Fall 2014

An Investigation of Sensitivity of an F Test in Locating Change Points in Linear Regression

Jing Sun

AN INVESTIGATION OF SENSITIVITY OF AN F TEST IN LOCATING CHANGE POINTS IN LINEAR REGRESSION

by

JING SUN

(Under the Direction of Patrica Humphrey)

ABSTRACT

Change point is a statistic phenomenon, which has many direct applications in climatology, bioinformatics, finance, oceanography and medical imaging. In this thesis, we investigate the sensitivity of the F-test for detecting change points in linear regression, using a two-phase linear regression model. The two-phase regression model was first introduced by Lund and Reeves [9]; it offers an effective method to detect "undocumented" change points using a form of an F-test. Using simulated data, we explore its sensitivity and accuracy with respect to different parameters in the model.

Key Words: change-point; linear regression; sensitivity; F test

2009 *Mathematics Subject Classification*: 62J05,62J99

AN INVESTIGATION OF SENSITIVITY OF AN F TEST IN LOCATING CHANGE POINTS IN LINEAR REGRESSION

by

JING SUN

B.S. in Electrical Engineering - Beijing University of Posts and Telecommunications
(2008)

A Thesis Submitted to the Graduate Faculty of Georgia Southern University
in Partial Fulfillment
of the Requirement for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

2014

AN INVESTIGATION OF SENSITIVITY OF AN F TEST IN LOCATING CHANGE POINTS IN LINEAR REGRESSION

by

JING SUN

Major Professor: Patrica Humphrey

Committee: Broderick Oluyede
Arpita Chatterjee

Electronic Version Approved:

December 17, 2014

DEDICATION

I dedicate my work to my beloved parents. Because of their continuous support, I was able to pursue my dream of being a professional statistician in the future. They always encourage and have faith in me even when I was a kid; they also never give me any pressure and force me to do anything. I express my most sincere appreciation to them, because they gave me the happiest and most memorable childhood. Without their love and continuing support, I would not make it this far.

ACKNOWLEDGMENTS

I would like to thank my direct academic advisor Dr. Patricia Humphrey for making it possible for me to finish this thesis in a desired amount of time. Many discussions and the interactions with Dr. Humphrey had a direct impact on the final form and quality of my thesis.

I would also like to thank Dr. Broderick Oluyede for his valuable comments and time toward my thesis, as well as his wonderful lectures during my time as a graduate student.

I express my sincere thanks to Dr. Arpita Chatterjee for being my committee member.

My special thanks go to Dr. Hua Wang for his helpfulness and wonderful advices to many different problems in both college life and real life.

I am very grateful for my parents, who have supported me for all my life. They provide a persistent inspiration for my journey.

I am also very thankful to Anh Tran for his patience and support during the master's period.

TABLE OF CONTENTS

Page

DEDICATION v

ACKNOWLEDGMENTS vi

LIST OF FIGURES ix

LIST OF SYMBOLS xi

CHAPTER

1 Introduction 1

1.1 Change Point Detection in Linear Regression 1

1.2 Linear Regression Model 2

1.3 The method of Least Squares 3

1.4 Basic Notation 4

1.4.1 $SSTO$ 4

1.4.2 SSE 5

1.4.3 SSR 6

1.5 Change point models 6

1.5.1 Discontinuous change point model 7

1.5.2 Continuous change point model 7

1.5.3 The sensitivity of a test 10

1.6 Work outline 10

2 Methodology and arguments 11

2.1 Two-phase linear model 11

2.2 The Full Model 12

2.3 The Reduced Model 13

2.4 An introduction to F_{max} statistics 14

3 Parametric study: Numerical simulation and results analysis . . . 18

3.1 Design of numerical simulation and its purpose 18

3.2 Simulation Procedure Description 20

4 Simulation results analysis 22

4.1 Slope change 22

4.2 The proportion of sample size change 24

4.3 The sample size change 26

4.4 The Standard Deviation change 28

4.5 Sensitivity as a function of two variables 31

5 Summary 34

5.1 Discussion 34

5.2 Conclusion 35

5.3 Future work 35

REFERENCES 37

A SAS Source Codes 39

LIST OF FIGURES

Figure Page

1.1 An example of a Regression Line 3

1.2 An illustration of the Total Sum of Squares. 5

1.3 An illustration of the Error Sum of Squares. 6

1.4 An illustration of the Regression Sum of Squares. 7

1.5 Discontinuous change point model with only one change point. . . 8

1.6 Discontinuous change point model with multiple change points. . . 8

1.7 Continuous change point model with only one change point 9

1.8 Continuous change point model with multiple change points. . . . 9

2.1 An illustration for the full model. 13

2.2 An illustration for the reduced model. 14

2.3 Step change in constant term α_i: $\sigma = 12; \alpha_1 = 20; \alpha_2 = -20; \beta_1 = \beta_2 = 20$ 15

2.4 Trend change in slope term β_i: $\sigma = 12; \beta_1 = 20; \beta_2 = -20$ 15

4.1 $\hat{\beta}_2$ vs. C_mean in Slope change simulation 23

4.2 $\hat{\beta}_2$ vs. C_std dev in Slope change simulation 24

4.3 n_2 vs. C_mean in Proportion sample size change simulation 25

4.4 n_2 vs. C_std dev in Proportion sample size change simulation . . . 26

4.5 n_2 vs. C_std dev in Sample size change simulation 28

4.6 Std dev vs. C_mean in Standard deviation change simulation . . . 29

4.7 Std dev vs. C_std dev in Standard deviation change simulation . . 30

4.8 C_mean vs. Std dev in Proportion Sample Size and Variance change 31

4.9 C_std dev vs. Std dev in Proportion Sample Size and Std dev change 33

LIST OF SYMBOLS

- α: Intercept of the population regression line
- β: Slope of the population regression line
- X: Predictor variable
- Y: Response variable
- σ: Standard deviation of points about the regression model
- ε: Population random error term with a normal distribution
- $\hat{\alpha}$: Estimated intercept of the regression line
- $\hat{\beta}$: Estimated slope of the regression line
- $\bar{X}$: The mean of the predictor variable
- $\bar{Y}$: The mean of the response variable
- $SSTO$: The sum of the squared deviations of the y values from the mean
- SSE: The error sum of squares (unexplained variation in the response variable)
- SSR: The regression sum of squares (amount of variation in the response variable explained by the model)
- c: The position of the change point
- n_1: Number of points before the change
- n_2: Number of points after the change

CHAPTER 1

INTRODUCTION

1.1 Change Point Detection in Linear Regression

Linear regression analysis is perhaps the most widely used statistical technique. It has many important applications, including finance, clinical trials, economics, management, biology, physics and many other areas. As stated in [4], there are three main reasons for its popularity:

- the relative easiness and usefulness
- the wide integration into every statistical software package
- the power and flexibility in tackling big data

The reason why we assume the relationship between variables is linear is not only because this is the simplest relationship between quantitative variables, but also the true relationship between variables is often approximately linear over the range of observed values. Even if the ”true” relationship is not linear, it might be possible to transform the data into a linear form.

However, it is not always true that the same linear model holds for the whole data set. The model may change after a specific point which partitions the data into two segments which have different models. Thus, a linear model with a change point is fitted for data sets where the structure of the linear model changes after a specific point.

According to Killick[6], change point detection means estimating *the point at which the statistical properties of a sequence of observations change.* There is a growing need to be able to identify the change point in a linear regression model. Detecting such changes is widely used in climatology, bioinformatic applications, finance, oceanography and medical imaging. In 2002, Lund and Reeves [9] proposed a revision of the two-phase linear regression test for change point detection at an undocumented time. The main interest of this thesis is to investigate the sensitivity of this F test in locating a change point in linear regression.

1.2 Linear Regression Model

Generally, linear regression analysis is categorized according to the number of predictor variables. The simple linear regression model only has one predictor variable and is linear in the parameters. It means that no parameter appears as an exponent or a multiplication or division with another parameter. Thus, we typically present a simple linear regression model using the following function:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{1.1}$$

where

- Y_i is the response variable value of an observation.
- X_i is the predictor variable value.
- α is the intercept, β is the slope; both of them are parameters.
- ε_i is a random error term with a normal distribution: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The error terms are assumed to be independent, with a mean of zero and constant variance, and normally distributed. The concept is illustrated in the Figure 1.1 below as a line on a scatterplot.

A multiple linear regression model, as indicated by its name, has several predictor variables and a linear regression function. The multiple regression model has the form:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \tag{1.2}$$

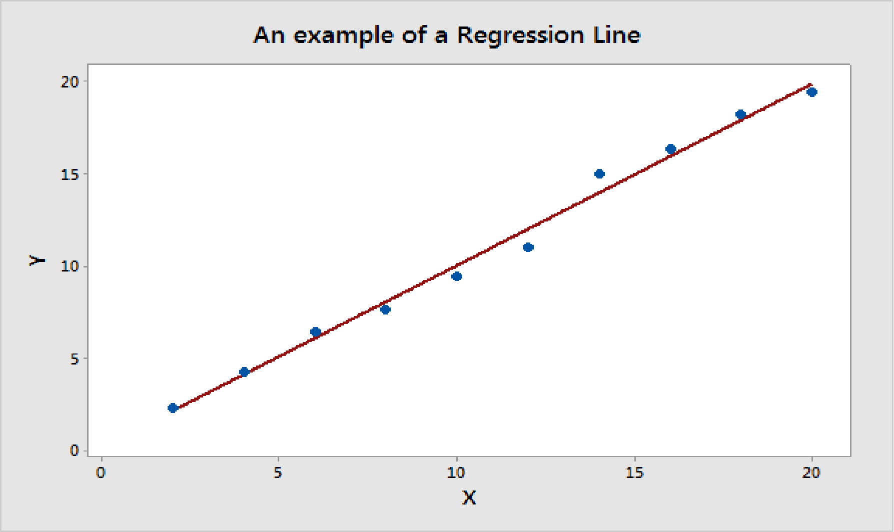

Figure 1.1: An example of a Regression Line

where ε is also $\mathcal{N}(0, \sigma^2)$. As in the simple linear regression model, the error terms in the multiple linear regression model are also assumed to be independent.

