 0.0 | 97.0 |
| | C.TP2 | | | >20 | 271.5 | 9.4 | 48.7 | 39.3 | | 96.0 | | 86.0 |
| | SR1 | 14.5 | 5.0 | | | | | | | | | |
| | SR2 | | 4.0 | | | | | | | | | |
| | SR1 -3 | 14.3 | 4.0 | | | | | | | | | |

*Continued on next page*

Table D.11 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR2 -3 | | 6.5 | | | | | | | | | |
| | SR1 +3 | 14.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 3.5 | | | | | | | | | |
| | JP1 | 12.9 | 38.0 | 8.4 | 5.4 | 2.2 | 15.8 | 13.6 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 39.0 | 12.1 | 5.3 | 5.1 | 18.8 | 13.7 | | 96.0 | 0.0 | 100.0 |

See Table 6.4 for the definition of $\beta_0$, **S.M, Q.TP, C.TP, SR, JP, DV, SG%, ETP, SD, LCL, UCL, WD, CI1%, CI2%, NO%, IN%**.

## D.5.2 Two change points occur close to the beginning and end of data

### Number of data points=35 and change at 7 and 27

Table D.12: Number of data points=35 and the true turning points at 7 and 27.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 35.4 | 7.5 | >35 | >1000 | -120.0 | 164.9 | 284.9 | 92.0 | 98.0 | 1.5 | 50.5 |
| | C.TP1 | 30.9 | 56.5 | 7.0 | 48.8 | -15.4 | 14.6 | 30.0 | 96.0 | | 1.0 | 99.0 |
| | C.TP2 | | | 28.6 | 143.0 | 21.4 | 46.0 | 24.6 | | 98.0 | | 97.0 |
| | SR1 | 29.0 | 18.0 | | | | | | | | | |
| | SR2 | | 25.0 | | | | | | | | | |
| | SR1 -3 | 29.5 | 9.0 | | | | | | | | | |
| | SR2 -3 | | 34.5 | | | | | | | | | |
| | SR1 +3 | 29.2 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 13.5 | | | | | | | | | |
| | JP1 | 26.8 | 77.5 | 14.4 | 9.0 | 3.0 | 27.1 | 24.1 | 95.5 | | 0.0 | 100.0 |
| | JP2 | | 77.0 | 22.7 | 8.4 | 9.6 | 32.9 | 23.3 | | 95.5 | 0.0 | 100.0 |
| 3 | Q.TP | 32.5 | 4.0 | 12.2 | 497.8 | -56.0 | 96.8 | 152.8 | 95.0 | 96.0 | 0.5 | 75.5 |
| | C.TP1 | 31.0 | 9.5 | 9.4 | 127.0 | -49.1 | 19.2 | 68.3 | 96.0 | | 1.0 | 98.0 |
| | C.TP2 | | | 34.0 | 191.0 | 15.9 | 83.8 | 67.9 | | 99.5 | | 82.5 |
| | SR1 | 29.0 | 5.5 | | | | | | | | | |
| | SR2 | | 8.0 | | | | | | | | | |
| | SR1 -3 | 29.4 | 4.5 | | | | | | | | | |
| | SR2 -3 | | 10.5 | | | | | | | | | |
| | SR1 +3 | 29.1 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.5 | | | | | | | | | |
| | JP1 | 26.9 | 63.0 | 15.4 | 10.2 | 2.5 | 29.3 | 26.9 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 61.5 | 21.0 | 10.3 | 6.9 | 33.6 | 26.7 | | 97.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.9 | 4.0 | 13.4 | >1000 | -43.3 | 78.8 | 122.2 | 95.5 | 95.5 | 0.0 | 83.0 |
| | C.TP1 | 32.9 | 6.5 | 10.6 | 181.7 | -54.3 | 19.8 | 74.1 | 96.5 | | 2.0 | 97.0 |
| | C.TP2 | | | >35 | 265.7 | 15.5 | 89.5 | 74.0 | | 96.5 | | 82.5 |
| | SR1 | 31.0 | 3.5 | | | | | | | | | |
| | SR2 | | 5.5 | | | | | | | | | |
| | SR1 -3 | 31.2 | 2.0 | | | | | | | | | |
| | SR2 -3 | | 5.0 | | | | | | | | | |

