A Comparison of Linear Regression Models for Heteroscedastic and Non-Normal Data

Raksmey Thinh

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Mathematics and Statistics

Prince of Songkla University

2018

A Comparison of Linear Regression Models for Heteroscedastic and Non-Normal Data

Raksmey Thinh

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Mathematics and Statistics

Prince of Songkla University

2018

Thesis Title	A Comparison of Linear Regression Models for Heteroscedastic and Non-Normal Data
Author	Mr. Raksmey Thinh
Major Program	Mathematics and Statistics

Major Advisor

.....

(Dr. Klairung Samart)

Co-advisor

.....

(Asst. Prof. Dr. Naratip Jansakul)

Examining Committee:

.....Chairperson

(Assoc. Prof. Dr. Wararit Panichkitkosolkul)

.....Committee

(Dr. Klairung Samart)

.....Committee

(Asst. Prof. Dr. Naratip Jansakul)

.....Committee

(Dr. Thammarat Panityakul)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Master of Science Degree in Mathematics and Statistics.

.....

(Prof. Dr. Damrongsak Faroongsarng)

Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

..... Signature

(Dr. Klairung Samart)

Major Advisor

..... Signature

(Asst. Prof. Dr. Naratip Jansakul)

Co-advisor

..... Signature

(Mr. Raksmey Thinh)

Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

..... Signature

(Mr. Raksmey Thinh)

Candidate

Thesis Title	A Comparison of Linear Regression Models for Heteroscedastic and Non-Normal Data
Author	Mr. Raksmey Thinh
Major Program	Mathematics and Statistics
Academic Year	2018

Abstract

In common practices, heteroscedasticity and non-normality are frequently encountered when fitting linear regression models. Several methods have been proposed to handle these problems. In this research, we compared four different estimation methods: ordinary least squares (OLS), transform both sides (TBS), power of the mean function (POM) and exponential variance function (VEXP), dealing with three different forms of the non-constant variances under four symmetric distributions. In order to study the performance of the four methods in estimating the studied model parameters, a simulation study with various sample sizes of 20, 50, 100, and 200 was conducted. To determine the models with the best fit, relative bias, mean squared error (MSE) and coverage probability of the nominal 95% confidence interval were applied. The simulation results and application to real life data suggest that each estimation method performed differently on different variance structures and different distributions whereas the sample size did not give much effect on each estimation method except in the case of extreme heteroscedasticity. In overall, the TBS method performed best in terms of smallest bias and MSE, especially under extreme heteroscedasticity. On the other hand, the OLS method was very accurate in maintaining the nominal coverage probabilities although it had relatively poor performance in terms of bias.

Acknowledgements

I would like to acknowledge and extend my deepest gratitude to my advisor, Dr. Klairung Samart, and co-advisor, Asst. Prof. Dr. Naratip Jansakul who always open their office's door whenever I ran into a trouble spot or had a question about my research. They consistently allowed this research to be my own work, but steered me in the right direction whenever they thought I need it to accomplish this thesis.

I would like to express my appreciation to the chairperson and member of my committee, Assoc. Prof. Dr. Wararit Panichkitkosolkul and Dr. Thammarat Panityakul for their helpful comments, guidance and advice concerning this thesis.

How could I forget all my teachers and staff of the Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University for sharing their knowledge and support so that I can obtain this Master degree.

I am grateful to the Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries Project Office of Higher Education Commission, contract no. TEH-AC 099/2016 for financial support throughout my Master degree.

Finally, I would like to express my profound gratitude to my beloved parents, siblings, and to my friends for their love and support throughout my life. Thank you for giving me strength and understanding in many moments of crisis. Your love and friendship make my life fulfill wonderful experiences.

Raksmey Thinh

Contents

List of Tables

List of Figures

List of Abbreviations

Bias	Relative bias
i.i.d	Independently identically distributed
MSE	Mean square error
OLS	Ordinary least squares
POM	Power of the mean function
SE	Standard error
TBS	Transform both sides
VEXP	Exponential function

Chapter 1

Introduction

1.1 Background and Motivation

In statistical analysis, linear regression is applied for modeling the relationship between the response variable and one or more predictor variables. This model is implemented by assuming that the response variable is a linear function of the predictor variables and an unknown term or error. In common practice, the error is strongly assumed to be an independent random variable having normal distribution with the mean of zero and constant variance. The measure of predictor variables can be either qualitative or quantitative. Thus the response variable is a random variable having normal distribution with constant variance and its mean is the linear combination of predictor variables. If there is only one predictor variable, it is called **simple linear regression**, while the regression with more than one predictors is known as **multiple linear regression**. The linear regression model is commonly used in many applied research fields. For example, in Finland, the linear regression model was used to predict berry yields, cowberry yields, potential yields in 'Golden Delicious' and 'Royal Gala' apple before bloom, potential yields in peach before bloom, potential yields of pear cultivar 'Blanquilla' and 'Conference' before bloom (Ihalainen & Pukkala, 2001; Jiménez & Díaz, 2004, 2003b,a), respectively. As well as, in eastern Canada, linear regression model of low-bush blueberry was conducted by using cli-

mate to be an explanatory variables to predict the yields (Hall et al., 1982). In Chile, the model for high bush blueberry was developed to estimate the potential yield by using linear regression which included number of flower buds and climate variables (Salvo et al., 2012; Ávila et al., 2013). Moreover, to observe the relationship between cardiovascular risk factors and atherosclerotic disease, Nicholls et al. (2006) used linear regression of rank-transform for univariate predictors and used two selection procedures to select the most significant variables and removed least significant variables for multiple linear regression model. The linear regression was also used to investigate the predictor of plaque burden; and used to establish explanatory factors for Decayed Missing Filled Teeth (DMFT) score and unstimulated salivary flow rate (Gebhard et al., 2018; Äyräväinen et al., 2018), respectively.

In reality, however, the assumptions of homogeneity and normality are always violated. For example, when the numbers of predictor increase, the prediction error seem to increase as well which indicates that the variance in the prediction error is non-constant and this effect is known as heteroscedasticity (Ávila et al., 2013). Ketelaere et al. (2006) claimed that heteroscedasticity often occurred in many storage experiments due to the earlier harvest, one should consider the amount of variance in quality of the products. For the data with the heteroscedastic response, it has been noted that the point estimates of the model parameters are inefficient and may lead to inaccurate inference for the model parameters (Sidik & Jonkman, 2016). Covering to these problems, the previous studies have proposed several methods which are possible to deal up with this situation. There is one common method claimed by Giltinan et al. (1986); Carroll & Ruppert (1988); Kutner (2005); Ávila et al. (2013) namely data transformation using logarithm transform both sides (TBS). This method had been commonly used in modeling blueberry fruits and flower buds, red raspberry growth and yield, Pharmacokenetics data, and fish length-weight relationship (Ávila et al., 2013; Clark et al., 2012; Giltinan & Ruppert, 1989; Le Cren, 1951), respectively. Moreover, it was considered as the best transformation among others attempted transforms, claimed in berry yield prediction, bilberry and cow-

berry yield prediction, $\log(y_i + 1)$ was applied to avoid logarithm of zero in both Ihalainen & Pukkala (2001) and Ihalainen et al. (2003). The transformation is assuredly reasonable in many applied situations. Nevertheless, one might be uncertain of the transformation because it can be difficult to make inference about the original scale (Giltinan et al., 1986). Meanwhile, Giltinan & Ruppert (1989); Ávila et al. (2013); Hao et al. (2015); Sidik & Jonkman (2016) also used two kinds of so-called residual variance functions: power of the mean function (POM) and exponential function (VEXP) (see also Pinheiro & Bates (2000)) to reduce heteroscedasticity.

However, based on the previous studies, we observed that the results were commonly questioned on occurring of non-constant variance. That is they only focused on heteroscedasticity where normality assumption was assumed to verify (Giltinan et al., 1986; Beal & Sheiner, 1988; Giltinan & Ruppert, 1989; Ávila et al., 2013; Sidik & Jonkman, 2016). This leads to the aim of this thesis, i.e., to explore the difference of linear regression parameters using different estimation methods under heteroscedasticity and non-normality of data.

1.2 Objectives

To compare and discuss the performance of four different estimation methods: OLS, TBS, POM, and VEXP methods in simulation study and real life data.

1.3 Scope and Methodology

This research focuses on comparing and discussing four different estimation methods namely OLS, TBS, POM, and VEXP under three different forms of heteroscedastic response and four symmetric distributions through the simulation study with the sample sizes of 20, 50, 100, and 200.

The application to real life dataset will also be conducted to compare the performance of the four estimation methods.

1.4 Advantages

1. Fulfill the knowledge of linear regression model and their properties;

2. Understand and be able to apply a linear regression model to real life situations;

3. Be able to find out a better way in fitting linear regression under the problems of heteroscedastic and non-normal data.

1.5 Thesis Overview

We have described the motivation, objectives, scope, and advantages of the study. We now organize and summarize the remaining parts of this thesis where the exploration of reviews, the assessment of results and discussion, and the conclusion are as follows.

In Chapter 2, we explore the basic concepts of linear regression model including model assumptions, estimation methods, models comparison and confidence intervals. We will deeply review on four estimation methods namely OLS, TBS, POM, and VEXP followed by three criteria namely: bias, mean square error (MSE), and coverage probability which are used to evaluate and assess each estimation method. This thesis will mainly concentrate on the performance of the four estimation methods with three studied different heteroscedastic patterns under four studied symmetric distributions.

Chapter 3 consists of two parts. In the first, we explain the procedure of simulation study to assess and compare the performance of the four estimation methods. It will bring up with the idea of getting the independent variable X and dependent variable Y and how to incorporate random error with the studied four distributions and three studied non-constant variances. Second, we show how to request and retrieve a real life data. The detail about the data set will be described, the assumption test to confirm homogeneity and normality of data will also performed.

Chapter 4 presents and explains the results and discussion of each estimation method in the simulation study and real life data set.

Finally, a summary of the thesis and some suggestions for further research will be presented in Chapter 5.

Chapter 2

Reviews of Literature

2.1 Basic concepts of regression analysis

Statistical modeling is an important aspect to reflect the objective of study with some level of realism. The statistical relationship give a starting point for further research. A linear regression model is used to discuss and explain the relationship between variables. The simplest way is when there are only two variables (x_i, y_i) with n observations and a linear regression model is assumed. Such model is used to study how just one variable called response variable (Y) relies on the other, called predictor variable (X). Once there exists a linear relationship, then try to model the relationship mathematically and use the model for prediction (Seber & Lee, 2012).

Let Y_i and $X_i, i = 1, 2, \cdots, n$ be respectively response and predictor variables. A simple linear regression model (Kutner, 2005) can be defined as

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.1}$$

where

β_0 is the y-intercept of regression line,

β_1 is the slope of regression line,

ϵ_i is a random error and assumed to be independent and identically distributed (i.i.d). For a special case of normality assumption, we have $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. This gives $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, where $\mu_i = \mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$.

In general, to model the relationship between two or more predictor variables and a response variable, a multiple linear regression is applied. Given a p predictor variables of n observations, the multiple linear regression model can be expressed as follows

$$Y_i = \beta_0 + \beta_1 \mathrm{x}_{1i} + \beta_2 \mathrm{x}_{2i} + \cdots + \beta_p \mathrm{x}_{pi} + \epsilon_i. \tag{2.2}$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. Again, this provides $Y \sim \mathrm{N}(\mu_i, \sigma^2)$, where $\mu_i = \mathrm{E}(Y_i) = \beta_0 + \beta_1 \mathrm{x}_{1i} + \beta_2 \mathrm{x}_{2i} + \cdots + \beta_p \mathrm{x}_{pi}$.

In this thesis, however, we will focus on the simple linear regression model. In order to use this approach sufficiently, there are four fundamental assumptions: linearity and independence, homoscedasticity, and normality of the error term that must be verified. Moreover, the parameter estimation and related statistical inferences of a linear regression model are usually based on the assumption that the error term is homoscedastic and normally distributed (Moore et al., 2009; Anderson et al., 2017). The results of the linear relationship, parameter estimation and statistics inferences may be unreliable, if the assumptions of the error term ϵ_i appear questionable.