1.3 The method of Least Squares

Following [7], we use the least squares method to find good estimators for the regression parameters α and β. Define the deviation as the difference between Y_i and its expected value $(\alpha + \beta X_i)$, and denote Q as the sum of the n squared deviations; we have

$$Q = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 \tag{1.3}$$

The goal is to find the value of α and β such that the sum of the n squared deviations for the given sample observations is minimized. In order to minimize Q, we should find the partial derivatives with respect to α and β.

$$\frac{\partial Q}{\partial \alpha} = -2 \sum (Y_i - \alpha - \beta X_i), \tag{1.4}$$

$$\frac{\partial Q}{\partial \beta} = -2 \sum X_i (Y_i - \alpha - \beta X_i). \tag{1.5}$$

Letting the partial derivatives equal zero and using $\hat{\alpha}$ and $\hat{\beta}$ to represent the estimators of α and β gives

$$-2\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \tag{1.6}$$

$$-2\sum X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0. \tag{1.7}$$

Dividing both sides by the constant term yields

$$\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \tag{1.8}$$

$$\sum X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0. \tag{1.9}$$

Expanding the total sum of the quantities, we have

$$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum X_i = 0 \tag{1.10}$$

$$\sum X_iY_i - \hat{\alpha}\sum X_i - \hat{\beta}\sum x_i^2 = 0. \tag{1.11}$$

The solutions $\hat{\alpha}$ and $\hat{\beta}$ are as follows:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{1.12}$$

$$\hat{\alpha} = \frac{1}{n}\Big(\sum Y_i - \hat{\beta}\sum X_i\Big) = \bar{Y} - \hat{\beta}\bar{X}. \tag{1.13}$$

1.4 Basic Notation

Considering a set of observations (X_i, Y_i), we set up a few simple notations and terminologies for the sake of simplicity of later use.

1.4.1 *SSTO*

SSTO stands for the total sum of squares, which is a measure of the total variation in the response variable. It is the sum of the squared distances between the observations Y_i and the mean of the observed values $\bar{Y}$. The $\bar{Y}$ is calculated as

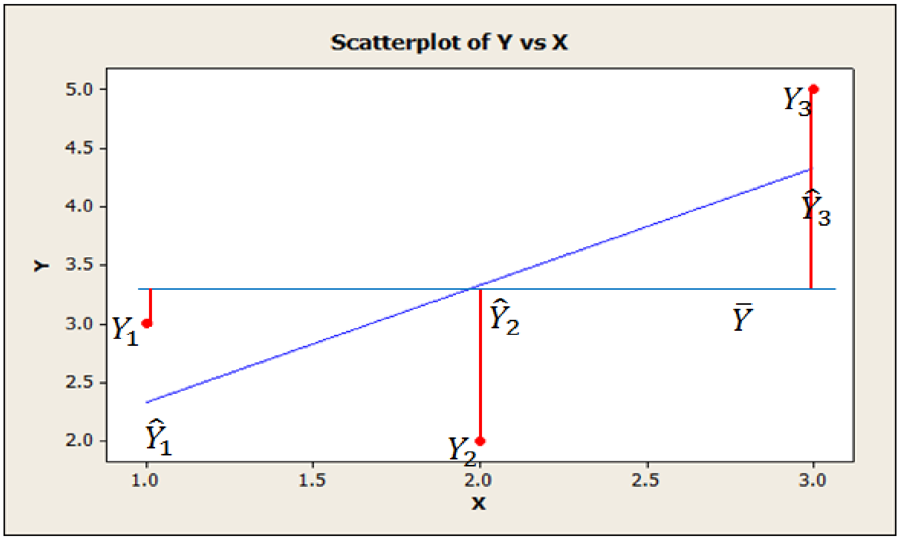

Figure 1.2: An illustration of the Total Sum of Squares.

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}, \tag{1.14}$$

$$SSTO = \sum (Y_i - \bar{Y})^2. \tag{1.15}$$

The larger the variation is, the larger the $SSTO$. See Figure 1.2 for an illustration.

1.4.2 SSE

SSE is an abbreviation for the error sum of squares; it measures the deviation between the observation Y_i and the expected value $\hat{Y}$. The mathematical expression for SSE is

$$SSE = \sum (Y_i - \hat{Y}_i)^2. \tag{1.16}$$

If the observations of Y_i have larger variation around the fitted regression line, we get a larger SSE, which indicates a poorer fit of the model to the data. For illustration, see Figure 1.3.

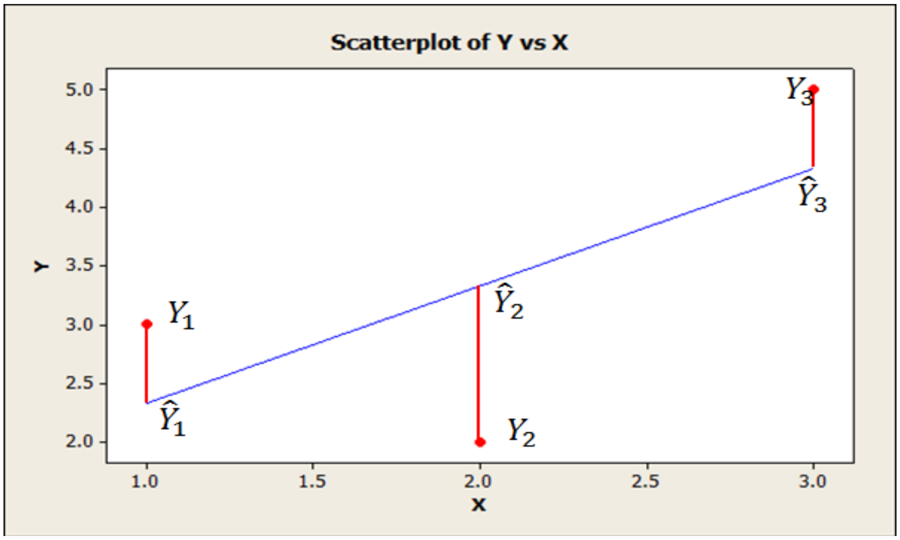

Figure 1.3: An illustration of the Error Sum of Squares.

1.4.3 SSR

SSR is an abbreviation for the regression sum of squares; it measures the deviation between the expected value $\hat{Y}$ and the mean value Y_i, which is the amount of variation in Y explained by the model. For illustration, see Figure 1.4. SSR is given by

$$SSR = \sum(\hat{Y}_i - \bar{Y})^2. \tag{1.17}$$

It can be proved that the total sum of squares is the sum of the regression sum of squares and the error sum of squares, i.e.

$$SSTO = SSR + SSE. \tag{1.18}$$

1.5 Change point models

Johannes Hofrichter, when writing his Ph.D. dissertation [5], distinctly separated two main types of change points regarding the continuity of the data set: discontinuous and continuous change points.

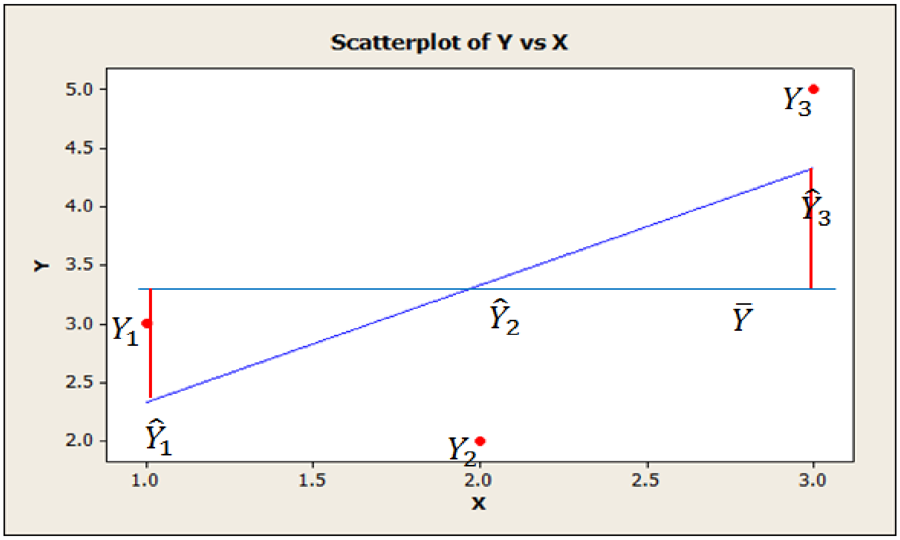

Figure 1.4: An illustration of the Regression Sum of Squares.

1.5.1 Discontinuous change point model

Denote Y as the response variable and X as the explanatory variable. It is assumed that the relationship between X and Y is a simple linear regression in a neighborhood. In this model, after the change point, the regression equation abruptly changes. Therefore, the discontinuous change point model has no continuity constraint at the point of change. The concept is illustrated in Figure 1.5.

This model can be further extended to the case of multiple discontinuous change points. Again, α and β are the parameters, while ε are independent errors with normal distribution $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The concept is illustrated in Figure 1.6.