*Continued on next page*

Table D.12 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR1 +3 | 30.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.5 | | | | | | | | | |
| | JP1 | 28.2 | 55.0 | 16.1 | 10.6 | 2.4 | 29.7 | 27.3 | 98.0 | | 0.0 | 100.0 |
| | JP2 | | 58.5 | 19.6 | 10.6 | 6.8 | 33.5 | 26.7 | | 97.5 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 32.0 | 15.5 | 16.3 | 483.8 | -57.6 | 87.0 | 144.6 | 90.0 | 81.5 | 0.5 | 74.5 |
| | C.TP1 | 30.4 | 16.0 | 11.0 | 370.2 | -48.5 | 19.3 | 67.8 | 96.5 | | 3.0 | 98.5 |
| | C.TP2 | | | >35 | 157.1 | 13.6 | 78.2 | 64.6 | | 92.0 | | 86.0 |
| | SR1 | 28.6 | 5.0 | | | | | | | | | |
| | SR2 | | 14.5 | | | | | | | | | |
| | SR1 -3 | 28.9 | 2.0 | | | | | | | | | |
| | SR2 -3 | | 20.0 | | | | | | | | | |
| | SR1 +3 | 28.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 7.5 | | | | | | | | | |
| | JP1 | 26.1 | 61.5 | 16.8 | 10.2 | 2.6 | 29.2 | 26.5 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 63.0 | 21.9 | 9.4 | 8.1 | 33.5 | 25.4 | | 97.0 | 0.0 | 100.0 |
| 3 | Q.TP | 32.5 | 5.0 | 8.7 | 664.6 | -43.6 | 76.4 | 120.0 | 94.0 | 95.0 | 1.0 | 83.5 |
| | C.TP1 | 31.4 | 6.5 | 11.1 | 165.0 | -54.7 | 20.2 | 74.9 | 96.5 | | 2.5 | 97.0 |
| | C.TP2 | | | >35 | 198.7 | 15.4 | 91.5 | 76.1 | | 98.5 | | 82.5 |
| | SR1 | 29.4 | 6.0 | | | | | | | | | |
| | SR2 | | 5.0 | | | | | | | | | |
| | SR1 -3 | 29.2 | 8.5 | | | | | | | | | |
| | SR2 -3 | | 6.5 | | | | | | | | | |
| | SR1 +3 | 29.5 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.0 | | | | | | | | | |
| | JP1 | 26.7 | 61.5 | 14.6 | 10.6 | 2.5 | 29.1 | 26.6 | 97.0 | | 0.0 | 100.0 |
| | JP2 | | 61.0 | 19.5 | 10.6 | 6.5 | 33.3 | 26.8 | | 97.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.3 | 6.0 | 16.6 | 439.3 | -43.1 | 80.2 | 123.3 | 90.5 | 94.0 | 0.0 | 81.5 |
| | C.TP1 | 32.3 | 6.0 | 11.1 | 170.0 | -49.2 | 20.0 | 69.2 | 96.0 | | 0.5 | 98.5 |
| | C.TP2 | | | >35 | 272.8 | 16.3 | 85.7 | 69.4 | | 99.5 | | 86.0 |
| | SR1 | 30.4 | 3.0 | | | | | | | | | |
| | SR2 | | 5.0 | | | | | | | | | |
| | SR1 -3 | 30.3 | 1.5 | | | | | | | | | |
| | SR2 -3 | | 7.5 | | | | | | | | | |
| | SR1 +3 | 30.4 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 3.0 | | | | | | | | | |
| | JP1 | 27.8 | 54.5 | 15.2 | 10.8 | 2.5 | 29.8 | 27.3 | 97.0 | | 0.0 | 100.0 |
| | JP2 | | 52.5 | 20.1 | 11.1 | 6.5 | 33.6 | 27.1 | | 97.0 | 0.0 | 100.0 |