2.1.1 Linearity assumption

In many situations, such a prior to performing linear regression analysis, researchers want to test their data for linearity. Linearity means that two variables, "y" and "x", are related based on a mathematical equation "$y = cx,$" where c is any constant number. In order to detect linearity of the variables in our data, the scatter plot of the independent variable and dependent variable must be performed.

2.1.2 Independence assumption

The independence of the error term is one of the assumptions. To verify whether the data meet this assumption Durbin-Watson statistics is used. The corresponding hypothesized statement is as follows.

$$\mathrm{H}_0 : \rho = 0$$

$$\mathrm{H}_1 : \rho \neq 0,$$

where ρ is the population correlation coefficient between values of ϵ_i and ϵ_{i+1}, $i = 1, \cdots, n$.

The Durbin-Watson test is used to confirm whether $\rho = 0$. To be simply, the i^{th} residual, the estimated value of ϵ_i, denoted by $\epsilon_i = y_i - \hat{y}_i$, where y and $\hat{y}$ are observed and predicted value of the response variables for individual i respectively, is used for calculating the Durbin-Watson test statistic. The test statistics given in Weiers (2010) is expressed as

$$D = \frac{\sum_{i=1}^{n} (\epsilon_{i+1} - \epsilon_i)^2}{\sum_{i=1}^{n} \epsilon_i^2}, \tag{2.3}$$

The smaller values of D compared to D_U and D_L, suggest the strong correlation. Upper and lower critical values, D_U and D_L have been listed in table shown in (Weiers, 2010). The decision making related to the hypothesis is that if D $<$ D_L, H_0 is rejected. If D $>$ D_U, H_0 is not rejected. If D_L $<$ D $<$ D_U test, it is inconclusive.

2.1.3 Homoscedasticity assumption

To see whether the variance of the error term are equal, Levene's Test and Bartlett's Test are used to test if k samples have equal variances. If we assume

that our data are normally distributed then Bartlett's Test is more appropriate than Levene's Test (Layard, 1973; Brown & Forsythe, 1974).

Levene's test

Levene's test statistic, Levene (1960) is used to test whether k samples have equal variances. The hypothesis for Levene's test according to Brown & Forsythe (1974) is given as follows

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

$$\mathrm{H}_1 : \sigma_j^2 \neq \sigma_l^2 \quad \text{for at least one pair, } (j \neq l).$$

Given an $\mathrm{N} = \sum_{j=1}^{k} n_j$ observations of y random variable contains k subgroups and denoted n_j is sample size of j^{th} subgroup.

$$\mathrm{W} = \frac{\sum_{j=1}^{k} n_j(\bar{z}_j - \bar{z}_{..})/(k-1)}{\sum_{j=1}^{k}\sum_{l=1}^{n_j}(z_{jl} - \bar{z}_j)^2 / \sum_{j=1}^{k}(n_j - 1)}, \tag{2.4}$$

where

$z_{jl} = |y_{jl} - \bar{y}_j|$,

$\bar{z}_j = \sum_{j=1}^{k} z_{jl}/n_j$ is the mean of the z_{jl} for group j,

$\bar{z}_{..} = \sum_{j=1}^{k}\sum_{l=1}^{n_j} z_{jl} / \sum_{j=1}^{k} n_j$ is the overall mean of the z_{jl}.

The critical region of rejecting the null hypothesis is when $\mathrm{W} > \mathrm{F}_{\alpha,k-1,n-k}$, where $\mathrm{F}_{\alpha,k-1,n-k}$ is the upper critical value of the F distribution with $k-1$ and $n-k$ degrees of freedom at a significant level of α.

Bartlett's test

Bartlett's test is used to test for homogeneity of variance. Assume we have k independent samples with sizes n_j of j^{th} subgroup then the hypotheses

are given by

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$$

$$\mathrm{H}_1 : \sigma_j^2 \neq \sigma_l^2 \quad \text{for at least one pair, } (j \neq l).$$

The Bartlett's test statistic is as follows

$$\chi^2 = \frac{q}{c}, \tag{2.5}$$

where

$q = (\mathrm{N} - k)\log(s_p^2) - \sum_{j=1}^{k} \big(n_j - 1\big)\log(s_j^2),$

$\mathrm{N} = \sum_{j=1}^{k} n_j,$

$s_p^2 = \frac{1}{\mathrm{N}-k} \sum_{j=1}^{k} (n_j - 1)s_j^2,$

$s_j^2 = \frac{1}{\mathrm{N}-1} \sum_{j=1}^{k} (y_j - \bar{y})^2,$ and

$c = 1 + \frac{1}{3(k-1)} \left(\sum_{j=1}^{k} \left(\frac{1}{n_j - 1}\right) - \frac{1}{\mathrm{N} - k} \right).$

Bartlett's test has been shown in Layard (1973) that it is approximately distributed as χ^2 with $k - 1$ degrees of freedom. Therefore, the critical value is $\chi^2_{\alpha,k-1}$. If χ^2 is greater than the critical value, we reject the null hypothesis. Then there exist at least one pair of population variance different from the others.

2.1.4 Normality assumption

Normally distributed assumption is quite important to provide the statistical inference about the ϵ_i. When the normality are not verified, then the interpretation about the inference may not valid or reliable (Razali & Wah, 2011). Kolmogorov-Smirnov test and Shapiro-Wilk test are commonly used to confirm whether the normality of the residuals are met. However, Mendes & Pala (2003)

and Razali & Wah (2011) have suggested that Shapiro-Wilk test is the most appropriate normality test for all types of sample sizes while Kolmogorov-Smirnov test is least powerful test.

The hypothesis for testing normality is given by

H_0 : The sample was drawn from a normal distribution

H_1 : The sample was not drawn from a normal distribution

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov statistic gives a mean of testing a set of data points that are from continuous distribution (Lilliefors, 1967). Mendes & Pala (2003) and Razali & Wah (2011) used the statistic test proposed by Kolmogorov (1933) then developed by Smirnov (1939) as follows. Given an n order of observations, $y_1 < y_2 < \cdots < y_n$,

$$T = \max_y |F^*(y) - F_n(y)| \tag{2.6}$$

where

$F_n(y)$ is the sample cumulative distribution function,

$F^*(y)$ is the cumulative normal distribution function with $\bar{y}$: the sample mean, and s^2 : the sample variance.

If the value of T exceeds the critical value, one rejects H_0 at level of significance α that the observations are from normal distribution.

Shapiro-Wilk test

The Shapiro-Wilk test is used to detect the violation of normality (Razali & Wah, 2011). Shapiro-wilk test statistic (Shaphiro & Wilk, 1965) was firstly suitable for the sample size of less than 50. It has become an appealing choice since it had good power properties, see the detail for example in Mendes & Pala (2003)

and Razali & Wah (2011). Given an order of n random sample, $y_1 < y_2 < \cdots < y_n$, Shapiro-Wilk test statistics referencing is defined as follows.

$$\mathrm{W} = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{2.7}$$

where

y_i is the i^{th} order statistic,

$\bar{y}$ is the sample mean,

$a_i = (a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$, and

$m = (m_1, m_2, \cdots, m_n)^T$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics.

The value of W lies between zero and one. Small values of W lead to the rejection of normality (Mendes & Pala, 2003; Razali & Wah, 2011).

2.2 Estimation methods

2.2.1 Ordinary least squares estimation (OLS)

Under model (2.1), the OLS estimators of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ can be obtained by minimizing the sum of squared error; $Q = \sum_{i=1}^{n} \epsilon_i^2$, where

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \Big(y_i - \beta_0 - \beta_1 x_i\Big)^2. \tag{2.8}$$

For a sample of n observations (x_i, y_i), the score functions can be derived by differentiating (2.8) with respect to β_0 and β_1.

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} \big(y_i - \beta_0 - \beta_1 x_i\big)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \big(y_i - \beta_0 - \beta_1 x_i\big)$$

Then set the partial derivatives equal to zero, using $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the particular estimate values of β_0 and β_1 that minimize Q.

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x^2 = 0$$

Simplifying and solving the above equations, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained as follows.

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right) = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.9}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}. \tag{2.10}$$

where $\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \neq 0$.

Gauss-Markov theorem in (Kutner, 2005) stated that the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are more efficient among all unbiased linear estimators.

2.2.2 The transform-both-sides estimation (TBS)

One way to remedy heteroscedasticity and non-normality is to apply transform both sides model (Ávila et al., 2013; Carroll & Ruppert, 1984, 1988; Miller, 1984; Newman, 1993). The transformation allows the parameters to be estimated efficiently. In common practice, the model of the logarithmic transformation on response and explanatory variables has been used and a TBS model is defined as

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i. \tag{2.11}$$

The parameters estimates $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is obtained by using OLS methods, where the sum of square error of (2.11) is given by.

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \Big(\log(y_i) - \beta_0 - \beta_1 \log(x_i) \Big)^2. \tag{2.12}$$

The score functions can be derived by differentiating (2.12) with respect to β_0 and β_1.

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \Big(\log(y_i) - \beta_0 - \beta_1 \log(x_i) \Big)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} \log(x_i) \Big(\log(y_i) - \beta_0 - \beta_1 \log(x_i) \Big)$$

Then set both partial derivatives to zero, using $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the particular estimate values of β_0 and β_1 that minimize S.

$$\sum_{i=1}^{n} \log(y_i) - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=0}^{n} \log(x_i) = 0$$

$$\sum_{i=1}^{n} \log(x_i) \log(y_i) - \hat{\beta}_0 \sum_{i=1}^{n} \log(x_i) - \hat{\beta}_1 \sum_{i=1}^{n} [\log(x_i)]^2 = 0$$

After simplifying and solving the above equations the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, respectively are

$$\hat{\beta}_0 = \frac{1}{n} \Big(\sum_{i=1}^{n} \log(y_i) - \hat{\beta}_1 \sum_{i=1}^{n} \log(x_i) \Big) = \overline{\log(y)} - \hat{\beta}_1 \overline{\log(x)}, \tag{2.13}$$

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} \log(x_i) \log(y_i) - \overline{\log(y)} \sum_{i-1}^{n} \log(x_i)}{\displaystyle\sum_{i=1}^{n} \log^2(x_i) - \overline{\log(x)} \sum_{i=1}^{n} \log(x_i)}, \tag{2.14}$$

where,

$\overline{\log(x)} = \frac{1}{n} \displaystyle\sum_{i=1}^{n} \log(x_i)$, and

$\overline{\log(y)} = \frac{1}{n} \displaystyle\sum_{i=1}^{n} \log(y_i)$.

2.2.3 Variance functions estimation

Another way to control heteroscedasticity is by modifying the constant error variance assumption (Ávila et al., 2013). A commonly applicable assumption about the model error variance is the power of the mean variance function model (POM) as given in Pinheiro & Bates (2000) and Sidik & Jonkman (2016). That is

$$\text{Var}(y_i) = \sigma^2 \hat{y}_i^{\theta}, \tag{2.15}$$

where σ and θ are the unknown parameters of the variance function model. This variance function is useful to represent the situation that the observations increase with $\hat{y}_i$. In general, when $\theta = 0$ it refers to homoscedasticity or constant variance model. $\theta = 1$ represents the variance proportional to the mean or the data come from a Poisson distribution and when $\theta = 2$ it corresponds to constant coefficient of variation (CV) model (Pinheiro & Bates, 2000).

Another extended variance function model is an exponential variance function (VEXP)

$$\text{Var}(y_i) = \sigma^2 \text{exp}(2\theta \hat{y}_i). \tag{2.16}$$

In practice, one may have prior knowledge about the value of θ in certain cases. Therefore, the parameter $\boldsymbol{\beta}$ may be estimated by the weighted least squares method with weights $1/\hat{y}^{\theta}$ or $1/\text{exp}(2\theta\hat{y}_i)$ for models (2.15) and (2.16) respectively (Sidik & Jonkman, 2016). Nevertheless, Giltinan & Ruppert (1989) stated that one must be clear about the true value of θ, thus the right weight scheme can be incorporated in fitting linear models.