1.5.2 Continuous change point model

The second type of change point model is the continuous change point model. At the point of change, the mean of the data set does not jump, but instead, smoothly follows another regression line. A continuity constraint is thus imposed on the point of change. Mathematically, if we denote the point of change of change as c, the model

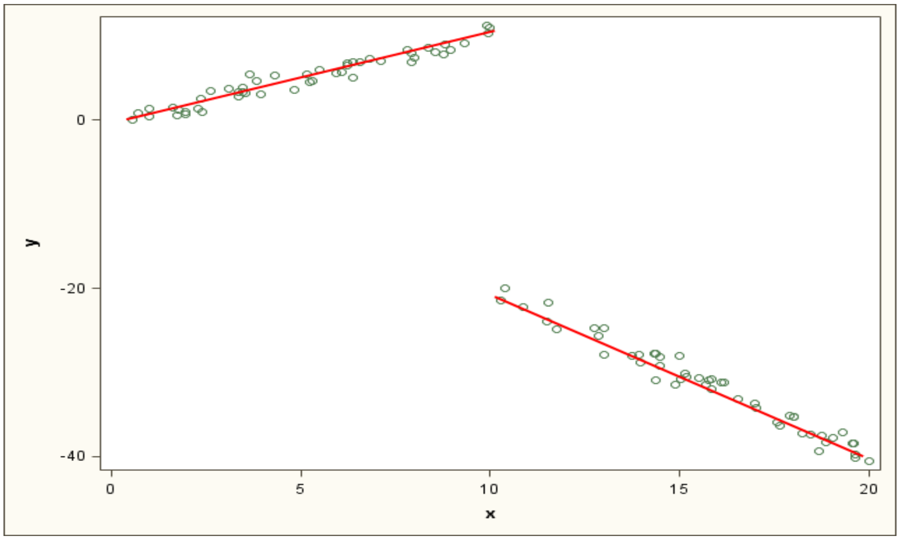

Figure 1.5: Discontinuous change point model with only one change point.

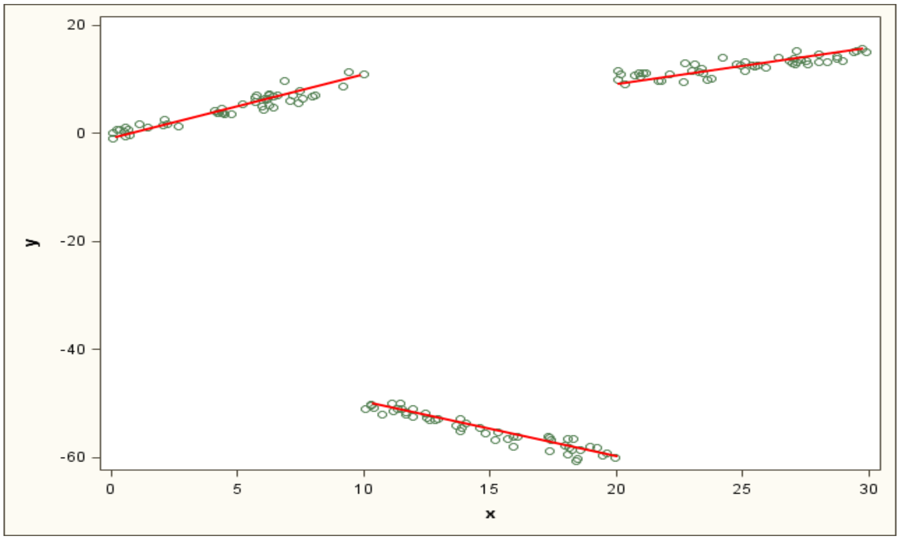

Figure 1.6: Discontinuous change point model with multiple change points.

can be rewritten as

$$Y_i = \begin{cases} \alpha_1 + \beta_1 X_i + \varepsilon_i, & X_i < c, \\ \alpha_2 + \beta_2 X_i + \varepsilon_i, & X_i > c \end{cases} \tag{1.19}$$

with continuity constraint at the point of change c

$$\alpha_1 + \beta_1 c = \alpha_2 + \beta_2 c. \tag{1.20}$$

This model can also be extended to multiple continuous change points. At each change point, an additional continuity constraint is imposed. The concepts are illustrated in Figure 1.7 and Figure 1.8.

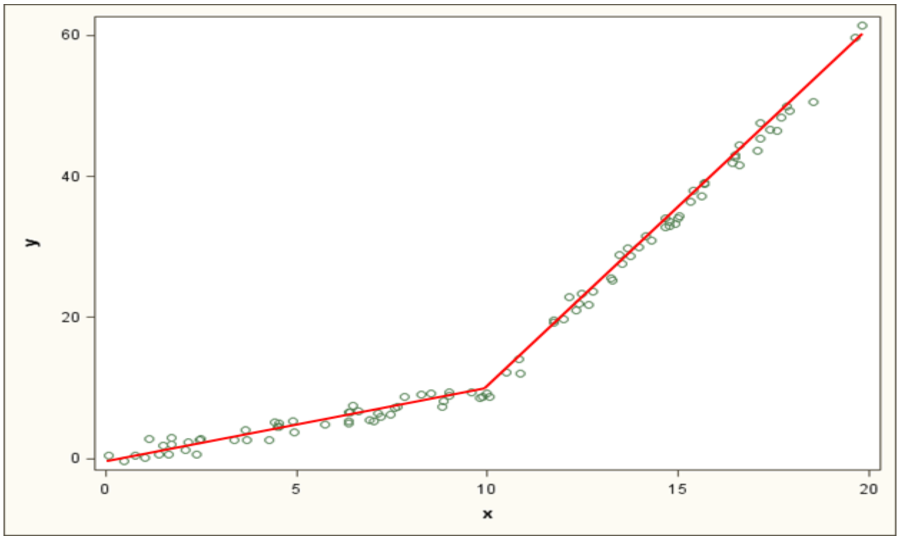

Figure 1.7: Continuous change point model with only one change point

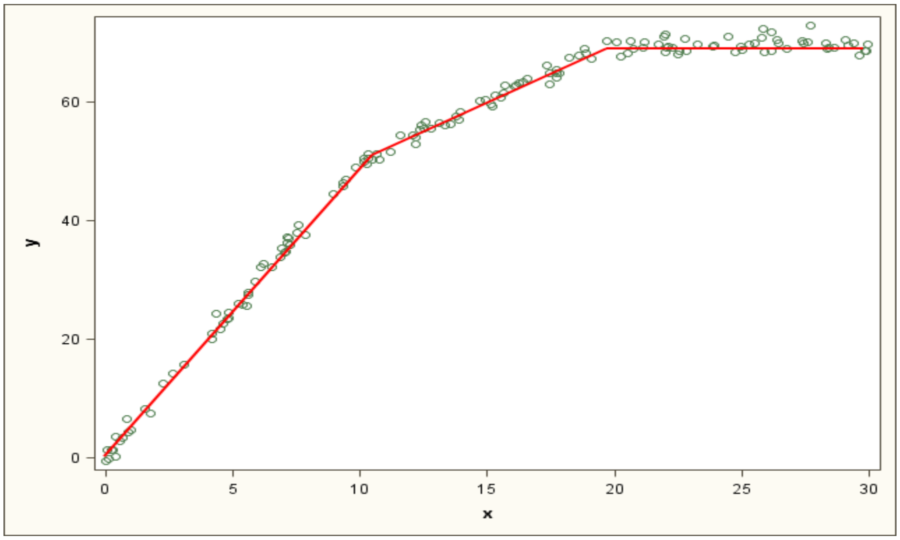

Figure 1.8: Continuous change point model with multiple change points.

1.5.3 The sensitivity of a test

The discontinuous change point model is quite obvious, compared to the continuous model, due to the abrupt change of the mean at the point of change.

Strictly speaking, the sensitivity to detect the point of change c depends largely on three quantities: the slopes β_i before and after the change, the standard deviation of the error $\sigma : \varepsilon \sim \mathcal{N}(0, \sigma^2)$ and the sample size, i.e. $n_1 = |\{x : x < c\}|$ and $n_2 = |\{x : x > c\}|$.

Because of numerous permutations of different parameters involved in the test, we limit our work to only one change point model, as in Figure 1.7.

1.6 Work outline

In Chapter 1, we introduced the readers to the relative concepts of change point in regression analysis. Chapter 2 provides a theoretical background of the F-test, first derived by Lund and Reeves [9], to detect the change point. In Chapter 3, we propose a parametric study framework to investigate the sensitivity of the F-test, with respect to different parameters. Chapter 4 presents the simulation results analysis based on the designed guideline of Chapter 3. The SAS codes used in this thesis to run the statistical simulation is documented in Appendix A for future reference.

CHAPTER 2

METHODOLOGY AND ARGUMENTS

2.1 Two-phase linear model

Recently, the area of regression change point analysis has drawn much attention from statistical researchers. Numerous papers have been published with respect to many different methods to detect the change point. In 2002, Lund and Reeves [9] proposed a revision of the two-phase linear regression test for change point detection at an undocumented time. Statistically speaking, change point is defined as the position where the series shifts dramatically.

The model we considered can be succinctly expressed as

$$Y_i = \begin{cases} \alpha_1 + \beta_1 X_i + \varepsilon_i, & X_i < c \\ \alpha_2 + \beta_2 X_i + \varepsilon_i, & X_i > c \end{cases} \tag{2.1}$$

where α_i, β_i and σ are all parameters, and thus, subject to change under the user's will in a simulation setting. We will investigate the successful rate by detection by varying the parameters α_i, β_i and σ.

Suppose that c is known to be the only point of change, then the least squares estimates of the trend parameters in Equation 2.1 are

$$\hat{\beta}_1 = \frac{\sum_{X_i<c}(X_i - \bar{X}_1)(Y_i - \bar{Y}_1)}{\sum_{X_i<c}(X_i - \bar{X}_1)^2} \tag{2.2}$$

and

$$\hat{\beta}_2 = \frac{\sum_{X_i>c}(X_i - \bar{X}_2)(Y_i - \bar{Y}_2)}{\sum_{X_i>c}(X_i - \bar{X}_2)^2}, \tag{2.3}$$

where

$$\bar{Y}_1 = \frac{\sum_{X_i<c} Y_i}{n_1} \tag{2.4}$$

and

$$\bar{Y_2} = \frac{\sum_{X_i > c} Y_i}{n_2} \tag{2.5}$$

are the average series Y values before and after c, respectively; and

$$\bar{X_1} = \frac{\sum_{X_i < c} X_i}{n_1} \tag{2.6}$$

and

$$\bar{X_2} = \frac{\sum_{X_i > c} X_i}{n_2} \tag{2.7}$$

are the average X values before and after c, respectively.