Original model is combined model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 39.2 | 17.0 | 27.3 | >1000 | -193.7 | 242.1 | 435.8 | 85.0 | 99.0 | 1.0 | 36.5 |
| | C.TP1 | 31.5 | 81.0 | 5.9 | 37.3 | -8.4 | 10.6 | 19.0 | 93.5 | | 0.0 | 98.5 |
| | C.TP2 | | | 27.9 | 12.6 | 24.8 | 35.9 | 11.1 | | 87.0 | | 99.0 |
| | SR1 | 29.6 | 22.5 | | | | | | | | | |
| | SR2 | | 55.0 | | | | | | | | | |
| | SR1 -3 | 29.9 | 8.0 | | | | | | | | | |
| | SR2 -3 | | 70.5 | | | | | | | | | |
| | SR1 +3 | 30.0 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 19.5 | | | | | | | | | |
| | JP1 | 27.2 | 83.0 | 14.0 | 9.3 | 3.8 | 26.1 | 22.4 | 88.5 | | 0.0 | 100.0 |

*Continued on next page*

Table D.12 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JP2 | | 91.5 | 24.6 | 7.0 | 13.8 | 32.2 | 18.4 | | 93.5 | 0.0 | 100.0 |
| 3 | Q.TP | 32.4 | 7.0 | <1 | >1000 | -73.8 | 116.6 | 190.4 | 91.5 | 96.0 | 0.0 | 67.5 |
| | C.TP1 | 30.5 | 16.0 | 9.6 | 195.6 | -43.2 | 19.8 | 63.0 | 96.0 | | 2.0 | 95.5 |
| | C.TP2 | | | >35 | 320.7 | 16.4 | 75.9 | 59.5 | | 96.5 | | 90.0 |
| | SR1 | 28.4 | 11.0 | | | | | | | | | |
| | SR2 | | 13.5 | | | | | | | | | |
| | SR1 -3 | 28.2 | 7.5 | | | | | | | | | |
| | SR2 -3 | | 18.5 | | | | | | | | | |
| | SR1 +3 | 28.8 | 0.5 | | | | | | | | | |
| | SR2 +3 | | 3.5 | | | | | | | | | |
| | JP1 | 26.0 | 67.0 | 15.7 | 10.6 | 2.6 | 29.0 | 26.4 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 62.5 | 22.4 | 10.1 | 7.8 | 33.4 | 25.6 | | 98.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 33.5 | 5.0 | 24.5 | >1000 | -51.2 | 85.1 | 136.3 | 95.0 | 94.0 | 0.0 | 75.5 |
| | C.TP1 | 32.3 | 7.5 | 10.0 | 112.3 | -48.4 | 19.8 | 68.2 | 98.0 | | 0.5 | 98.5 |
| | C.TP2 | | | 30.0 | 297.6 | 15.6 | 83.6 | 68.0 | | 100.0 | | 88.0 |
| | SR1 | 30.0 | 3.0 | | | | | | | | | |
| | SR2 | | 6.5 | | | | | | | | | |
| | SR1 -3 | 30.5 | 2.0 | | | | | | | | | |
| | SR2 -3 | | 5.5 | | | | | | | | | |
| | SR1 +3 | 30.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.5 | | | | | | | | | |
| | JP1 | 27.8 | 51.5 | 16.4 | 10.2 | 2.4 | 29.5 | 27.1 | 98.5 | | 0.0 | 100.0 |
| | JP2 | | 51.5 | 21.8 | 9.7 | 7.1 | 33.6 | 26.5 | | 96.0 | 0.0 | 100.0 |

See Table 6.4 for the definition of $\beta_0$**, S.M, Q.TP, C.TP, SR, JP, DV, SG%, ETP, SD, LCL, UCL, WD, CI1%, CI2%, NO%, IN%.**

## Number of data points=20 and change at 5 and 15

Table D.13: Number of data points=20 and the true turning points at 5 and 15.