To estimate the parameters of model (2.15) and (2.16), several methods have been proposed in Beal & Sheiner (1988). However, it has been proved that the generalized least square (GLS) is better and simpler than that produced by another in certain situation, for more detail, see Carroll & Ruppert (1982), Davidian & Carroll (1987) and Beal & Sheiner (1988). Consequently, the most common method to estimate parameters $\boldsymbol{\beta}$ is GLS method. The GLS procedure has been previously

described in Carroll & Ruppert (1982), Giltinan et al. (1986), Davidian & Carroll (1987), Giltinan & Ruppert (1989), and Sidik & Jonkman (2016).

Assume that the responses are normally distributed, the parameter estimates can be obtained based on the three following steps:

1. Obtain a preliminary unweighted least squares to estimate $\hat{\beta}_p$ of β,

2. Estimate parameters (σ, θ) from maximum likelihood function, letting that $\beta = \hat{\beta}_p$ and form estimated standard deviations $\hat{\sigma}_i = \sigma^2 \hat{y}_i^{\theta}$ and $\hat{\sigma}_i = \sigma^2 \text{exp}(2\theta \hat{y}_i)$ for model (2.15) and (2.16) respectively,

3. Repeat step 1 with the transformed y variables $y_i^* = y_i/\hat{\sigma}_i$, $x_i^* = x_i/\hat{\sigma}_i$ to obtain the final parameter estimates of $\hat{\theta}$, $\hat{\sigma}^2$, and $\hat{\beta}$.

2.3 Model comparisons and confidence intervals

2.3.1 Bias

Let $\hat{\theta}$ be a point estimator for parameter θ. If the difference between the expected value of point estimates and the parameter equals to zero (i.e $\text{E}[\hat{\theta}] - \theta = 0$), then $\hat{\theta}$ is said to be an unbiased estimator. If $\text{E}[\hat{\theta}] - \theta \neq 0$, the bias of the point estimator is given by (Wackerly et al., 2014)

$$\text{Bias}_{\theta}(\hat{\theta}) = \text{E}_{\theta}[\hat{\theta} - \theta] = \text{E}_{\theta}(\hat{\theta}) - \theta. \tag{2.17}$$

2.3.2 Mean squared error (MSE)

Mean squared error (MSE) is used to describe average model performance error. MSE of an estimator measures the average of the square difference between estimated values and underlying parameter estimated. Thus, in addition to preferring unbiasedness, we want the error of an estimator and its target parameter to be as small as possible. The MSE of an estimator is given by (Wackerly et al., 2014)

$$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2]. \tag{2.18}$$

2.3.3 Confidence intervals

To observe data that might contain the true value of unknown population parameter, one may wish to obtain interval predictors for θ of linear regression coefficients. We assume that the data are approximately normally distributed. As a result, we can produce an approximate $100(1-\alpha)\%$ confidence interval for parameters θ as the form

$$\text{Lower}(\hat{\theta}) = \hat{\theta} - t_{\alpha/2,n-2}\text{SE}_{\hat{\theta}}, \tag{2.19}$$

$$\text{Upper}(\hat{\theta}) = \hat{\theta} + t_{\alpha/2,n-2}\text{SE}_{\hat{\theta}}. \tag{2.20}$$

where $t_{\alpha/2,n-2}$ is the upper quantile of the t-distribution with $n-2$ degrees of freedom and $\text{SE}_{\hat{\theta}}$ is the standard error of each parameter estimates $\hat{\theta}$ (Davidian & Carroll, 1987; Moore et al., 2009).

In this thesis, there are two quantities associated with confidence intervals:

Coverage probability refers to a procedure for constructing a random probability region that produce an interval covering or containing the true value of the parameters of interest θ divided by the number of replications.

$$\frac{\overset{\text{N}}{\underset{r=1}{\#}}\left(\text{Lower}(\hat{\theta}_r) \leq \theta \leq \text{Upper}(\hat{\theta}_r)\right)}{\text{N}}. \tag{2.21}$$

where

Lower$(\hat{\theta}_r)$ is the lower bound of parameter θ at r^{th} replication,

Upper$(\hat{\theta}_r)$ is the upper bound of parameter θ at r^{th} replication,

N is the number of replications of parameter θ.

Average length is simply the mean of the difference in two end-points. We are most interested in obtaining confidence intervals that are narrow as possible. That is, it gives more specific range of realism of the population θ.

$$\text{Average lenght}(\hat{\theta}) = \frac{\sum_{r=1}^{\text{N}} \left(\text{Upper}(\hat{\theta}_r) - \text{Lower}(\hat{\theta}_r)\right)}{\text{N}} \tag{2.22}$$

where

Upper$(\hat{\theta}_r)$ is the upper bound of parameter θ at r^{th} replication,

Lower$(\hat{\theta}_r)$ is the lower bound of parameter θ at r^{th} replication,

N is the number of replications of parameter θ.

Chapter 3

Methodology

In this chapter, we will describe about the methods applied to carry out the study. We will begin with the simulation process incorporate with different distributions and variance patterns. Then we will demonstrate how to apply the studied methods to the real life data.

3.1 The simulation study

This section contains a process of simulation study to assess and compare the performance of the four studied estimation methods discussed in the previous chapter. The assessment of the estimates is based on bias, MSE and coverage probability of the nominal 95% confidence interval.

In the simulation study, random sample sizes n (= 20, 50, 100, and 200) were generated and values of X were drawn independently from the uniform distribution over [0,1] with corresponding values of Y given by

$$\mathrm{Y}_i = 3 + x_i + \epsilon_i, \tag{3.1}$$

where

$\beta_0 = 3$ and $\beta_1 = 1$,

ϵ_i are independently drawn from the following four studied symmetric distributions.

1. Standard normal distribution,
2. Laplace distribution,
3. Location scale of t-distribution, and
4. Logistic distribution.

All distributions were set with the mean of 0, $\big(\mathrm{E}(\epsilon_i) = 0\big)$ and three different variance structures in order to form heteroscedasticity.

The three variance structures consist of power of the predictor $\mathrm{Var}(\epsilon_i) = \sqrt{x_i}$, exponential function of the predictor $\mathrm{Var}(\epsilon_i) = e^{x_i}$, and a constant plus fitted values $\mathrm{Var}(\epsilon_i) = 0.5 + \hat{y}_i$. The simulation process was replicated $\mathrm{N} = 1,000$ times. The parameters estimates, biases, MSEs and coverage probabilities were then calculated.

Under model (3.1), we want to make a random variable (Y_i) to form heteroscedasticity and to avoid producing negative values of Y_i. We observed that, when β_1 is large model (3.1) does not generate heteroscedasticity. Moreover, we also suggest to choose any appropriate values of β_1 which is not much greater than β_0, to avoid the negative values of Y_i. In this thesis, however, we choose the values 3 and 1 for β_0 and β_1, respectively.

Moreover, in order to visualize the simulated data, we plot the scatter plot of original data of (X_i, Y_i) and estimated line using OLS method for each case of distributions and sample sizes. The non-constant variance structure of $\sqrt{x_i}$ are shown in Figure 3.1-3.4 under standard normal, laplace, location scale of t , and logistic distribution, respectively. The non-constant variance of e^{x_i} are shown in Figure 3.5-3.8 under standard normal, laplace, location scale of t, and logistic distribution, respectively. The non-constant variance structure of $0.5 + \hat{y}_i$ are shown in Figure 3.9-3.12 under standard normal, laplace, location scale of t, and logistic distribution, respectively.

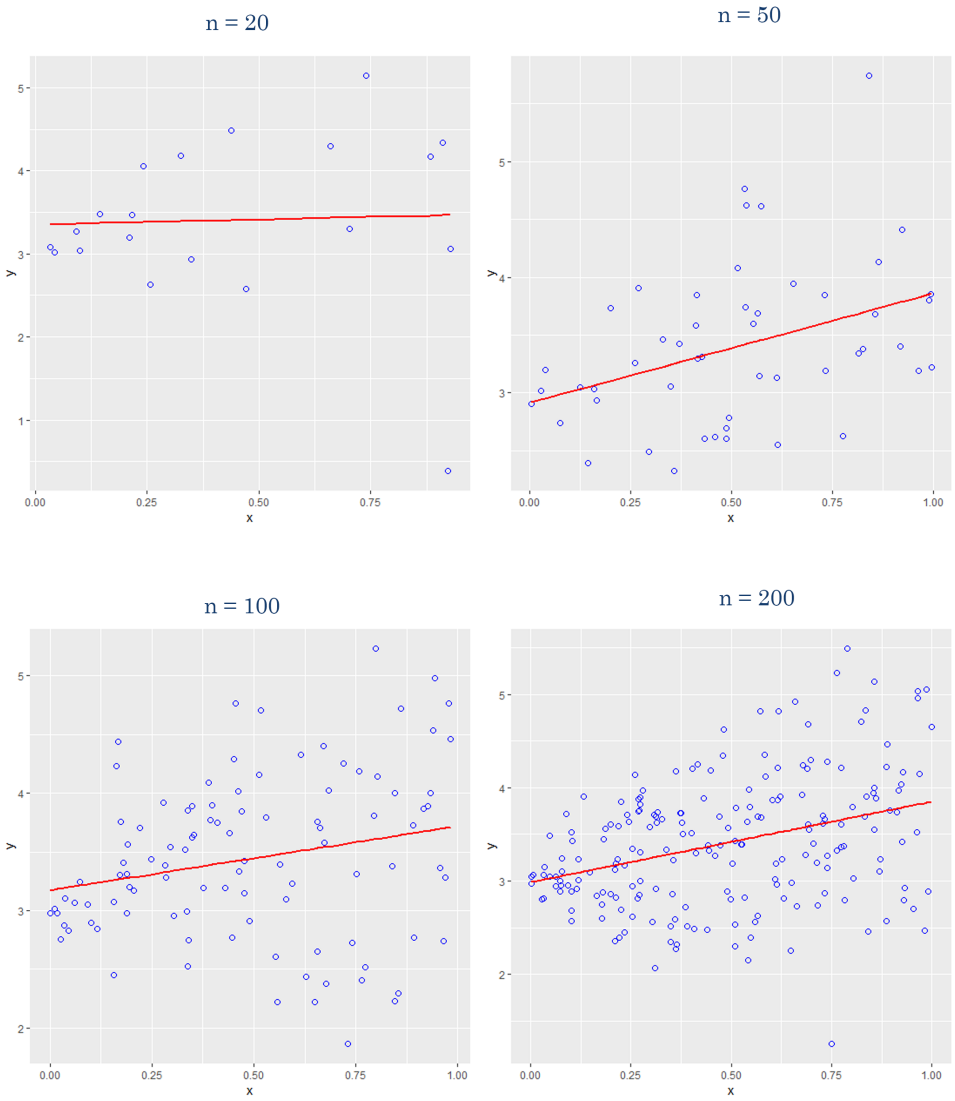

Figure 3.1: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = \sqrt{x_i}$ under standard normal distribution.