The least squares estimates of the location parameters α_1 and α_2 in (2.1) are

$$\hat{\alpha}_1 = \bar{Y_1} - \hat{\beta}_1 \bar{X}_1 \tag{2.8}$$

and

$$\hat{\alpha}_2 = \bar{Y_2} - \hat{\beta}_2 \bar{X}_2. \tag{2.9}$$

2.2 The Full Model

In the simple linear regression change point case, the full model is the normal error regression model with two lines:

$$Y_i = \begin{cases} \alpha_1 + \beta_1 X_i + \varepsilon_i, & X_i < c \\ \alpha_2 + \beta_2 X_i + \varepsilon_i, & X_i > c \end{cases} \tag{2.10}$$

Figure 2.1 shows a typical graph of two regression llines, which demonstrates the fundamental concept underlying the full model.

We fit this full model by the method of least squares and obtain the error sum of squares. The error sum of squares is the sum of the squared deviations of each observation Y_i around its estimated expected value. In this context, we shall denote

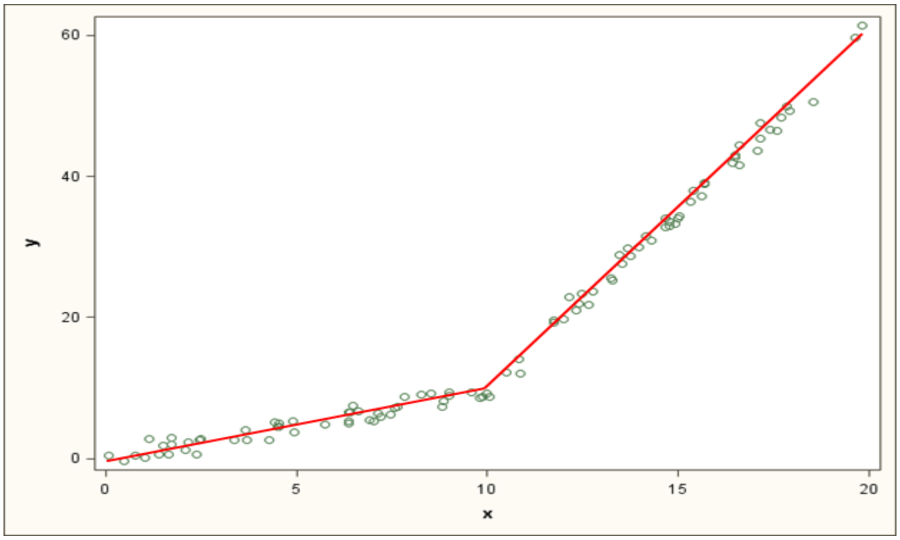

Figure 2.1: An illustration for the full model.

this sum of squares by $SSE(F)$ to indicate this it is the error sum of squares for the full model. Here, we have:

$$\begin{aligned} SSE_{Full} &= \sum_{X_i<c}(Y_i - \hat{\alpha}_1 - \hat{\beta}_1 X_i)^2 + \sum_{X_i>c}(Y_i - \hat{\alpha}_2 - \hat{\beta}_2 X_i)^2 \\ SSE_{Full} &= SSE_1 + SSE_2. \end{aligned} \tag{2.11}$$

Thus, for the full model, the error sum of squares is $(SSE_1 + SSE_2)$, which measures the variability of the Y_i observations around the fitted regression model. The sum of squares SSE_1 has $(n_1 - 2)$ associated degrees of freedom. The sum of squares SSE_2 has $(n_2 - 2)$ associated degrees of freedom. Two degrees of freedom are lost because both α_i and β_i had to be estimated in obtaining the estimated means $\hat{Y}_i$.

2.3 The Reduced Model

Next, we consider H_0. In this situation, we have:

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 \tag{2.12}$$

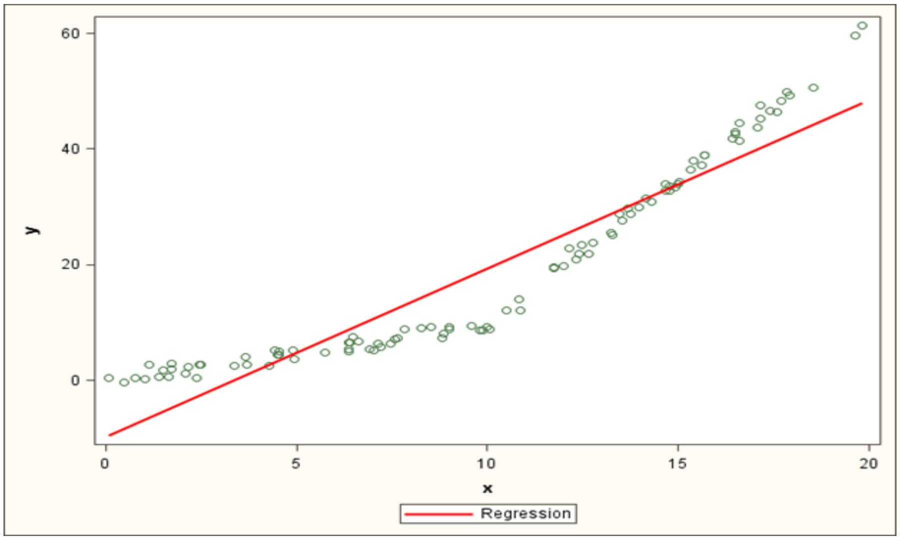

Figure 2.2: An illustration for the reduced model.

$$H_a : \alpha_1 \neq \alpha_2 \text{ or } \beta_1 \neq \beta_2. \tag{2.13}$$

The model when H_0 holds is called the reduced model. When $\alpha_1 = \alpha_2, \beta_1 = \beta_2$, the model is reduced to:

$$Y_i = \alpha + \beta X_i + \varepsilon_i. \tag{2.14}$$

Figure 2.2 shows a typical graph of one regression lline, which demonstrates the fundamental concept underlying the reduced model.

We fit this reduced model, by the method of least squares, obtain the error sum of squares for this reduced model, denoted by $SSE(R)$. Mathematically,

$$SSE(R) = \sum_{x=1}^{n}[Y_i - (\alpha + \beta x_i)]^2 = \sum_{x=1}^{n}[Y_i - \hat{Y}_i]^2 = SSE \tag{2.15}$$

2.4 An introduction to F_{max} statistics

Equation 2.1 is considered as the simplest two-phase regression model. There are two types of change points associated with this model, namely step change point (sudden jump in intercept) and trend change point (sudden jump in slope). We illustrate these two concepts by Figure 2.3 and Figure 2.4.

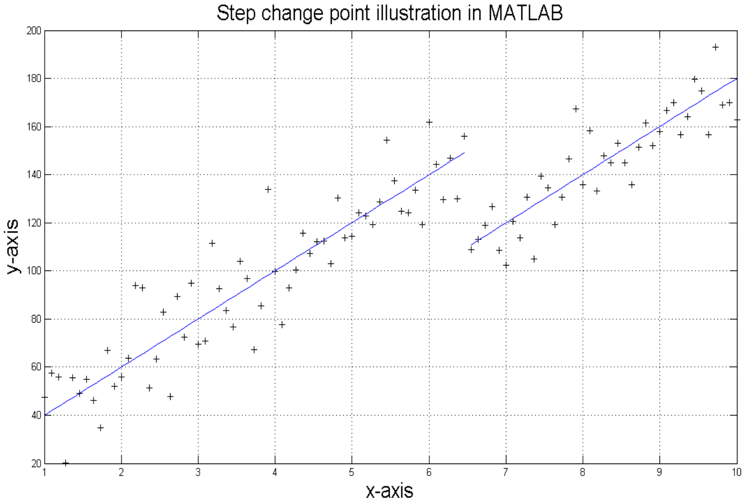

Figure 2.3: Step change in constant term α_i: $\sigma = 12; \alpha_1 = 20; \alpha_2 = -20; \beta_1 = \beta_2 = 20$

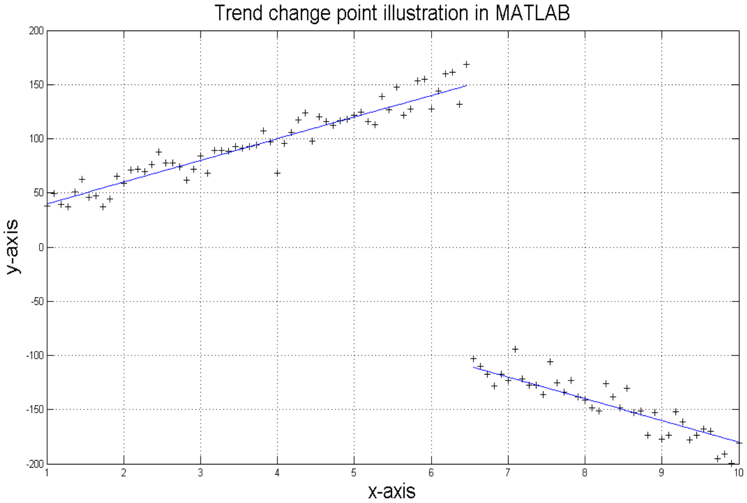

Figure 2.4: Trend change in slope term β_i: $\sigma = 12; \beta_1 = 20; \beta_2 = -20$

The null hypothesis H_0 is that there is no change point, and the alternative hypothesis is that there is an undocumented change point. These expressions are mathematically written as,

$$H_0 : \alpha_1 = \alpha_2, \beta_1 = \beta_2 \tag{2.16}$$

and

$$H_a : \alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2. \tag{2.17}$$

The aforementioned model allows both intercept ($\alpha_1 \neq \alpha_2$) and slope ($\beta_1 \neq \beta_2$) types of change points. Therefore, the formal definition of change point c if either intercept $\alpha_1 \neq \alpha_2$ or slope $\beta_1 \neq \beta_2$.

A comprehensive change point detection method will check for both intercept and slope changes.