Original model is cubic model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 16.7 | 4.5 | 7.6 | >1000 | -24.4 | 44.0 | 68.4 | 94.5 | 94.5 | 0.0 | 83.5 |
| | C.TP1 | 15.7 | 4.5 | 6.0 | >1000 | -27.4 | 11.4 | 38.8 | 99.0 | | 0.5 | 97.0 |
| | C.TP2 | | | >20 | 146.9 | 9.4 | 47.8 | 38.4 | | 100.0 | | 86.5 |
| | SR1 | 13.5 | 7.5 | | | | | | | | | |
| | SR2 | | 6.0 | | | | | | | | | |
| | SR1 -3 | 13.4 | 5.0 | | | | | | | | | |
| | SR2 -3 | | 3.5 | | | | | | | | | |
| | SR1 +3 | 13.7 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 4.5 | | | | | | | | | |
| | JP1 | 11.8 | 44.5 | 8.7 | 5.2 | 2.4 | 15.1 | 12.8 | 92.5 | | 0.0 | 100.0 |
| | JP2 | | 49.5 | 12.2 | 5.4 | 5.7 | 18.4 | 12.7 | | 93.5 | 0.0 | 100.0 |
| 3 | Q.TP | 17.0 | 4.0 | 5.9 | 283.5 | -22.6 | 43.8 | 66.4 | 93.0 | 94.0 | 0.0 | 83.0 |
| | C.TP1 | 16.1 | 5.5 | 6.0 | 100.5 | -31.6 | 11.5 | 43.1 | 97.5 | | 0.5 | 96.0 |
| | C.TP2 | | | >20 | 165.3 | 9.1 | 49.3 | 40.2 | | 98.5 | | 86.5 |
| | SR1 | 14.1 | 1.5 | | | | | | | | | |
| | SR2 | | 4.0 | | | | | | | | | |
| | SR1 -3 | 14.2 | 5.0 | | | | | | | | | |

*Continued on next page*

Table D.13 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SR2 -3 | | 4.0 | | | | | | | | | |
| | SR1 +3 | 13.7 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 7.0 | | | | | | | | | |
| | JP1 | 12.3 | 43.5 | 8.6 | 5.1 | 2.3 | 15.8 | 13.5 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 48.5 | 12.0 | 5.4 | 5.5 | 18.7 | 13.2 | | 96.0 | 0.0 | 100.0 |
| 1.5 | Q.TP | 18.0 | 5.0 | >20 | 142.8 | -22.5 | 43.3 | 65.8 | 91.5 | 95.0 | 0.0 | 85.5 |
| | C.TP1 | 17.1 | 5.5 | 6.4 | 61.4 | -24.8 | 11.4 | 36.2 | 97.0 | | 0.0 | 96.5 |
| | C.TP2 | | | 19.5 | 190.6 | 9.4 | 47.0 | 37.6 | | 99.0 | | 89.0 |
| | SR1 | 15.2 | 3.5 | | | | | | | | | |
| | SR2 | | 3.5 | | | | | | | | | |
| | SR1 -3 | 15.0 | 1.0 | | | | | | | | | |
| | SR2 -3 | | 3.0 | | | | | | | | | |
| | SR1 +3 | 14.8 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 2.5 | | | | | | | | | |
| | JP1 | 13.4 | 40.0 | 9.4 | 5.1 | 2.2 | 15.8 | 13.6 | 97.5 | | 0.0 | 100.0 |
| | JP2 | | 38.0 | 12.8 | 4.9 | 5.1 | 18.8 | 13.7 | | 98.0 | 0.0 | 100.0 |