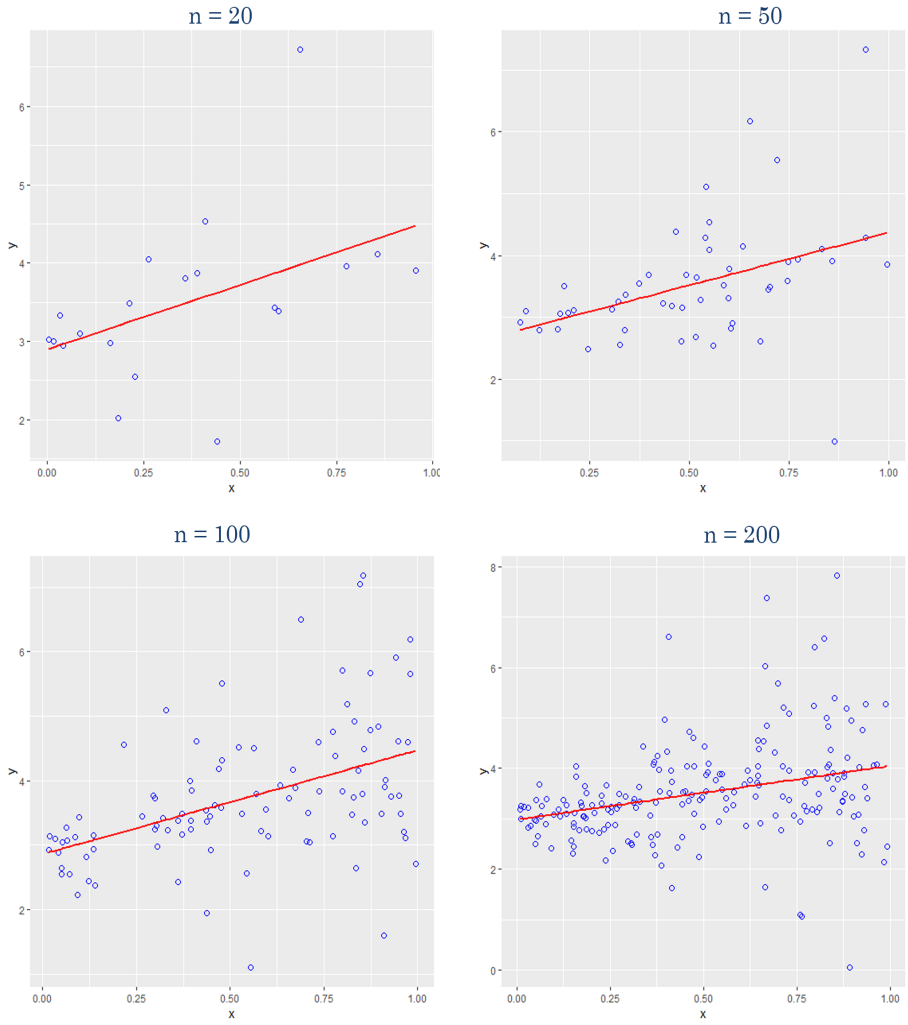

Figure 3.2: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = \sqrt{x_i}$ under laplace distribution.

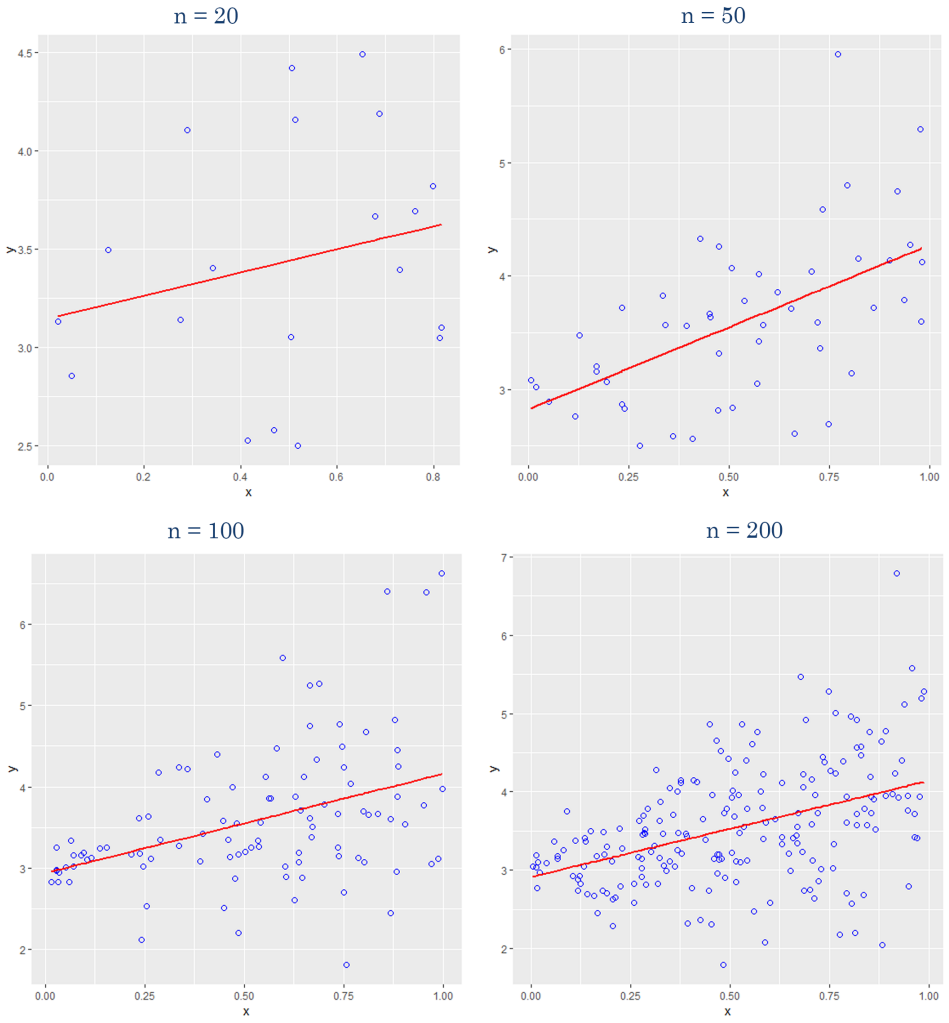

Figure 3.3: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = \sqrt{x_i}$ under location scale of t distribution.

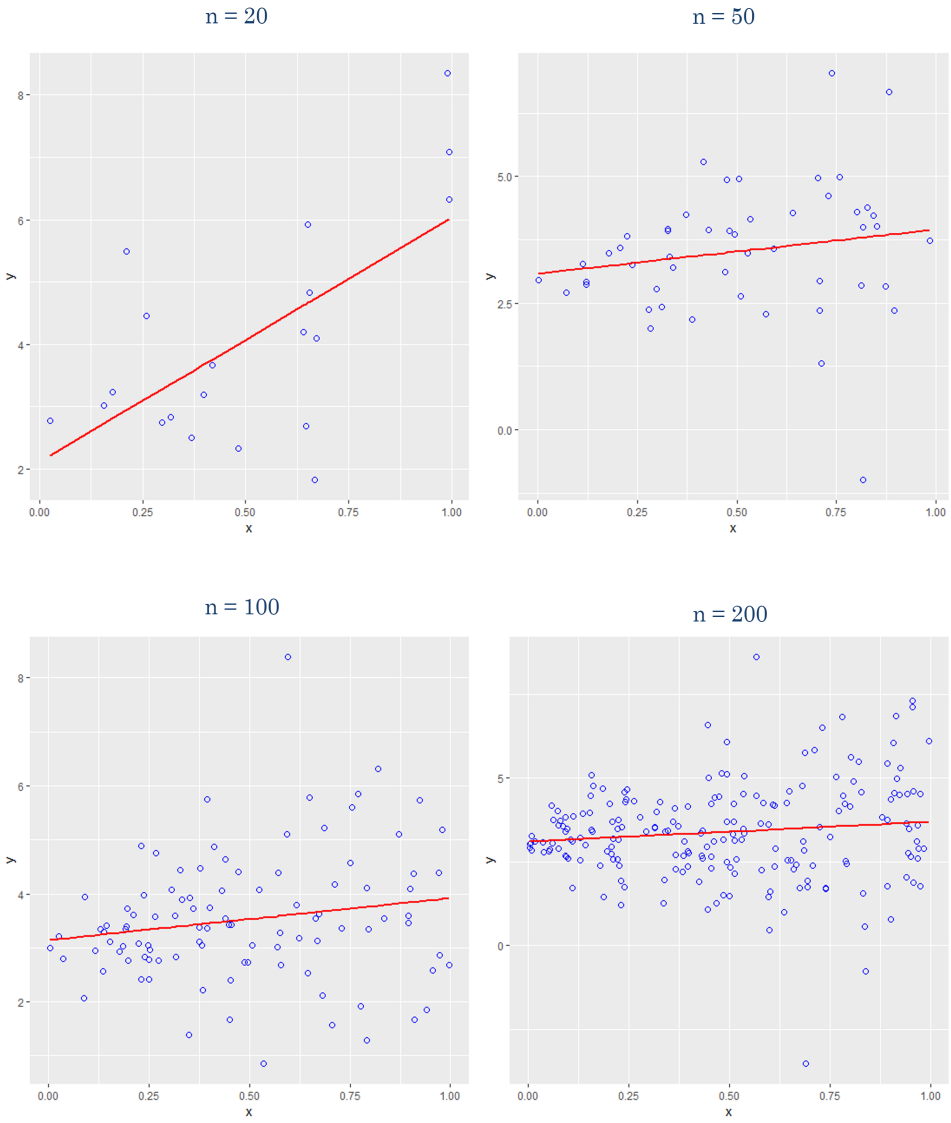

Figure 3.4: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = \sqrt{x_i}$ under logistic distribution.

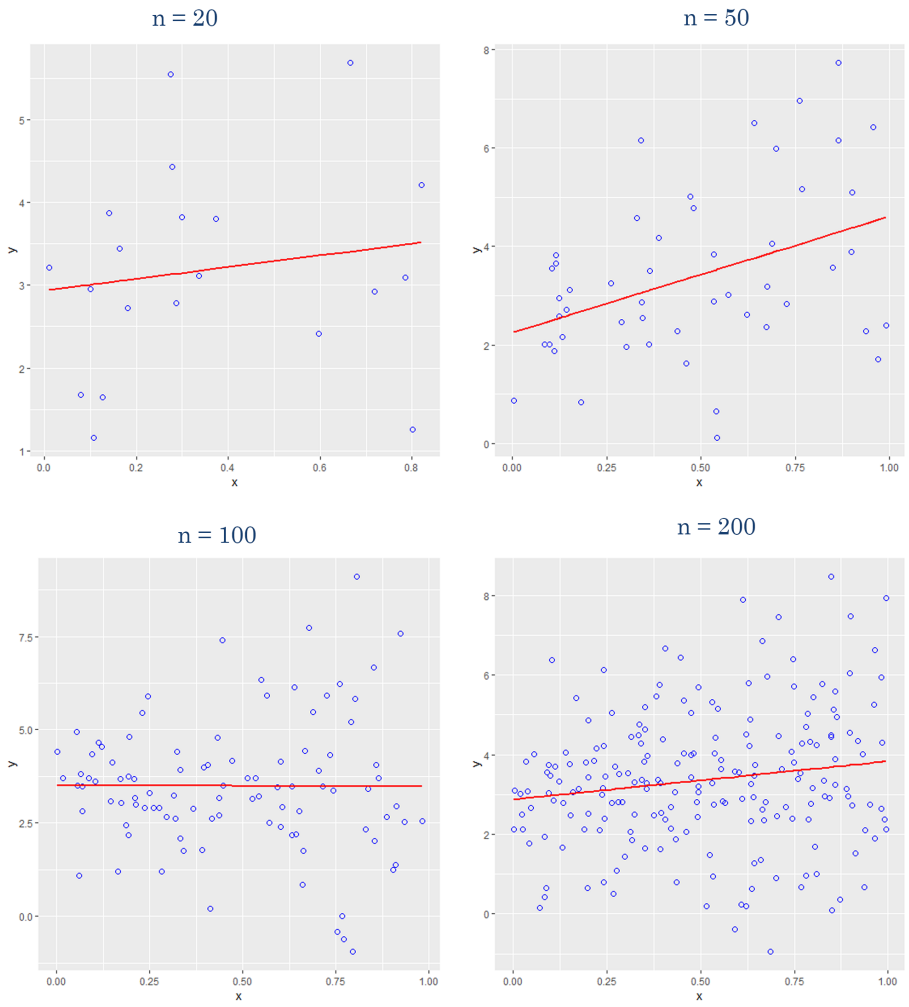

Figure 3.5: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = e^{x_i}$ under standard normal distribution.