The null hypothesis model of one regression line can be stated as $\hat{\alpha}_1 - \hat{\alpha}_2$ and $\hat{\beta}_1 - \hat{\beta}_2$ should be statistically close to zero for each $X_{i\in\{1,\cdots,n\}}$, and a general linear test statistic

$$F = \frac{(SSE_{Red} - SSE_{Full})/2}{SSE_{Full}/(n-4)} \tag{2.18}$$

should be small when there is no change point. So the goal is to figure out when F achieves its maximum value by testing each X_i as the change point. Based on the definition of change point c, we can see that

$$c = \underset{x\in\{1,2,\cdots,n\}}{\arg\max}\ F \tag{2.19}$$

A natural question is how the two error sums of squares $SSE(F)$ and $SSE(R)$ are compared relative to each other. In fact, $SSE(F)$ is never greater than $SSE(R)$. The main reason is that if more parameters are included in the model, it helps the process of fitting the data and reduces the standard deviation around the fitted regression function. When $SSE(F)$ is not much less than $SSE(R)$, the full model does not yield much more explanation of the Y_i than the reduced model does, in which case the data suggests that the reduced model is adequate (i.e. H_0 holds). On the other hand, a large difference suggests that H_a holds because the additional parameters in the model do help to reduce substantially the variation of the observations Y_i around the fitted regression function.

CHAPTER 3

PARAMETRIC STUDY: NUMERICAL SIMULATION AND RESULTS ANALYSIS

3.1 Design of numerical simulation and its purpose

As mentioned in the previous chapters, we examine the model

$$\begin{cases} \hat{Y}_1 = \hat{\alpha}_1 + \hat{\beta}_1 X_i + \varepsilon_i, & X_i < c, \\ \hat{Y}_2 = \hat{\alpha}_2 + \hat{\beta}_2 X_i + \varepsilon_i, & X_i > c \end{cases} \tag{3.1}$$

and

$$\varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{3.2}$$

and its associated statistic for the ability to detect a change point. We consider the following terms to be varying parameters:

1. β_2: while keeping $\beta_1 = 1$ as a constant, the second slope value β_2 is considered to be one of these values

   $$\beta_2 \in \{1.5, 2, 3, 5\}$$

   We vary β_2 to study the sensitivity of the proposed method to the amount of change in the slope. The numerical experiment results are analyzed in Section 4.1.

2. n_2: parameter n_2 gives us a general idea about the efficiency of the proposed method. The second sample size (number of points after the change) was allowed to take these following values:

   $$n_2 \in \{10, 20, 30, 40, 50\}$$

   while n_1 is held constant at 50. The numerical results are collected and analyzed in Section 4.2.

3. (n_1, n_2): while keeping the ratio of $\dfrac{n_1}{n_2} = \dfrac{5}{2}$ as a constant, increased the overall sample size to examine whether detection of the change point was influenced by the number of points before the change. The following values were used for (n_1, n_2):

   $$(n_1, n_2) \in \{(50, 20), (100, 40), (150, 60), (200, 80)\}.$$

   Its results are collected and tabulated in Section 4.3.

4. σ is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$: we also vary σ to investigate the sensitivity of the proposed method with respect to variability in the data around the true model.

   $$\sigma \in \{1, 2, 3, 4\}.$$

   The numerical results of σ change are listed in Section 4.4.

Due to the considerable number of different permutations in the nature of the parametric study, a baseline for all the numerical simulations is proposed. In fact, we only consider changing one value at a time; the results are to be compared with the baseline and other possibilities with respect to change in only one parameter. In this work, the standard baseline for simulation is:

$$\begin{cases} n_1 = 50 \\ n_2 = 20 \\ \beta_1 = 1 \\ \beta_2 = 2 \\ \sigma = 1 \end{cases} \tag{3.3}$$

3.2 Simulation Procedure Description

It is noted that this F_{max} test statistic no longer has an exact F distribution. The simulation is executed as the following procedures:

Procedures A: Generating data

Step-1: Generate n_1 simulated data with a regression line

$$\hat{Y}_1 = \hat{\alpha}_1 + \hat{\beta}_1 X_i + \varepsilon_i \tag{3.4}$$

for the first interval

$$0 \leq X_i \leq 10 \tag{3.5}$$

with $\hat{\alpha}_1 = 0$.

Step-2: Generate n_2 simulated data with a regression line

$$\hat{Y}_2 = \hat{\alpha_2} + \hat{\beta}_2 X_i + \varepsilon_i \tag{3.6}$$

for the second interval

$$10 \leq X_i \leq 10\frac{n_2}{n_1}. \tag{3.7}$$

Step-3: Order the simulated data

Procedures B: Analyzing data

Step-1: Fit the reduced model for all the data points $X_1, X_2, \cdots, X_c, \cdots, X_n$ with one regression line.

Step-2: Compute SSE for the reduced model.

$$SSE_{Red} = \sum_{X_i=1}^{n} (Y_i - \hat{\alpha}_{Red} - \hat{\beta}_{Red} X_i)^2. \tag{3.8}$$

Note that the degrees of freedom of the reduced model are $(n-2)$.

Step-3: Fit the full model for the first interval $X_1, X_2, \cdots, X_c$ and the second interval $X_{c+1}, \cdots, X_n$ with two different regression lines.

Step-4: Compute SSE_1 and SSE_2 for the full model as

$$\begin{aligned} SSE_{Full} &= \sum_{X_i<c} (Y_i - \hat{\alpha}_1 - \hat{\beta}_1 X_i)^2 + \sum_{X_i>c} (Y_i - \hat{\alpha}_2 - \hat{\beta}_2 X_i)^2 \\ SSE_{Full} &= SSE_1 + SSE_2. \end{aligned} \tag{3.9}$$

Note that the degrees of freedom of the full model is

$$n_1 - 2 + n_2 - 2 = n - 4.$$

Step-5: Put Step-3 and Step-4 inside a *for* loop, to include all possible two line models.

Step-6: Compute the F statistic for each of substep as

$$F = \frac{(SSE_{Red} - SSE_{Full})/2}{SSE_{Full}/(n-4)}. \tag{3.10}$$

Step-7: Find F_{max} and the index in X_i that gives F_{max}.

Procedures C: Sample the statistical process by repeating procedures A and B described above 1000 times to get a stable empirical distribution of F_{max}.

CHAPTER 4

SIMULATION RESULTS ANALYSIS

4.1 Slope change

Constant parameters:

$$\begin{cases} n_1 = 50 \\ n_2 = 20 \\ \beta_1 = 1 \\ \sigma = 1 \end{cases} \tag{4.1}$$

Varying parameters:

$$\beta_2 \in \{1.5, 2, 3, 5\} \tag{4.2}$$

We generate the first portion of a sample dataset with sample size $n_1 = 50$, and slope $\beta_1 = 1$, the regression line of these data is:

$$\hat{Y_1} = X_i + \varepsilon_i, \quad 0 \leq X_i \leq 10. \tag{4.3}$$

We then generate the second portion with sample size $n_2 = 20$, with the slope β_2 separately, i.e.

$$\hat{Y_2} = \alpha_2 + \beta_2 X_i + \varepsilon_i, \quad 10 < X_i \leq 14, \tag{4.4}$$

where $\beta_2 \in \{1.5, 2, 3, 5\}$.

The numerical results are tabulated in Table 4.1.

In order to recognize the pattern, Minitab is employed to graphically plot the tabulated results, as β_2 varies.

When β_2 was changed to 1.5, the method failed 8 times (i.e., The change point index c was 1 or 2) to detect a change point in 1000 trials. As the difference of two slopes $|\beta_1 - \beta_2|$ is larger, the mean of change point index C_{mean} is moving closer to 50

n_1	n_2	Slope1	Slope2	Std dev	C_mean	C_std dev	C_min	C_max
50	20	1	1.5	1	43.54	15.578	1	67
50	20	1	2	1	47.381	8.302	4	67
50	20	1	3	1	48.477	3.963	34	60
50	20	1	5	1	48.835	2.237	42	57

Table 4.1: Slope change - Tabulated simulations inputs and outputs

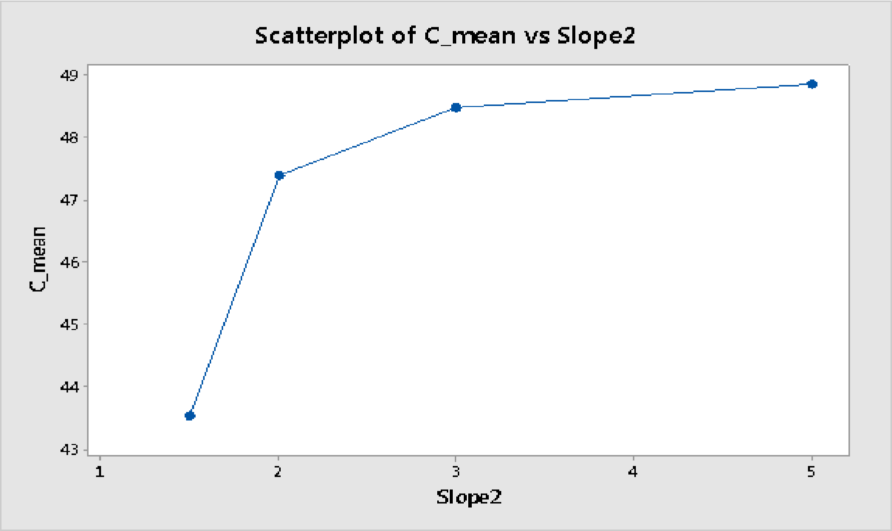

Figure 4.1: $\hat{\beta}_2$ vs. C_mean in Slope change simulation

which is the number of points before the change; also, the standard deviation of the position of point of change $C_{std\ dev}$ is smaller. It means the detection of change point is more precise. Concluding, the slope change will affect the accuracy of detection: the bigger the difference of the slopes is, the more accurate and precise it can get.