Original model is segmented regression model

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 17.7 | 7.0 | 13.3 | 574.0 | -24.8 | 44.9 | 69.7 | 91.0 | 89.5 | 0.0 | 78.5 |
| | C.TP1 | 16.4 | 7.0 | 6.9 | 144.0 | -27.0 | 11.6 | 38.6 | 99.0 | | 1.0 | 96.5 |
| | C.TP2 | | | >20 | 79.0 | 9.0 | 48.6 | 39.6 | | 99.0 | | 85.5 |
| | SR1 | 14.2 | 8.0 | | | | | | | | | |
| | SR2 | | 8.0 | | | | | | | | | |
| | SR1 -3 | 14.3 | 4.0 | | | | | | | | | |
| | SR2 -3 | | 8.5 | | | | | | | | | |
| | SR1 +3 | 14.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 4.5 | | | | | | | | | |
| | JP1 | 12.7 | 54.5 | 8.9 | 5.1 | 2.4 | 15.4 | 13.0 | 96.5 | | 0.0 | 100.0 |
| | JP2 | | 53.5 | 12.2 | 5.2 | 5.7 | 18.7 | 12.9 | | 95.5 | 0.0 | 100.0 |
| 3 | Q.TP | 17.5 | 5.5 | 8.8 | 330.0 | -22.9 | 43.9 | 66.8 | 95.5 | 93.0 | 0.0 | 81.5 |
| | C.TP1 | 16.5 | 5.5 | 6.2 | 467.0 | -28.6 | 11.5 | 40.1 | 97.5 | | 0.0 | 96.0 |
| | C.TP2 | | | >20 | 135.8 | 9.3 | 47.8 | 38.5 | | 99.0 | | 85.5 |
| | SR1 | 14.3 | 5.5 | | | | | | | | | |
| | SR2 | | 5.5 | | | | | | | | | |
| | SR1 -3 | 14.6 | 3.0 | | | | | | | | | |
| | SR2 -3 | | 6.0 | | | | | | | | | |
| | SR1 +3 | 14.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 1.5 | | | | | | | | | |
| | JP1 | 12.6 | 54.0 | 9.6 | 5.6 | 2.4 | 15.4 | 13.0 | 94.5 | | 0.0 | 100.0 |
| | JP2 | | 51.0 | 12.7 | 5.2 | 5.5 | 18.6 | 13.1 | | 95.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 18.3 | 3.0 | 8.7 | 185.9 | -20.9 | 42.7 | 63.5 | 91.5 | 96.5 | 0.0 | 83.0 |
| | C.TP1 | 17.3 | 5.5 | 6.4 | 140.9 | -27.2 | 11.3 | 38.5 | 97.0 | | 0.0 | 96.5 |
| | C.TP2 | | | 18.2 | 95.1 | 9.5 | 49.3 | 39.8 | | 99.0 | | 82.0 |
| | SR1 | 15.2 | 4.0 | | | | | | | | | |
| | SR2 | | 4.0 | | | | | | | | | |
| | SR1 -3 | 14.8 | 2.5 | | | | | | | | | |
| | SR2 -3 | | 4.5 | | | | | | | | | |
| | SR1 +3 | 15.3 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 0.0 | | | | | | | | | |
| | JP1 | 13.2 | 46.0 | 8.8 | 5.1 | 2.4 | 15.3 | 12.9 | 93.0 | | 0.0 | 100.0 |
| | JP2 | | 44.5 | 11.9 | 5.1 | 5.6 | 18.5 | 12.9 | | 94.5 | 0.0 | 100.0 |