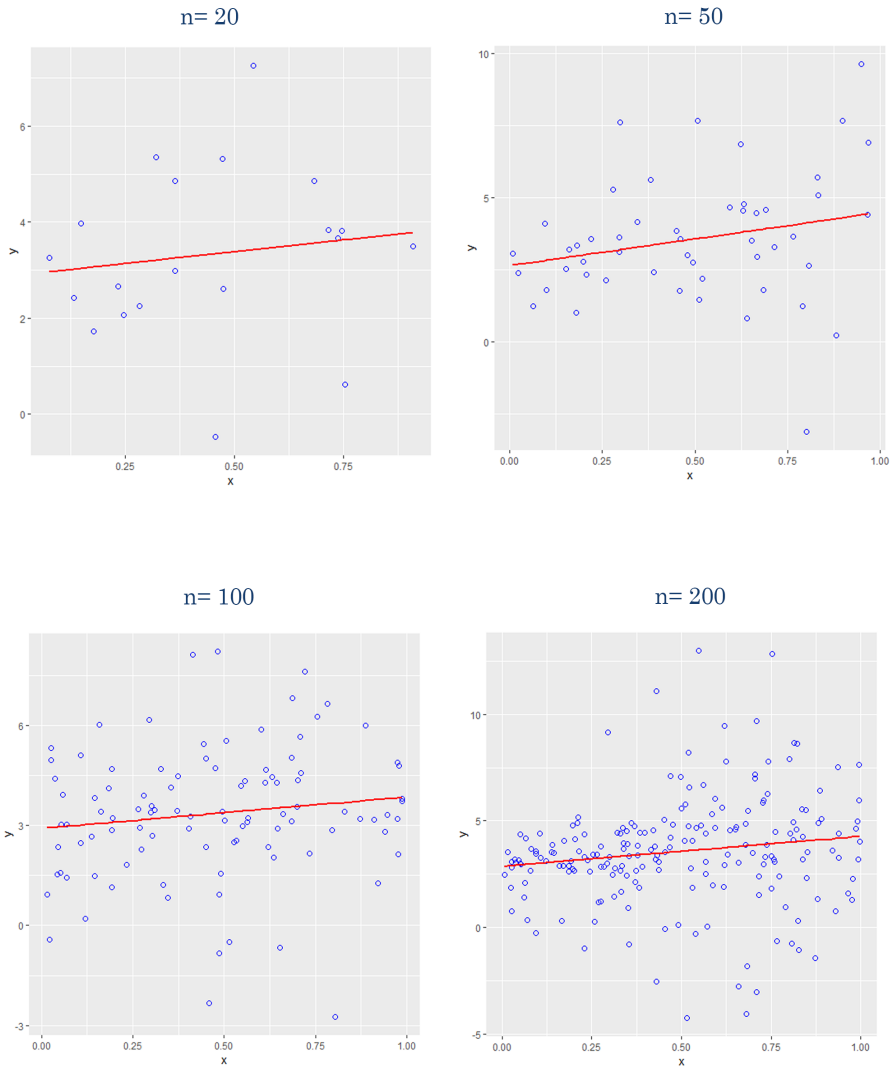

Figure 3.6: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = e^{x_i}$ under laplace distribution.

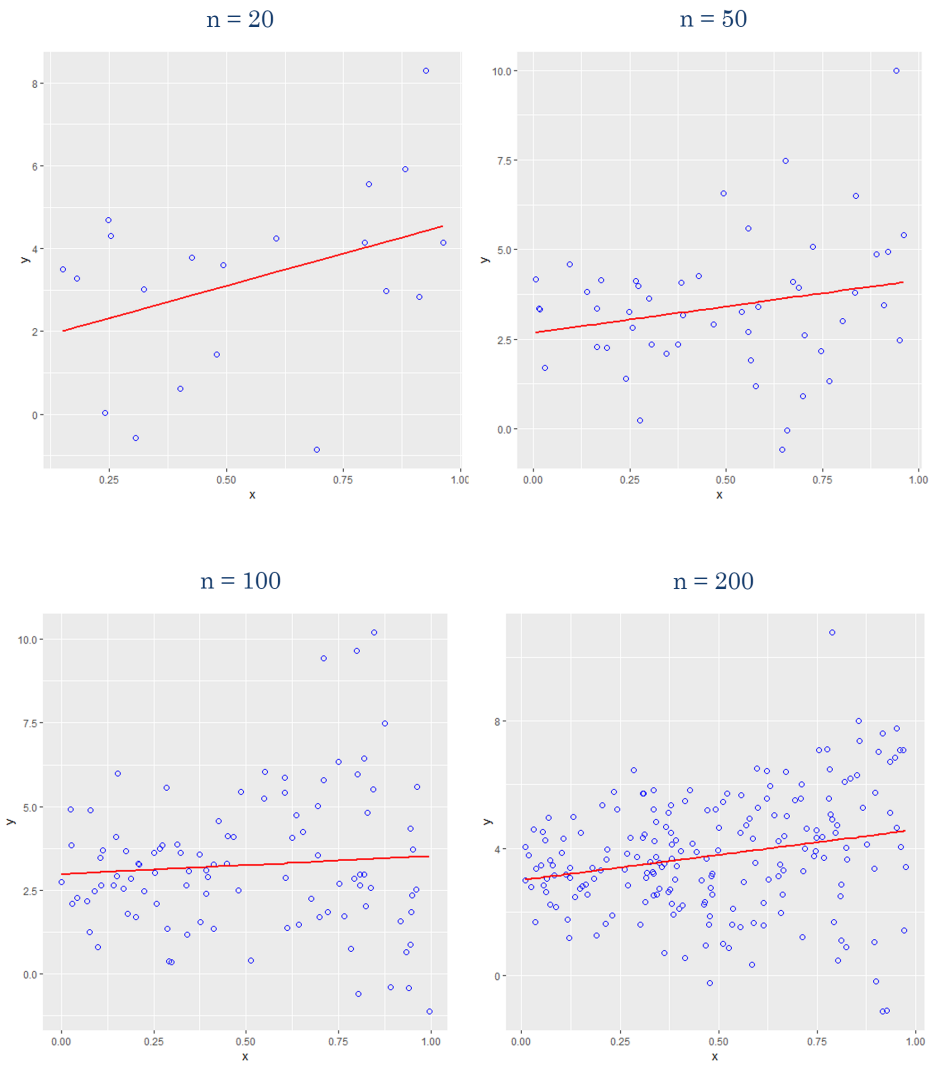

Figure 3.7: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = e^{x_i}$ under location scale of t distribution.

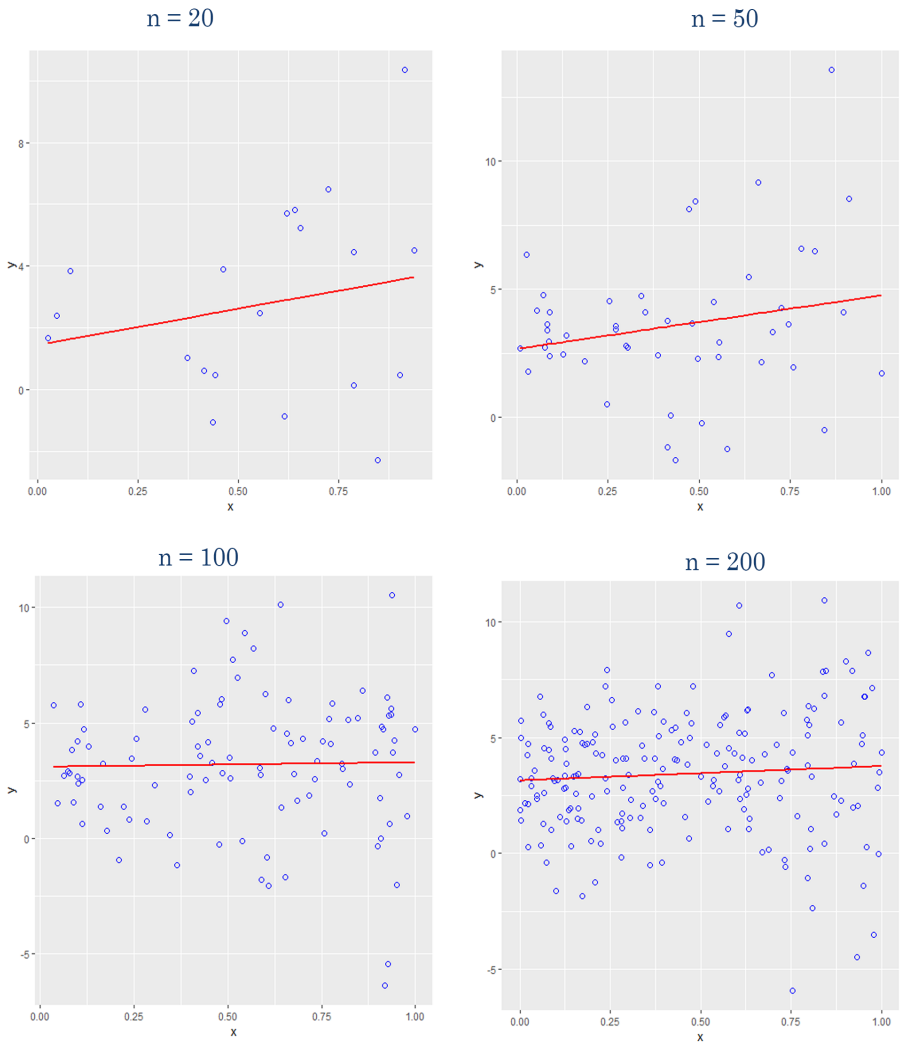

Figure 3.8: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = e^{x_i}$ under logistic distribution.

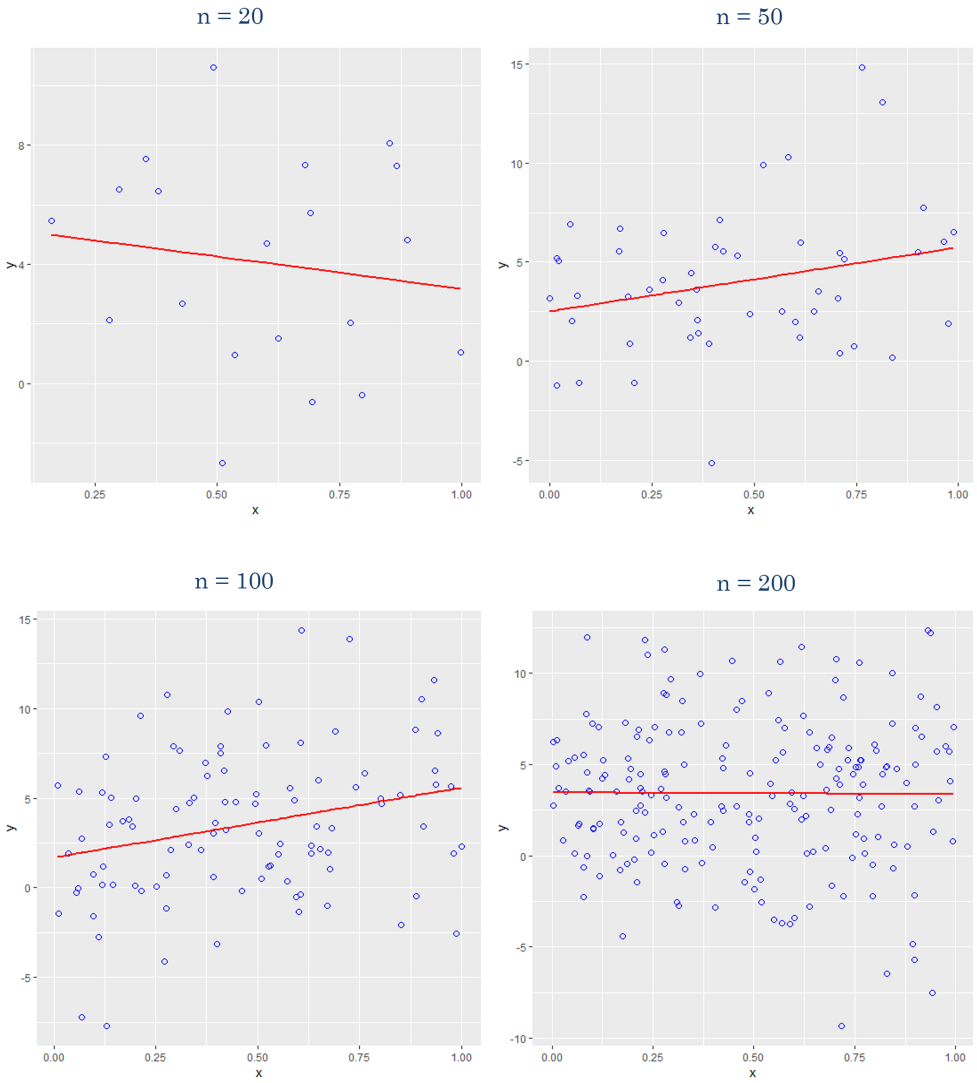

Figure 3.9: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = 0.5 + \hat{y}_i$ under standard normal distribution.