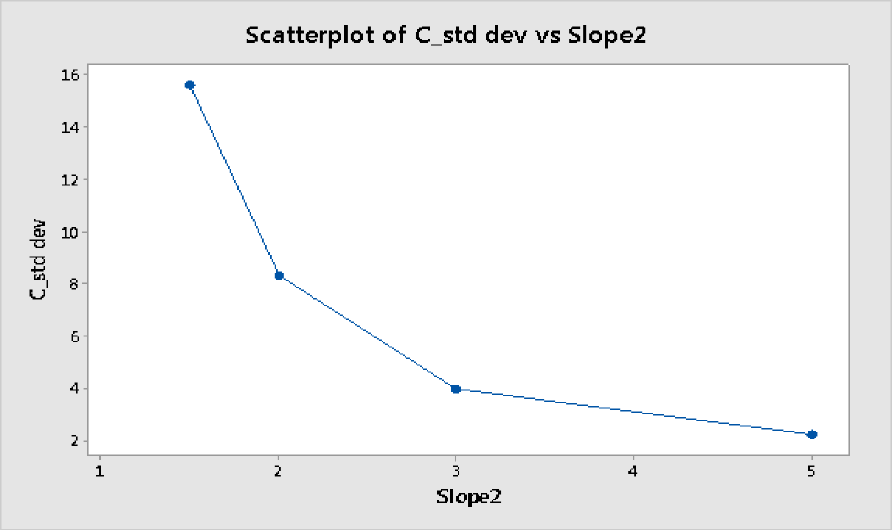

Figure 4.2: $\hat{\beta}_2$ vs. C_std dev in Slope change simulation

4.2 The proportion of sample size change

Constant parameters:

$$\begin{cases} n_1 = 50 \\ \beta_1 = 1 \\ \beta_2 = 2 \\ \sigma = 1 \end{cases} \tag{4.5}$$

Varying parameters:

$$n_2 \in \{10, 20, 30, 40, 50\}. \tag{4.6}$$

Here, we want to examine the dependency of the detection on the proportion of total sample size n_2. To do that, we generate the first portion of a dataset with slope $\beta_1 = 1$ with sample size $n_1 = 50$, the regression line of these data is:

$$\hat{Y}_1 = X_i + \varepsilon_i, \quad 0 \leq X_i \leq 10. \tag{4.7}$$

We then generate the second portion with slope of $\beta_2 = 2$ with sample size $n_2 \in$

$\{10, 20, 30, 40, 50\}$ separately. The regression lines of these data are

$$\hat{Y}_2 = -10 + 2X_i + \varepsilon_i, \quad 10 < X_i \leq 10 + 10\frac{n_2}{50} \tag{4.8}$$

where $n_2 \in \{10, 20, 30, 40, 50\}$.

The numerical results are tabulated in the following table:

n_1	n_2	Slope1	Slope2	Std dev	C_mean	C_std dev	C_min	C_max
50	10	1	2	1	39.774	14.644	1	57
50	20	1	2	1	47.381	8.302	4	67
50	30	1	2	1	48.605	6.993	4	66
50	40	1	2	1	48.616	6.236	21	72
50	50	1	2	1	49.168	6.137	34	67

Table 4.2: Proportion sample size change - Tabulated simulations inputs and outputs

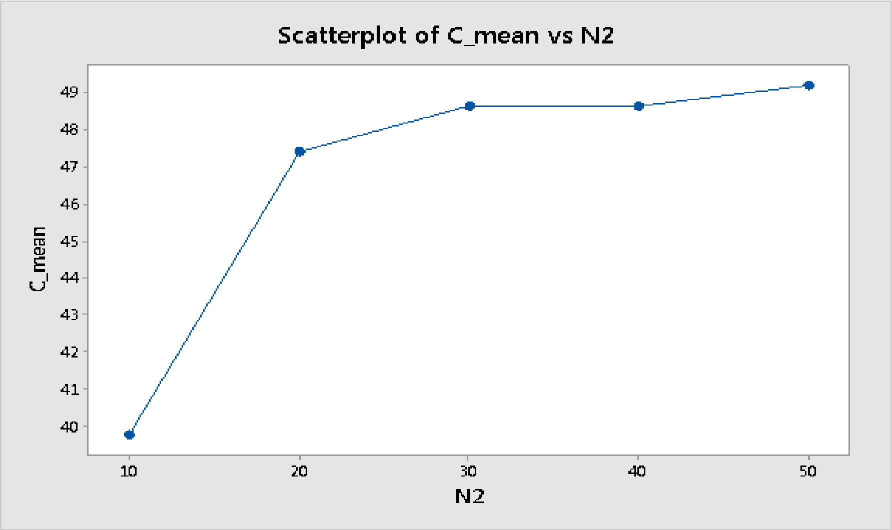

Figure 4.3: n_2 vs. C_mean in Proportion sample size change simulation

When n_2 was changed to 10, this F test failed 18 times (i.e., The change point index was 1 or 2) to detect a change point in 1000 trials. We observe that when the

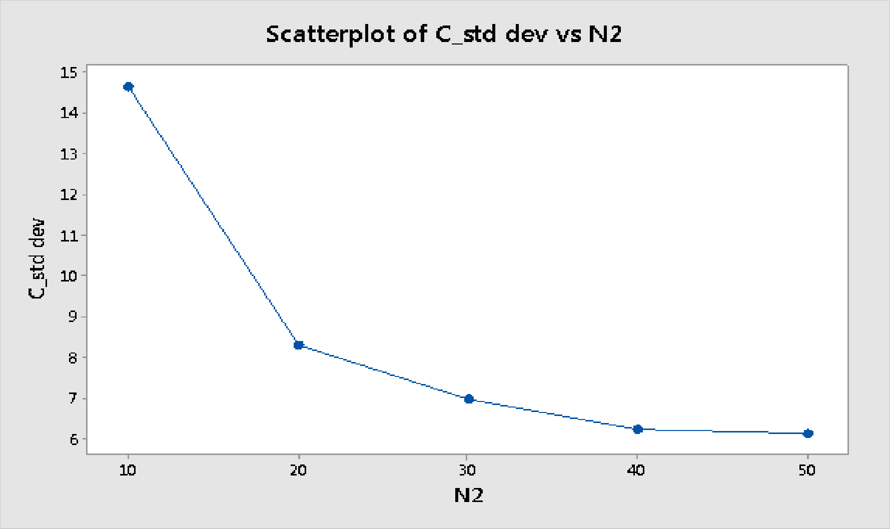

Figure 4.4: n_2 vs. C_std dev in Proportion sample size change simulation

proportion of the sample size n_2 is larger, the mean of change point position is closer to $n_1 = 50$; the standard deviation of change point gets smaller almost exponentially. It implies that the detection of change point is more accurate; so in this F test, the proportion of the total sample size after the change dramatically affects the accuracy of detection.

4.3 The sample size change

Constant parameters:

$$\begin{cases} \beta_1 = 1 \\ \beta_2 = 2 \\ \sigma = 1 \end{cases} \tag{4.9}$$

Varying parameters:

$$(n_1, n_2) \in \{(50, 20), (100, 40), (150, 60), (200, 80)\}. \tag{4.10}$$

Thirdly, we investigate the dependency of the detection on the sample size change, while keeping the sample ratio $\frac{n_1}{n_2} = \frac{5}{2}$. We generate the first portion of a sample dataset with the slope of $\beta_1 = 1$ with sample size of $n_1 \in \{50, 100, 150, 200\}$ separately. The regression line of these data is:

$$\hat{Y_1} = X_i + \varepsilon_i, \quad 0 \leq X_i \leq 10 \tag{4.11}$$

We then generate the second portion with the slope of $\beta_2 = 2$ with the sample size n_2 of $n_2 \in \{20, 40, 60, 80\}$ respectively. The regression line is

$$\hat{Y_2} = -10 + 2X_i + \varepsilon_i, \quad 10 < X_i \leq 14, \tag{4.12}$$

The simulation results regarding the position of the change point is tabulated below:

n_1	n_2	Slope1	Slope2	Std dev	C_mean	C_std dev	C_min	C_max	Bias
50	20	1	2	1	47.381	8.302	4	67	2.619
100	40	1	2	1	97.29	10.848	57	130	2.71
150	60	1	2	1	147.578	13.914	108	178	2.422
200	80	1	2	1	197.51	16.598	162	239	2.49

Table 4.3: Sample size change - Tabulated simulations inputs and outputs

We can see that as the sample size gets larger, the bias of change point fluctuates from around 2.4 to 2.7 and the standard deviation of the change point estimate increases almost linearly. It means the detection of change point is almost the same, but the variance of the change point estimate becomes larger and larger.

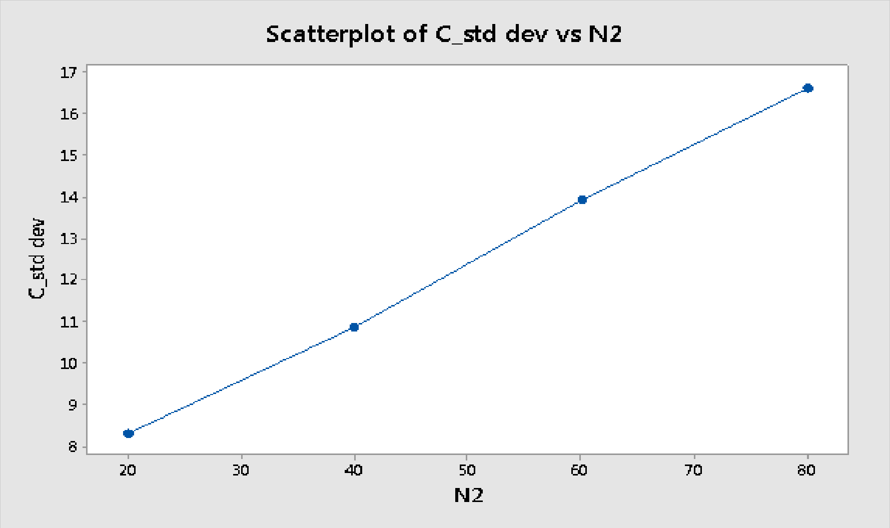

Figure 4.5: n_2 vs. C_std dev in Sample size change simulation

4.4 The Standard Deviation change

Constant parameters:

$$\begin{cases} n_1 = 50 \\ n_2 = 20 \\ \beta_1 = 1 \\ \beta_2 = 2 \end{cases} \tag{4.13}$$