Original model is combined model

*Continued on next page*

Table D.13 – *Continued from previous page*

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

| $\beta_0$ | S.M | DV | SG% | ETP | SD | LCL | UCL | WD | CI1% | CI2% | NO% | IN% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Q.TP | 17.7 | 5.0 | <1 | 566.4 | -45.8 | 68.2 | 114.0 | 89.5 | 99.0 | 0.0 | 63.5 |
| | C.TP1 | 16.1 | 8.0 | 5.7 | 281.0 | -26.6 | 11.4 | 38.0 | 99.0 | | 0.5 | 94.5 |
| | C.TP2 | | | >20 | 432.6 | 9.2 | 47.1 | 37.9 | | 95.5 | | 85.0 |
| | SR1 | 14.2 | 7.0 | | | | | | | | | |
| | SR2 | | 8.5 | | | | | | | | | |
| | SR1 -3 | 14.2 | 5.5 | | | | | | | | | |
| | SR2 -3 | | 9.5 | | | | | | | | | |
| | SR1 +3 | 14.0 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 3.5 | | | | | | | | | |
| | JP1 | 12.3 | 54.5 | 8.5 | 5.0 | 2.3 | 15.0 | 12.7 | 98.0 | | 0.0 | 100.0 |
| | JP2 | | 56.0 | 11.7 | 5.0 | 6.0 | 18.5 | 12.5 | | 97.0 | 0.0 | 100.0 |
| 3 | Q.TP | 17.1 | 5.5 | 12.4 | 311.2 | -30.6 | 50.7 | 81.3 | 95.0 | 94.0 | 0.0 | 76.5 |
| | C.TP1 | 15.9 | 5.5 | 6.5 | 123.5 | -29.8 | 11.7 | 41.5 | 99.0 | | 0.5 | 95.0 |
| | C.TP2 | | | >20 | 78.7 | 8.9 | 52.1 | 43.2 | | 99.0 | | 82.0 |
| | SR1 | 14.2 | 7.0 | | | | | | | | | |
| | SR2 | | 7.0 | | | | | | | | | |
| | SR1 -3 | 13.7 | 9.0 | | | | | | | | | |
| | SR2 -3 | | 5.5 | | | | | | | | | |
| | SR1 +3 | 14.2 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 5.0 | | | | | | | | | |
| | JP1 | 12.1 | 49.5 | 8.4 | 5.1 | 2.5 | 15.2 | 12.7 | 93.0 | | 0.0 | 100.0 |
| | JP2 | | 50.0 | 11.6 | 5.4 | 5.8 | 18.6 | 12.8 | | 93.5 | 0.0 | 100.0 |
| 1.5 | Q.TP | 17.6 | 5.5 | 4.4 | 506.8 | -23.4 | 43.5 | 66.9 | 94.5 | 94.0 | 0.0 | 84.0 |
| | C.TP1 | 16.6 | 5.5 | 5.9 | 130.8 | -31.8 | 11.2 | 43.0 | 99.0 | | 0.0 | 95.0 |
| | C.TP2 | | | 19.1 | 87.3 | 9.1 | 51.7 | 42.6 | | 98.0 | | 80.0 |
| | SR1 | 14.5 | 4.5 | | | | | | | | | |
| | SR2 | | 3.0 | | | | | | | | | |
| | SR1 -3 | 14.5 | 1.0 | | | | | | | | | |
| | SR2 -3 | | 6.5 | | | | | | | | | |
| | SR1 +3 | 14.6 | 0.0 | | | | | | | | | |
| | SR2 +3 | | 0.5 | | | | | | | | | |
| | JP1 | 13.0 | 39.5 | 8.4 | 5.3 | 2.2 | 15.7 | 13.5 | 98.5 | | 0.0 | 100.0 |
| | JP2 | | 38.0 | 11.9 | 5.3 | 5.0 | 18.8 | 13.8 | | 98.0 | 0.0 | 100.0 |

See Table 6.4 for the definition of $\beta_0$**, S.M, Q.TP, C.TP, SR, JP, DV, SG%, ETP, SD, LCL, UCL, WD, CI1%, CI2%, NO%, IN%**.