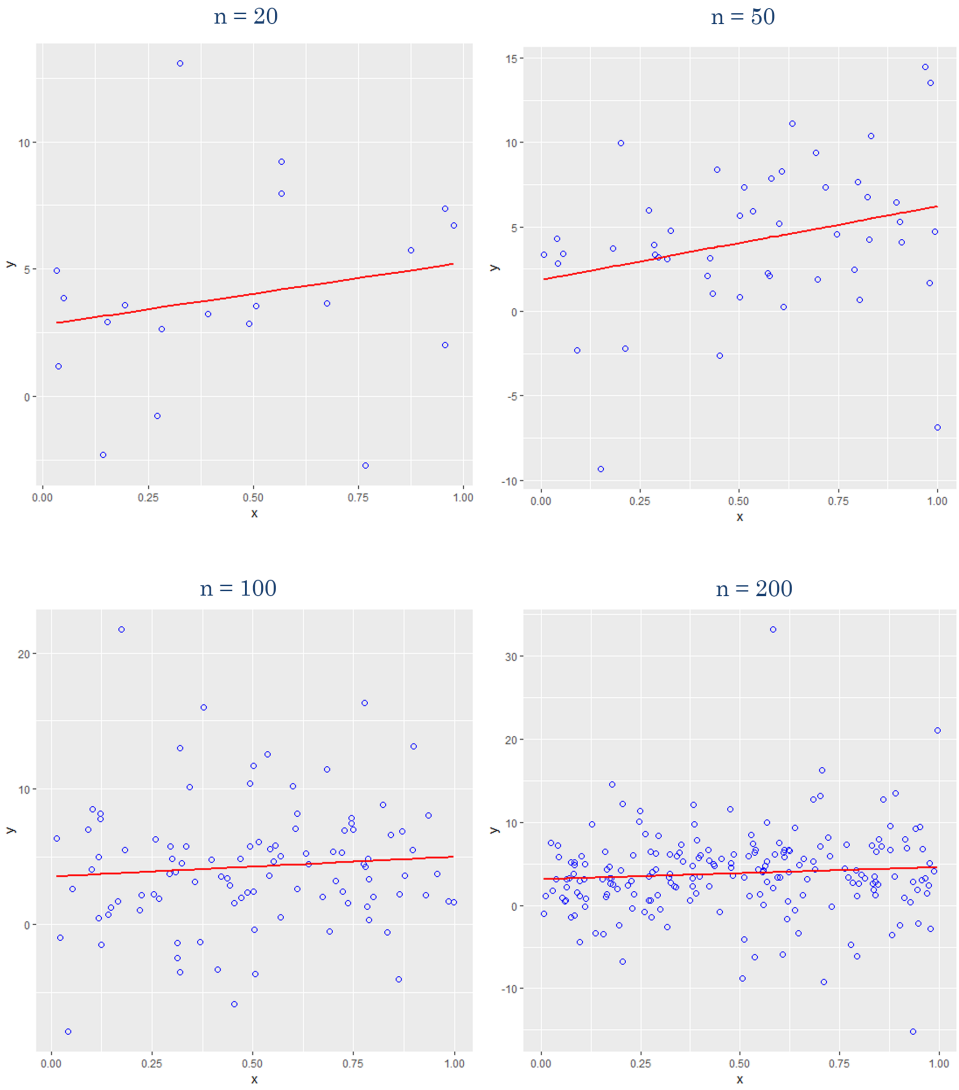

Figure 3.10: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = 0.5 + \hat{y}_i$ under laplace distribution.

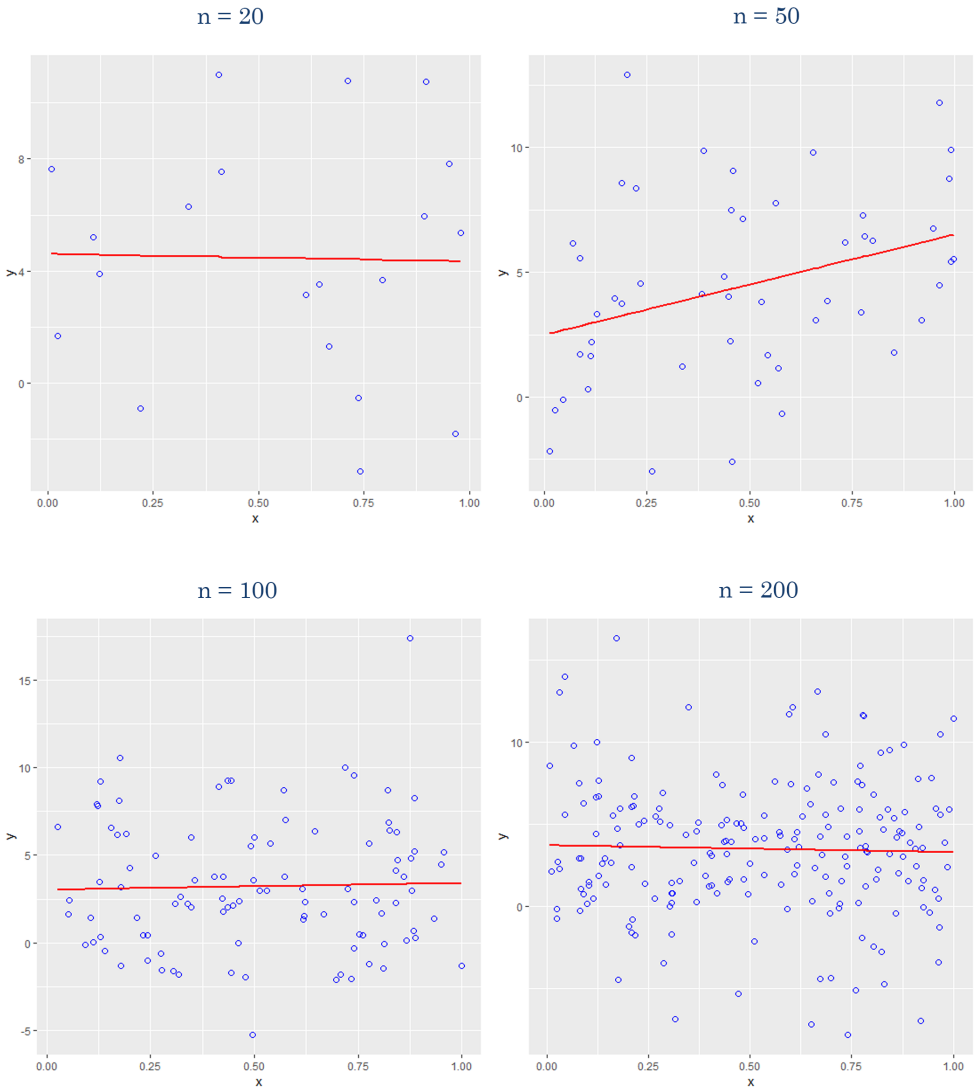

Figure 3.11: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = 0.5 + \hat{y}_i$ under location scale of t distribution.

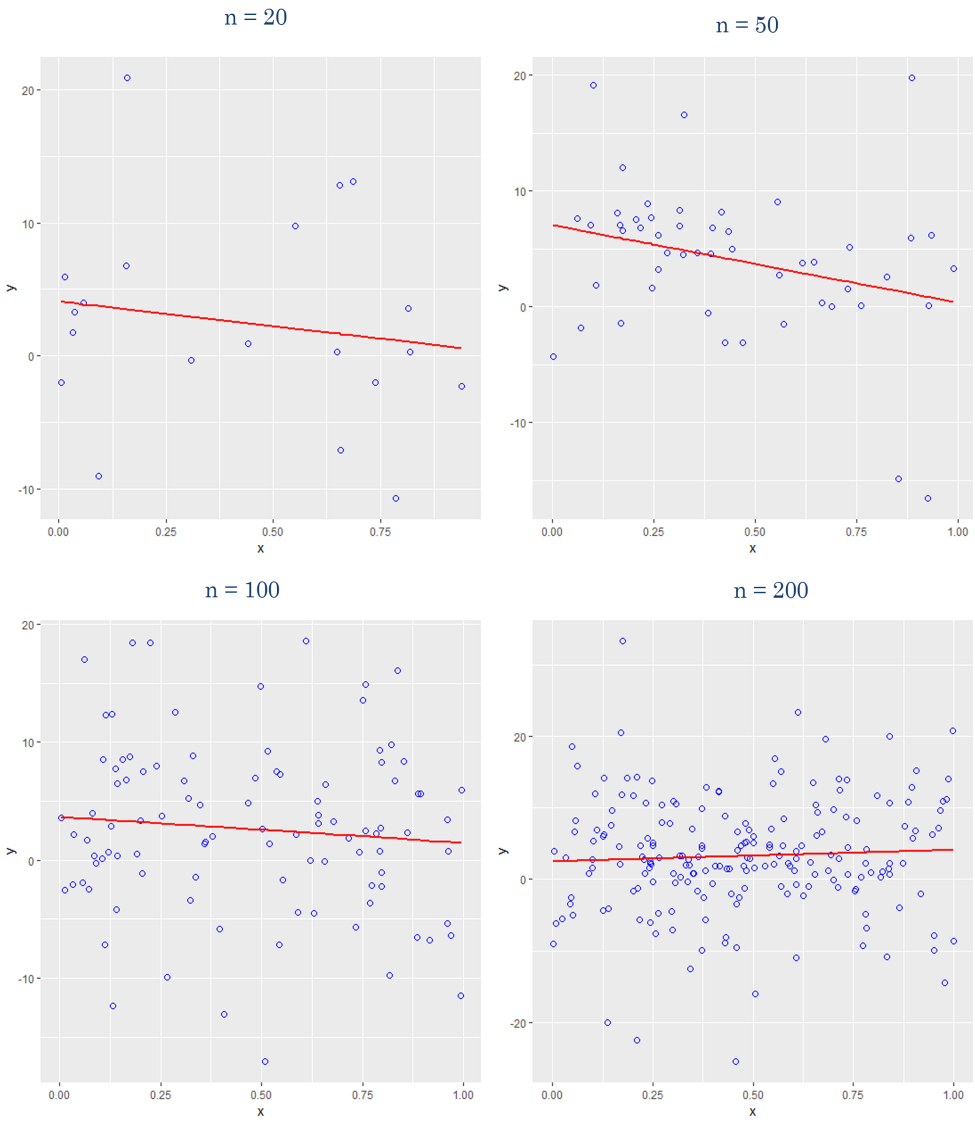

Figure 3.12: The scatter plot of original data as well as the estimated solid line of the linear variance function $\text{Var}(\epsilon_i) = 0.5 + \hat{y}_i$ under logistic distribution.

3.2 Practical applications

In this section, we apply all the studied methods to real dataset, named here is bike sharing data. The data was extracted from Fanaee T & Gama (2013). The data retrieved from https://archive.ics.uci.edu/ml/datasets/Bike+ Sharing+Dataset: March 14th 2018, 6:40PM. To form the relationship between the count of rental bikes (cnt) and count of registered users (registered) collected from the period 1 January 2011 to 31 December 2012 for Washington, D.C., USA, the regression analysis was used. The results will be shown in Figures 4.1 and 4.2. The models validation of the errors assumptions by using Shapiro-Wilk test for normality and Bartlett's test for homoscedasticity (Salvo et al., 2012) were conducted. The residual plots and prediction intervals were achieved to evaluate the performance of each studied method.

The results, reported in Table 3.1 and 3.2, have shown that this dataset is heteroscedastic with non-normal distribution.