Varying parameters:

$$\sigma \in \{1, 2, 3, 4\}. \tag{4.14}$$

Fourthly, we investigate whether the standard deviation of the random noise error will affect the accuracy of detection. We generate the first portion of a sample dataset with slope of $\beta_1 = 1$ with the sample size of $n_1 = 50$, the regression line is:

$$\hat{Y_1} = X_i + \varepsilon_i, \quad 0 \leq X_i \leq 10. \tag{4.15}$$

We then generate the second portion with the slope of $\beta_2 = 2$ with sample size

$n_2 = 20$. The regression line is:

$$\hat{Y}_2 = -10 + 2X_i + \varepsilon_i, \quad 10 < X_i \leq 14. \tag{4.16}$$

The standard deviation for both regression lines are $\sigma \in \{1, 2, 3, 4\}$ respectively. We tabulated the results in the following table:

n_1	n_2	Slope1	Slope2	Std dev	C_mean	C_std dev	C_min	C_max
50	20	1	2	1	47.381	8.302	4	67
50	20	1	2	2	42.738	16.111	1	67
50	20	1	2	3	38.722	19.735	1	67
50	20	1	2	4	38.16	20.9237	1	67

Table 4.4: Standard deviation change - Tabulated simulations inputs and outputs

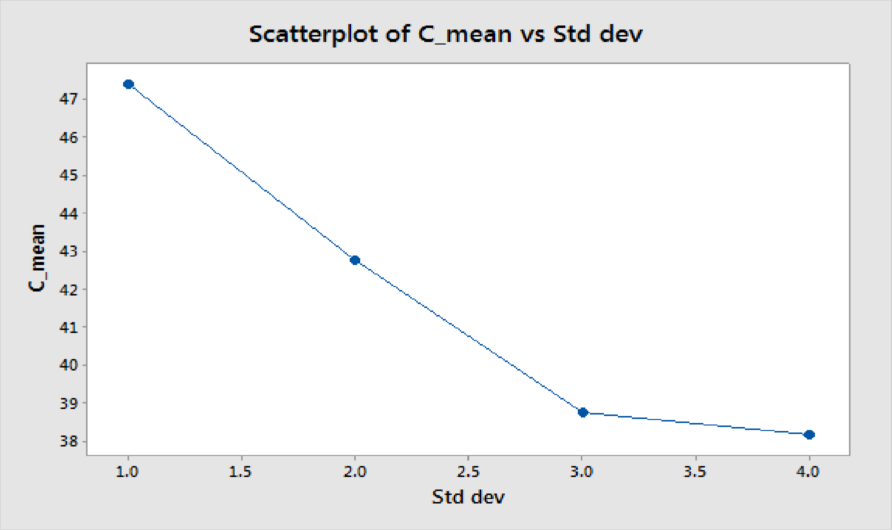

Figure 4.6: Std dev vs. C_mean in Standard deviation change simulation

As the standard deviation gets larger, the mean of change point distribution drops down dramatically and the method quickly loses its stability. The standard deviation of the change point distribution also increases at a relatively fast pace.

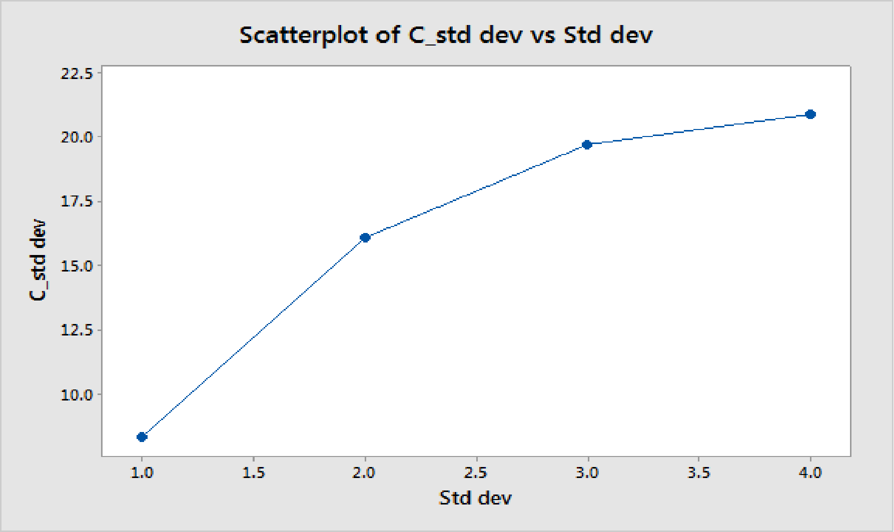

Figure 4.7: Std dev vs. C_std dev in Standard deviation change simulation

We see clearly that the efficiency of the proposed method strongly depends on the standard deviation of the noise error, thus the standard deviation strongly affects the accuracy of detection.

4.5 Sensitivity as a function of two variables

By previous results, increasing the variance of the error decreases the accuracy of the F-test; however, increasing the proportion sample size adds some degrees of accuracy toward estimating the change point. A natural question to ask is how accuracy is affected by varying two parameters at the same time.

Constant parameters:

$$\begin{cases} n_1 = 50 \\ \beta_1 = 1 \\ \beta_2 = 2 \end{cases} \tag{4.17}$$

Varying parameters:

$$n_2 \in \{10, 20, 30\}, \tag{4.18}$$

$$\sigma \in \{1, 2, 3, 4\}. \tag{4.19}$$

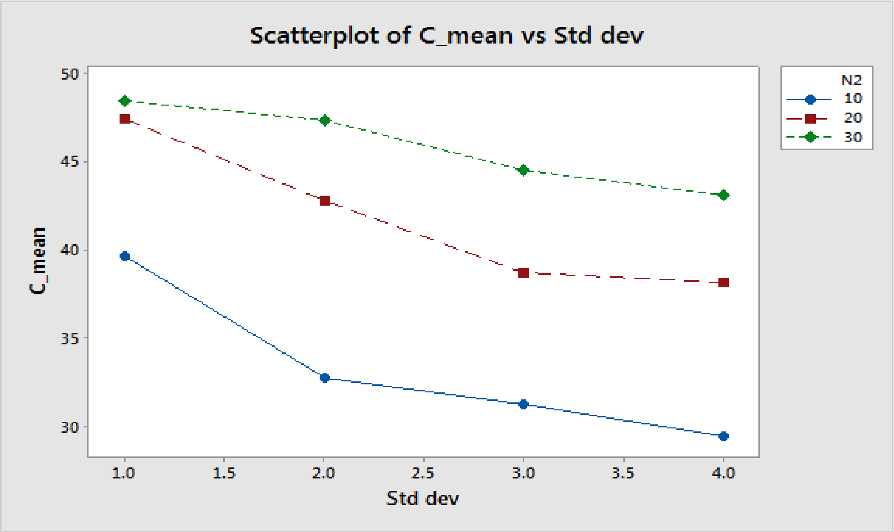

Figure 4.8: C_mean vs. Std dev in Proportion Sample Size and Variance change

We observe that (1) as the standard deviation grows larger, the more failed trials they have, and (2) as the proportion of the sample size n_2 gets larger, the mean

n_1	n_2	Slope1	Slope2	Std dev	C_mean	C_std dev	C_min	C_max	Fail times
50	10	1	2	1	39.668	14.721	1	57	18
50	10	1	2	2	32.752	18.173	1	57	52
50	10	1	2	3	31.255	18.804	1	57	69
50	10	1	2	4	29.397	19.163	1	57	74
50	20	1	2	1	47.381	8.302	4	67	0
50	20	1	2	2	42.738	16.111	1	67	8
50	20	1	2	3	38.722	19.735	1	67	26
50	20	1	2	4	38.16	20.924	1	67	50
50	30	1	2	1	48.415	6.647	26	69	0
50	30	1	2	2	47.33	13.057	1	76	2
50	30	1	2	3	44.463	17.812	1	77	8
50	30	1	2	4	43.059	21.417	1	77	28

Table 4.5: Standard deviation change and proportion of sample size change-Tabulated simulations inputs and outputs

of change point position is closer to n_1=50 for each value of variables, (3) if the standard deviation of the error σ grows larger, the mean of change point position is shifted further from the real change point n_1=50; it is true for every n_2 value, as in Figure 4.8.

On the other hand, the F-test quickly loses its stability as the variance grows larger, regardless of the proportion of sample size n_2, as illustrated in Figure 4.9. We explain the reasoning logic as follows:

- When the standard deviation of the error σ is relatively small, roughly around 1 and 2, the more information after the change point, the more accurate the test can be. This observation is made based on the unchanged order of n_2, when

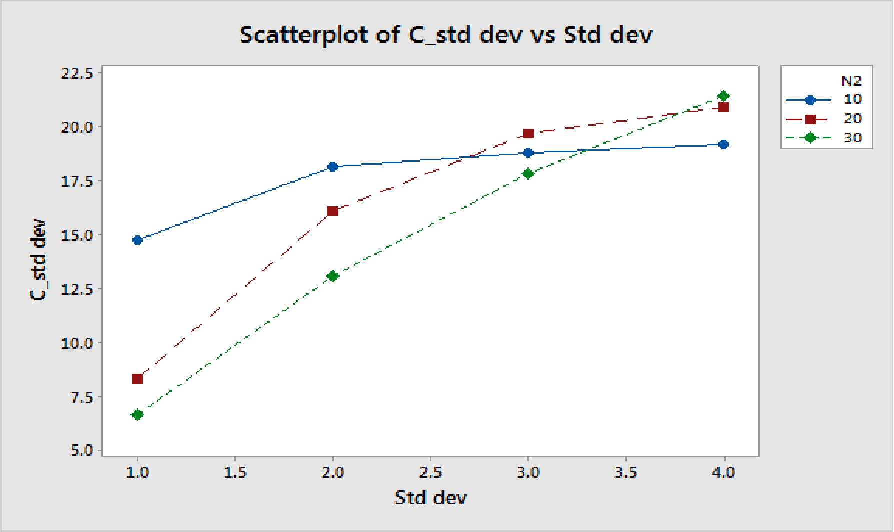

Figure 4.9: C_std dev vs. Std dev in Proportion Sample Size and Std dev change

$\sigma \in \{1, 2\}$.