# Appendix E

# Spline GAM Regression - Chapter 7

## E.1 R code for estimating change points from GAM model and their confidence intervals

See Section 7.2 for the algorithm.

```
library(XLConnect)
library(mgcv) # To fit a generalized additive model (GAM) to data
library(boot)


d1 <- readWorksheetFromFile("JoinpointDATA-uptoMarch2016.xlsx",
sheet=1, header=T)
d1$Qu <- factor(d1$Qu)

#To fit gam model for MRSA
z1 <- gam(no.MRSA1 ~  offset(log(aobd))+ s(time, bs="cr", k=7)+Qu ,
 family=poisson, data=d1)
summary(z1)

########## Estimate change points ##########
# Predicted values from spline function
z.pred <- predict(z1, type="terms", terms="s(time)")

dy <- diff(z.pred) # first derivative
d2y <- diff(diff(z.pred)) # second derivative

l.tp <- which(diff(sign(dy))!=0) # locations of first derivative
Time.tp <- (l.tp/4) + 2003  # time associated with first derivative
```

```
# To give a set includes all maximum points
otp <- which(sign(d2y) == -1)
if(l.tp[1] %in% otp) 1 else NA
if(l.tp[2] %in% otp) 1 else NA
if(l.tp[3] %in% otp) 1 else NA

########## Construct confidence intervals ##########
z.res <- residuals(z1,type="pearson")
z.predict <- predict(z1,type="response")# predicted counts
data <- cbind(d1, resid = z.res, fit = z.predict)
ch <- round(length(no.MRSA1)*10/100,0) # length of 10% of data
ltp <- length(Time.tp) # length of number of change points

n.boots <- 10000
out1 <- matrix(NA, nrow=n.sim, ncol=8)

for(i in 1:n.boots) {
data$n.res <- sample (z.res, rep=T)
data$new.n <- round(data$fit + data$n.res*sqrt(data$fit),0)

rz2 <- gam(new.n ~ offset(log(aobd))+ s(time, bs="cr", k=7)+ Qu ,
family=poisson , data=data)
zb.pred <- predict(rz2, type="terms", terms="s(time)")
dyb <- diff(zb.pred)
d2yb <- diff(diff(zb.pred))

# Location when first derivative change sign,
#(location of all possible change points)

l.tpb <- which(diff(sign(dyb))!=0)
le <- length(l.tpb)
if(le < ltp) next
if (le > ch) next

# The values of s(time) ((zb.pred)) at change points.
st.pred <- zb.pred[l.tpb+1]

# The maximum value from s(time) at change points
max1 <- which.max(st.pred)

# Give the location of the maximum from the set of all possible change points
fmax <- l.tpb[max1]

# Calculate the time of the largest value
t.max <- ((fmax )/4) + 2003
Time.tpb <- (l.tpb/4) + 2003

otpb <- which(sign(d2yb) == -1)

mx1 <- if(l.tpb[1] %in% otpb) Time.tpb[1] else NA
mx2 <- if(l.tpb[2] %in% otpb) Time.tpb[2] else NA
mx3 <- if(l.tpb[3] %in% otpb) Time.tpb[3] else NA
mx4 <- if(l.tpb[4] %in% otpb) Time.tpb[4] else NA
mx5 <- if(l.tpb[5] %in% otpb) Time.tpb[5] else NA
mx6 <- if(l.tpb[6] %in% otpb) Time.tpb[6] else NA

result <- c(le, t.max, mx1, mx2, mx3, mx4, mx5, mx6 )
out1[i,] <- result
gc()
```

```
}

dimnames(out1)[[2]]<- c("no.TP", "max","mx1","mx2",
                        "mx3","mx4","mx5","mx6")

out11 <- data.frame(out1)

# To remove samples have NA,
#(i.e. length of le> ch or length of le< ltp)
OUT1.1 <- subset(out11, no.TP!="NA")
nrow(OUT1.1)

# Quantile bootstrap CI
CIL <- quantile(OUT1.1$max,c(0.025,0.975))

########## Plot fitted model,
# estimated change points and CIs ##########
plot(d1$time, z.pred, col="black", type="l",
xlab="time", ylab="s(time)", lwd=2)
abline(v=Time.tp, col=2, lty=1, lwd=2)
abline(v=CIL[1], col=2, lty=2, lwd=2)
abline(v=CIL[2], col=2, lty=2, lwd=2)
```