To calculate the Bartlett's test, the quartiles of response variable are used in order to manage our data into four distinct subgroups as follows: the random numbers of response variable that less than the first quartile are recorded as group 1, the numbers in the range between the first quartile and the second quartile are recorded as group 2, as well as the numbers that greater than the second quartile and less than or equal the third quartile are recorded as group 3, and for the rests are recorded as group 4.

Table 3.1: Bartlett test for homogeneity of variance

Bartlett's Test	Chi-squared	Degrees of freedom	p-value
	140.47	3	$<$ 2.2e-16

The p-value, reported in Table 3.1 is less than 0.05 which indicates strong evidence of heterogeneous variance.

Table 3.2: Shapiro-Wilk normality test

Shapiro-Wilk	W-test statistics	p-value
	0.98	2.081e-08

The p-value, reported in Table 3.2 is less than 0.05 indicating that the data are not normally distributed.

Chapter 4

Results and Discussion

This chapter comes up with the result interpretation and comparison of each model's performance in the simulation study and real life data. All results were achieved by using R software (R Core Team, 2017).

4.1 Results and discussion of simulation study

For each sample size and variance, the tables show the relative biases, MSEs, as well as the 95% empirical coverage probabilities and average lengths of confidence intervals (in parentheses). Tables 4.1-4.4 show the results of the non-constant variance Var$(\epsilon_i) = \sqrt{x_i}$ for sample sizes 20, 50, 100, and 200, respectively. Tables 4.5-4.8 is about the non-constant variance Var$(\epsilon_i) = e^{x_i}$ for the sample sizes 20, 50, 100, and 200, respectively. Consequently, Tables 4.9-4.12 obtained the results of non-constant variance Var$(\epsilon_i) = 0.5 + \hat{y}_i$ for sample sizes 20, 50, 100, and 200, respectively.

Table 4.1: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = \sqrt{x_i}$ with the sample size 20.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	0.00	0.06	0.00	0.33	0.99(1.38)	0.94(2.37)
	TBS	-0.10	0.08	-0.20	0.03	0.92(0.99)	0.99(0.78)
	POM	0.00	0.03	0.00	0.23	0.88(0.59)	0.94(0.59)
	VEXP	0.00	0.04	0.10	0.25	0.97(0.87)	0.95(0.87)
Laplace	OLS	-0.40	0.11	3.10	0.63	0.99(1.88)	0.95(3.27)
	TBS	0.30	0.14	0.70	0.04	0.92(1.36)	0.99(1.07)
	POM	-0.10	0.06	0.80	0.42	0.86(0.81)	0.94(2.49)
	VEXP	-0.40	0.08	2.20	0.48	0.97(1.21)	0.96(2.91)
Location scale	OLS	-0.10	0.07	0.60	0.37	0.99(1.42)	0.94(2.46)
of t	TBS	0.10	0.08	0.60	0.03	0.93(1.03)	0.98(0.81)
	POM	-0.30	0.04	1.90	0.25	0.86(0.81)	0.94(2.49)
	VEXP	-0.30	0.04	2.40	0.28	0.97(1.21)	0.96(2.91)
Logistic	OLS	-0.50	0.19	3.60	1.09	0.99(2.43)	0.95(4.23)
	TBS	0.30	0.24	0.70	0.08	0.93(1.78)	0.99(1.40)
	POM	-0.40	0.10	3.30	0.75	0.86(1.05)	0.94(3.25)
	VEXP	-0.30	0.13	1.10	0.87	0.97(1.54)	0.95(3.79)

The results presented in Table 4.1 show that when the sample size is small with normal distribution, all of the estimation methods produce reasonably accurate estimates. However, under non-normal distributions namely laplace, location scale of t, and logistic distributions, TBS tends to have smaller bias and MSE for β_1 compared to the other methods and POM has the smallest bias and MSE for β_0. On

the other hand, the OLS estimation obviously overestimates the slope parameter, especially under laplace and logistic distributions. In terms of coverage probabilities, the OLS and VEXP methods seem to be better than the others whereas the POM and TBS give the shortest average length for β_0 and β_1, respectively.

Table 4.2: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = \sqrt{x_i}$ with the sample size 50.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	0.20	0.02	-1.90	0.13	0.99(0.82)	0.95(1.41)
	TBS	-0.20	0.03	-0.30	0.01	0.92(0.58)	0.99(0.43)
	POM	0.00	0.01	-0.70	0.07	0.89(0.30)	0.95(1.02)
	VEXP	0.10	0.01	-1.00	0.09	0.97(0.50)	0.97(1.28)
Laplace	OLS	0.00	0.04	-0.70	0.24	0.99(1.14)	0.96(1.98)
	TBS	-0.20	0.06	-0.20	0.01	0.91(0.82)	0.99(0.60)
	POM	0.00	0.01	-1.10	0.14	0.88(0.42)	0.94(1.43)
	VEXP	0.10	0.02	-1.70	0.18	0.98(0.71)	0.97(1.77)
Location scale	OLS	0.00	0.02	0.90	0.14	0.99(0.83)	0.94(1.43)
of t	TBS	0.20	0.03	0.30	0.01	0.92(0.59)	0.99(0.44)
	POM	-0.10	0.01	1.30	0.07	0.89(0.30)	0.94(1.03)
	VEXP	-0.10	0.01	1.50	0.09	0.98(0.51)	0.97(1.29)
Logistic	OLS	-0.30	0.06	0.90	0.38	0.99(1.47)	0.95(2.54)
	TBS	0.00	0.09	0.40	0.02	0.91(1.05)	0.99(0.77)
	POM	-0.20	0.03	0.30	0.23	0.89(0.53)	0.94(1.83)
	VEXP	-0.20	0.04	0.20	0.28	0.99(0.89)	0.96(2.28)

Likewise, when the sample size is 50 as shown in Table 4.2, the results are similar to those in small sample size. However, it is quite obvious in almost all distributions that POM has the smallest bias and MSE for β_0 and TBS has the smallest bias and MSE for β_1.

Table 4.3: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = \sqrt{x_i}$ with the sample size 100.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-0.10	0.01	1.00	0.07	0.99(0.57)	0.94(0.98)
	TBS	0.10	0.01	0.10	0.00	0.92(0.40)	0.99(0.29)
	POM	-0.10	0.00	0.60	0.03	0.91(0.19)	0.94(0.69)
	VEXP	-0.10	0.01	0.50	0.04	0.98(0.35)	0.96(0.89)
Laplace	OLS	0.20	0.02	-2.10	0.11	0.99(0.79)	0.97(1.37)
	TBS	-0.30	0.02	-0.60	0.01	0.93(0.56)	0.99(0.41)
	POM	0.10	0.00	-1.70	0.06	0.93(0.26)	0.97(0.96)
	VEXP	0.10	0.01	-1.70	0.07	0.98(0.49)	0.98(1.24)
Location scale	OLS	0.10	0.01	0.30	0.06	0.99(1.02)	0.96(1.77)
of t	TBS	0.00	0.01	-0.10	0.00	0.92(0.73)	0.99(0.53)
	POM	0.00	0.00	0.20	0.03	0.92(0.19)	0.94(0.70)
	VEXP	0.00	0.00	-0.10	0.04	0.99(0.35)	0.98(0.90)
Logistic	OLS	-0.10	0.03	0.30	0.20	0.99(0.57)	0.94(0.99)
	TBS	0.00	0.04	0.10	0.01	0.92(0.41)	0.99(0.29)
	POM	-0.10	0.01	0.30	0.10	0.92(0.34)	0.95(1.24)
	VEXP	-0.10	0.02	0.70	0.13	0.98(0.63)	0.96(1.60)

Table 4.4: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = \sqrt{x_i}$ with sample size 200.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	0.00	0.01	0.10	0.03	0.99(0.40)	0.96(0.68)
	TBS	0.00	0.01	0.00	0.00	0.92(0.28)	0.99(0.20)
	POM	0.00	0.00	-0.10	0.02	0.92(0.12)	0.95(0.47)
	VEXP	0.00	0.00	0.10	0.02	0.99(0.24)	0.97(0.62)
Laplace	OLS	-0.10	0.01	0.70	0.06	0.99(0.56)	0.95(0.96)
	TBS	0.00	0.01	0.10	0.00	0.92(0.40)	0.99(0.28)
	POM	0.00	0.00	0.20	0.03	0.93(0.17)	0.96(0.66)
	VEXP	-0.10	0.01	0.80	0.04	0.99(0.34)	0.97(0.87)
Location scale	OLS	0.10	0.00	-1.00	0.03	0.99(0.40)	0.96(0.68)
of t	TBS	-0.10	0.01	-0.20	0.00	0.92(0.28)	0.99(0.20)
	POM	0.10	0.00	-0.50	0.01	0.92(0.12)	0.93(0.47)
	VEXP	0.10	0.00	-0.40	0.02	0.98(0.24)	0.97(0.63)
Logistic	OLS	0.10	4.10	-0.10	0.10	0.99(0.71)	0.95(1.24)
	TBS	0.10	0.02	0.00	0.00	0.93(0.51)	0.99(0.36)
	POM	0.00	0.00	0.40	0.05	0.92(0.21)	0.95(0.84)
	VEXP	0.10	0.01	-0.30	0.06	0.98(0.44)	0.98(1.12)

Moreover, when the sample size is large, the results of all distributions carried out in Table 4.3 and Table 4.4 are in agreement with those in Table 4.1. In almost all distributions, POM and VEXP perform best in terms bias and MSE for β_0 and TBS has the smallest bias and MSE for β_1. Again, OLS performs best in terms of coverage probability in maintaining an accuracy of the nominal confidence

level 95%. The bias and MSE of the OLS, POM, and VEXP estimation methods decrease as the sample size increase, though there is a small change of bias and MSE of TBS method.

Table 4.5: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = e^{x_i}$ with the sample size 20.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	1.00	0.49	-8.30	2.53	0.98(3.45)	0.93(6.00)
	TBS	-1.20	0.50	-2.50	0.20	0.91(2.50)	0.97(1.95)
	POM	1.20	0.45	-9.70	2.33	0.80(2.00)	0.88(4.90)
	VEXP	0.90	0.42	-7.50	2.23	0.92(6.05)	0.92(5.48)
Laplace	OLS	-1.00	0.86	6.90	4.64	0.98(4.72)	0.94(8.20)
	TBS	0.60	0.95	1.60	0.38	0.93(3.45)	0.97(2.72)
	POM	-1.00	0.78	7.60	4.09	0.82(2.66)	0.91(6.67)
	VEXP	-0.80	0.73	4.30	4.00	0.92(3.26)	0.93(7.40)
Location scale	OLS	-0.40	0.51	0.40	2.54	0.98(3.60)	0.94(6.24)
of t	TBS	-0.10	0.52	0.90	0.22	0.92(2.62)	0.97(2.07)
	POM	0.00	0.47	-2.10	2.32	0.80(2.09)	0.89(5.14)
	VEXP	0.00	0.43	-2.80	2.23	0.92(2.51)	0.92(5.72)
Logistic	OLS	0.00	1.46	-0.80	7.48	0.98(6.22)	0.95(10.80)
	TBS	-0.50	1.55	-0.80	0.60	0.93(4.50)	0.98(3.53)
	POM	-0.70	1.38	2.60	6.77	0.82(3.58)	0.90(8.79)
	VEXP	0.10	1.27	-2.50	6.58	0.92(4.32)	0.94(9.80)

The results displayed in Table 4.5, where the variance is in exponential form, show that when the sample size is small in all distributions, VEXP provides

the smallest bias and MSE for β_0 whereas TBS provides the smallest bias and MSE for β_1. However, the 95% CIs based on OLS are quite accurate in maintaining the nominal confidence level although it gains the accuracy with wider intervals. Under normal distribution, TBS clearly underestimates for both β_0 and β_1, whereas other estimation methods severely underestimate only for β_1.