- For relatively large variabilities, roughly around 3 or 4, increasing the number of data points after the change does not necessarily increase the accuracy of the test. This statement is most obvious when $\sigma = 4$.

  As the standard deviation gets larger, the mean of the change point distribution drops down dramaticallly and the method quickly loses its stability. On the other hand, the standard deviation of change point distribution increases at a relatively fast pace but then seems to stabilize. Concluding, the efficiency of the proposed method strongly depends on the standard deviation of the niose error; the standard deviation strongly affects the accuracy and precision of detection.

CHAPTER 5
SUMMARY

5.1 Discussion

The investigation into the sensitivity of the F-test reveals its strengths and weaknesses in detecting a change point in a single, change point model. The F-test is most applicable when change in slope is obvious, and is most ineffective when data are scattered with large deviation away from the mean. The change point detection is a fairly difficult technical problem with at least three main aspects: efficiency (detect the change point as fast as possible), accuracy (detect the change point at its true location) and precision (optimal deviation in change point distribution).

Besides this F-test, many other techniques can also be used to detect the change point position, for instance, CUSUM procedure [12], Shiryayev-Robert procedure [14], to list a few. Therefore, it might be necessary to do the sensitivity analysis to compare the robustness and reliability of these techniques. For an example of comparison of F-test with Empirical Likelihood Ratio (ELR), see [1]. Regardless of its power, the F-test method perhaps seems to be somewhat computationally heavy compared to other methods, and therefore, more likely geared toward post-processing of data which allows more time and requires more accuracy and precision, instead of online change point detection.

If detecting the change point as soon as it occurs is important, one automatically encounters the trade-off of decreasing detection delay and increasing the frequency of false alarms (type-I error). A popular method to resolve the problem is the maximum likelihood theory; for more information, see [11].

5.2 Conclusion

In this thesis, we studied the sensitivity of the F test with respect to the slope, the proportion of the sample size after the change, and the standard deviation. We conclude that the slope is an influential parameter to change point detection in the change point model: a larger difference in slopes yields a more accurate and precise detection (i.e. smaller deviation in the change point distribution). The numerical evidence also points out that the F-test loses most of its effectiveness when the data have much variability from increasing standard deviation error of the dataset. The proportion sample size change also has considerable effect on the F-test result: more information after the actual change point gives a more accurate and precise change point location. Regarding the sample size change, the bias of the change point location is relatively stable, but the variance of the change point location gets bigger as the sample size increases.

5.3 Future work

The work in this thesis points out to several directions, that might fit for future development:

(i) change point detection for intercept change

(ii) change point detection for the multiple change point model (both discontinuous and continuous)

(iii) sensitivity analysis of change point detection by F-test with different noise distribution, for examples, standardized Gamma distribution, exponential distribution and Student's t distribution with different degrees of freedom.

(iv) sensitivity analysis of change point detection by moving average control charts

(v) sensitivity analysis of change point detection by Page's CUSUM procedure [12]

(vi) sensitivity analysis of change point detection by the Shiryayev-Robert procedure [18]

(vii) compare the effectiveness of various change point detection techniques by using statistical simulations

REFERENCES

[1] Baragona, R., Battaglia, F. and Cucina,D. (2014) Maximum Empirical Likelihood Inference for Outliers in Autoregressive Time Series. *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, 17-20.

[2] Best, M., Neuhauser, D. (2006) Walter A Shewhart, 1924, and the Hawthorne factory. *Quality and Safety in Health Care,* **15** 142-143.

[3] Gill, D. (1970) Application of a statistical zonation method to reservoir evaluationand digitized log analysis. *American Association of Petroleum Geologists Bulletin,* **54**, 719-729.

[4] Hoffmann, John P (2010) Linear Regression Analysis: Applications and Assumptions. *Brigham Young University, Provo.*

[5] Hofrichter, J. (2007) Change Point Detection in Generalized Linear Models. *Ph.D. thesis.*

[6] Killick, R., Idris, A. E. (2011) Changepoint: an R package for changepoint analysis. *R package version 0.6, URL http://CRAN. R-project. org/package=changepoint*

[7] Kutner, M., Nachtsheim, C. J., Neter, J., and Li, W. (2005) *Applied Linear Statistical Models.* McGraw-Hill/Irwin, New York.

[8] Lorden, G. (1971) Procedure for reacting to a change in distribution. *Ann. Math. Statist.,* **42** 1897-1908.

[9] Lund, R., Reeves, J. (2002) Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate.,* **15**:17, 2547-2554

[10] Lund, R., Wang, X. L., Lu, Q. Q., Reeves, J., Gallagher, C. and Feng, Y. (2007) Changepoint detection in periodic and autocorrelated time series. *Journal of Climate.,* **20**:20, 5178–5190

[11] Mei, Y. (2003) Asymptotically optimal methods for sequential change-point detection. *Ph.D. thesis.*

[12] Page, E. S. (1955) A test for a change in a parameter occuring at an unknown point. *Biometrika,* **42**, 523-527.

[13] Roberts, S. W. (1966) A comparison of some control chart procedures. *Technometrics,* **8**, 411-430.

[14] Shiryayev, A. N. (1978) *Optimal Stopping Rules.* Springer-Verlag, New York.

[15] Snijders, A.M.; et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number *Nature Genetics,* **29** (3), 263-264.

[16] Sowa, Y.; Rowe, A. D., Leake, M. C., Yakushi, T., Homma, M., Ishijima, A., Berry, R. M.(2005) Direct observation of steps in rotation of the bacterial flagellar motor. *Nature,* **437** (7060), 916-919.

[17] Umbuahg, Scott E. (2010) Digital image processing and analysis : human and computer vision applications with CVIPtools (2nd ed.), CRC Press: Boca Raton, FL.

[18] Yakir, B. (1997) A note on optimal detection of a change indistribution. *Ann. Statist.,* **25**, 2117-2126.

Appendix A

SAS SOURCE CODES

```
options nonotes;
%let n1=50;
%let n2=20;
%let s1=1;
%let s2=2;
%let nsim=1000;

ods pdf file="C:\Users\js11718\Desktop\s5020121\s5020121.pdf";
%macro loop;
    %do a = 1 %to &nsim;

    /*get the first line y1=x1+error 0<x1<10*/
data Reg1(keep=x1 y1);
    call streaminit(0);
    do i=1 to &n1;
        x1= rand("Uniform");
        x1=10*x1;
        eps=rand("Normal",0,1);
        y1=&s1*x1+eps;
        output;
    end;
run;

    /*get the second line y1=2x1-(s2-s1)*10+error 10<x1<20*/
data Reg2(keep=x1 y1);
    call streaminit(0);
    do i=1 to &n2;
        x1= rand("Uniform");
        x1=x1*10*(&n2/&n1);
        x1=x1+10;
        eps=rand("Normal",0,1);
        y1=&s2*x1-(&s2-&s1)*10+eps;
        output;
    end;
run;
```

```
    /*combined two lines*/
data combined;
    set Reg1 Reg2;
run;

proc sort data=combined;
    by x1;
run;

/*get the sse for the reduced model*/
proc reg data=combined noprint
    outest=result(drop=_model_ _type_ _depvar_ _rmse_ y1 _in_ _p_ _edf_ _rsq_);
    model y1=x1/sse;
run;
quit;

/*get the sse1 for the front of full model*/
%macro test;
data _null_;
    if 0 then set combined nobs=nobs;
    call symput ('n',nobs);
run;
    %do i=2 %to &n-2;
data want&i;
    set combined (obs=&i);
run;
proc reg data=want&i noprint
    outest=result&i(drop=_model_ _type_ _depvar_ _rmse_ y1 _in_ _p_ _edf_ _rsq_);
    model y1=x1/sse;
    ods exclude Nobs;
    proc append base=fulla data=result&i;
    proc append base=reduced data=result;
run;
quit;
%end;
%mend;
%test*/
run;
```

```
proc sort data=combined;
    by descending x1;
run;

/*get the sse2 for the back of full model*/
%macro test;
data _null_;
    if 0 then set combined nobs=nobs;
    call symput ('n',nobs);
run;
    %do i=2 %to &n-2;
data want&i;
    set combined (obs=&i);
run;
proc reg data=want&i noprint
    outest=result&i(drop=_model_ _type_ _depvar_ _rmse_ y1 _in_ _p_ _edf_ _rsq_);
    model y1=x1/sse;
    ods exclude Nobs;
    proc append base=fullb data=result&i;
run;
quit;
%end;
%mend;
%test*/
run;

/*calculate F values*/
data fulla;
    set fulla;
    indx = _n_;
run;
proc sort data=fulla;
    by descending indx;
run;
data r(drop=intercept x1 indx) ;
    set fulla(rename=(_sse_=sse1));
    set fullb(rename=(_sse_=sse2));
    set reduced(rename=(_sse_=sse _mse_=mse));
    E=sse1+sse2;
    F=((sse-E)*(&n1+&n2-4))/(E*2);
```

```
run;

/*get the maximum of F*/
data r;
   set r;
   indx = _n_;
run;

proc sql;
   create table Max AS
   select indx,F as F
   from r
   having F= max(F);
quit;

proc datasets lib=work nolist;
   delete fulla fullb reduced;
run;

proc append base=max1 data=max;
run;

%end;
%mend;
%loop;

proc means data=max1;
run;
proc print data=max1;
run;
libname s5020121 "C:\Users\js11718\Desktop\s5020121";
data s5020121.max1;
   set Max1;
run;
data s5020121.max;
   set Max;
run;
data s5020121.R;
   set R;
run;
```

```
data s5020121.combined;

   set combined;

run;

ods pdf close;

run;
```