Table 4.6: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure $\text{Var}(\epsilon_i) = e^{x_i}$ with the sample size 50.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	0.20	0.17	0.30	0.97	0.99(2.06)	0.92(3.57)
	TBS	0.30	0.18	0.30	0.06	0.92(1.48)	0.97(1.08)
	POM	-0.20	0.17	2.20	0.85	0.85(1.25)	0.88(2.94)
	VEXP	0.00	0.14	1.80	0.77	0.93(1.43)	0.94(3.31)
Laplace	OLS	-0.70	0.31	2.90	1.69	0.98(2.88)	0.94(4.98)
	TBS	0.10	0.35	1.10	0.12	0.91(2.06)	0.97(1.52)
	POM	-0.60	0.29	2.10	1.47	0.87(1.68)	0.89(4.04)
	VEXP	-0.60	0.27	1.70	1.40	0.94(1.97)	0.94(4.55)
Location scale	OLS	0.50	0.15	-2.40	0.80	0.99(2.10)	0.96(3.64)
of t	TBS	0.00	0.17	-0.30	0.05	0.92(1.50)	0.99(1.11)
	POM	0.30	0.15	-1.30	0.74	0.87(1.28)	0.92(2.99)
	VEXP	0.20	0.12	-0.40	0.66	0.95(1.47)	0.96(3.38)
Logistic	OLS	0.00	0.53	-2.30	2.74	0.99(3.70)	0.95(6.42)
	TBS	-0.60	0.59	-0.80	0.18	0.91(2.65)	0.98(1.94)
	POM	0.60	0.51	-5.60	2.46	0.85(2.22)	0.90(5.25)
	VEXP	0.40	0.45	-5.40	2.34	0.95(2.58)	0.95(5.93)

The sample size is 50, the results reported in Table 4.6 is in agreement with those in Table 4.5. However, under non-normal distribution, TBS performs best in terms of bias and MSE for both β_0 and β_1 while OLS, POM, and VEXP underestimate both parameters.

Table 4.7: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure $\text{Var}(\epsilon_i) = e^{x_i}$ with the sample size 100.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-0.30	0.08	2.40	0.44	0.99(1.42)	0.93(2.47)
	TBS	0.30	0.09	0.70	0.03	0.91(1.01)	0.98(0.73)
	POM	-0.20	0.08	2.10	0.41	0.86(0.90)	0.89(2.05)
	VEXP	-0.30	0.07	2.40	0.37	0.94(0.99)	0.95(2.30)
Laplace	OLS	-0.20	0.17	0.60	0.89	0.99(2.01)	0.93(3.48)
	TBS	-0.10	0.17	0.10	0.05	0.91(1.43)	0.97(1.03)
	POM	-0.20	0.15	0.70	0.75	0.86(1.23)	0.89(2.85)
	VEXP	-0.30	0.14	1.50	0.69	0.93(1.40)	0.94(3.22)
Location scale	OLS	0.10	0.44	-0.80	0.08	0.99(1.44)	0.94(2.50)
of t	TBS	-0.10	0.09	-0.30	0.03	0.91(1.02)	0.98(0.75)
	POM	-0.10	0.08	0.40	0.40	0.89(0.93)	0.89(2.08)
	VEXP	0.00	0.06	0.10	0.35	0.95(1.02)	0.94(2.33)
Logistic	OLS	0.00	0.28	0.50	1.45	0.99(2.60)	0.94(4.51)
	TBS	0.00	0.26	-0.20	0.08	0.93(1.85)	0.98(1.34)
	POM	0.10	0.28	-0.10	1.28	0.86(1.63)	0.91(3.73)
	VEXP	0.00	0.22	0.80	1.18	0.94(1.81)	0.95(4.19)

Table 4.8: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = e^{x_i}$ with sample size 200.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-0.10	0.04	0.60	0.23	0.99(1.00)	0.92(1.74)
	TBS	0.00	0.04	0.00	0.01	0.92(0.71)	0.98(0.51)
	POM	0.00	0.04	0.10	0.20	0.87(0.65)	0.88(1.45)
	VEXP	0.00	0.03	0.30	0.17	0.94(0.70)	0.95(1.62)
Laplace	OLS	0.20	0.08	-0.20	0.43	0.99(1.42)	0.94(2.45)
	TBS	0.20	0.08	0.00	0.02	0.93(1.00)	0.98(0.72)
	POM	0.00	0.08	0.90	0.37	0.89(0.89)	0.90(2.03)
	VEXP	0.20	0.06	-0.30	0.33	0.95(0.98)	0.96(2.28)
Location scale	OLS	0.30	0.04	-1.40	0.22	0.99(1.00)	0.94(1.73)
of t	TBS	0.00	0.05	-0.20	0.01	0.92(0.71)	0.98(0.51)
	POM	0.20	0.04	-1.10	0.21	0.88(0.65)	0.88(1.45)
	VEXP	0.20	0.03	-0.90	0.18	0.94(0.69)	0.94(1.62)
Logistic	OLS	-0.40	0.13	2.60	0.73	0.91(1.81)	0.98(3.13)
	TBS	0.20	0.14	0.50	0.04	0.99(1.28)	0.94(0.92)
	POM	-0.10	0.13	1.40	0.63	0.88(1.17)	0.90(2.61)
	VEXP	-0.30	0.10	2.60	0.55	0.95(1.27)	0.95(2.92)

When the sample size increases to the sizes 100 and 200, the results displayed in Table 4.7 and Table 4.8 reveal that the results are similar to those in Table 4.5. The VEXP still provides the smallest bias and MSE for β_0 whereas TBS provides the smallest bias and MSE for β_1. In terms of coverage probabilities, the OLS and VEXP method gain the accuracy with wider intervals in all cases. More-

over, the bias and MSE of each method seems to decrease as the sample size increase to 100 and 200, especially when data is not under normal distribution.

Table 4.9: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var(ϵ_i) $= 0.5 + \hat{y}_i$ with the sample size 20.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-0.70	3.26	5.00	10.74	0.96(7.81)	0.95(13.57)
	TBS	0.90	2.04	2.60	1.17	0.94(5.61)	0.95(4.40)
	POM	-2.60	3.73	15.40	12.39	0.87(6.40)	0.91(12.43)
	VEXP	-1.90	3.61	14.10	12.37	0.91(6.99)	0.93(13.08)
Laplace	OLS	-0.90	5.93	2.90	21.08	0.97(10.80)	0.95(18.91)
	TBS	-0.20	3.99	0.50	2.13	0.95(7.89)	0.96(6.08)
	POM	0.20	6.46	-4.60	22.78	0.87(8.47)	0.92(16.97)
	VEXP	0.30	6.28	-7.40	22.32	0.93(9.50)	0.93(17.89)
Location scale	OLS	2.80	3.37	-10.30	11.40	0.96(8.14)	0.96(14.18)
of t	TBS	0.70	2.10	-1.80	1.17	0.94(5.87)	0.96(4.62)
	POM	2.40	3.83	-6.20	12.82	0.87(6.64)	0.92(13.01)
	VEXP	2.80	3.75	-8.60	12.95	0.93(7.25)	0.93(13.67)
Logistic	OLS	-3.90	10.43	12.50	35.68	0.96(13.78)	0.94(24.11)
	TBS	-1.40	6.80	1.60	3.84	0.93(10.02)	0.95(7.78)
	POM	-4.70	11.58	16.10	38.70	0.87(11.34)	0.91(22.13)
	VEXP	-3.90	11.16	12.10	38.01	0.92(12.38)	0.93(23.16)

Table 4.10: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = 0.5 + \hat{y}_i$ with the sample size 50.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-1.90	1.16	11.10	4.15	0.98(4.62)	0.95(8.04)
	TBS	1.20	0.70	3.80	0.34	0.95(3.32)	0.97(2.41)
	POM	-2.20	1.19	12.90	4.21	0.94(4.10)	0.94(7.64)
	VEXP	-2.30	1.16	14.40	4.20	0.96(4.27)	0.95(7.94)
Laplace	OLS	-0.90	2.24	4.70	7.79	0.96(6.48)	0.95(11.22)
	TBS	0.50	1.40	1.90	0.65	0.95(4.61)	0.96(3.40)
	POM	-0.90	2.20	4.30	7.59	0.92(5.53)	0.93(10.45)
	VEXP	-0.70	2.16	3.30	7.51	0.94(5.91)	0.95(10.97)
Location scale	OLS	1.20	1.27	-6.20	4.20	0.96(4.68)	0.96(8.12)
of t	TBS	0.20	0.73	-0.20	0.36	0.95(3.35)	0.96(2.46)
	POM	1.10	1.33	-6.00	4.31	0.92(4.15)	0.89(7.71)
	VEXP	1.10	1.29	-5.80	4.22	0.94(4.33)	0.95(8.00)
Logistic	OLS	0.40	3.75	-15.60	13.08	0.97(8.33)	0.95(14.49)
	TBS	-4.00	2.40	-5.30	1.08	0.93(5.96)	0.96(4.37)
	POM	1.00	3.87	-19.60	13.44	0.92(7.22)	0.94(13.62)
	VEXP	0.50	3.76	-16.90	13.23	0.95(7.62)	0.95(14.24)

Table 4.11: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = 0.5 + \hat{y}_i$ with the sample size 100.

Distributions	Estimators	β_0		β_1		Coverage Probability (average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	-1.30	0.59	4.30	1.98	0.96(3.20)	0.95(5.55)
	TBS	-0.20	0.35	0.90	0.16	0.94(2.28)	0.96(1.65)
	POM	-1.30	0.61	4.40	2.01	0.93(2.88)	0.93(5.32)
	VEXP	-1.40	0.58	5.00	1.97	0.95(2.96)	0.95(5.50)
Laplace	OLS	1.30	1.26	-1.70	4.37	0.95(4.49)	0.94(7.80)
	TBS	0.70	0.74	-1.00	0.33	0.93(3.20)	0.96(2.30)
	POM	1.60	1.22	-3.50	4.27	0.92(3.95)	0.89(7.38)
	VEXP	1.40	1.22	-2.90	4.22	0.93(4.13)	0.94(7.69)
Location scale	OLS	-0.80	0.60	5.80	2.10	0.96(3.22)	0.95(5.59)
of t	TBS	0.90	0.36	2.30	0.16	0.96(2.30)	0.96(1.65)
	POM	-0.90	0.62	6.30	2.12	0.92(2.90)	0.52(2.68)
	VEXP	-0.90	0.60	6.60	2.08	0.94(2.98)	0.95(5.54)
Logistic	OLS	-3.10	1.96	19.40	6.52	0.95(5.77)	0.95(10.01)
	TBS	1.40	1.14	4.00	0.50	0.95(4.11)	0.96(2.97)
	POM	-3.40	2.02	20.80	6.64	0.92(5.19)	0.94(9.58)
	VEXP	-3.20	1.95	20.10	6.49	0.94(5.37)	0.95(9.91)

Table 4.12: Relative biases, MSEs, and coverage probabilities (average lengths) obtained from the simulation under variance structure Var$(\epsilon_i) = 0.5 + \hat{y}_i$ with sample size 200.

Distributions	Estimators	β_0		β_1		Coverage Probability(average length)	
		Bias	MSE	Bias	MSE	β_0	β_1
Normal	OLS	0.80	0.29	-5.60	0.99	0.97(2.25)	0.95(3.89)
	TBS	-0.80	0.17	-2.20	0.07	0.94(1.59)	0.96(1.14)
	POM	1.00	0.29	-6.90	0.99	0.95(2.06)	0.94(3.76)
	VEXP	0.90	0.29	-6.60	0.99	0.95(2.09)	0.95(3.87)
Laplace	OLS	-0.90	0.53	6.60	1.84	0.97(3.15)	0.96(5.48)
	TBS	0.90	0.33	2.10	0.13	0.95(2.24)	0.97(1.60)
	POM	-1.10	0.53	8.00	1.81	0.96(2.84)	0.95(5.24)
	VEXP	-1.00	0.52	7.80	1.80	0.96(2.91)	0.96(5.42)
Location scale	OLS	-0.60	0.28	5.20	0.97	0.97(2.25)	0.96(3.90)
of t	TBS	0.80	0.18	1.60	0.07	0.95(1.60)	0.97(1.14)
	POM	-0.70	0.28	5.50	0.98	0.94(2.06)	0.95(3.77)
	VEXP	-0.60	0.28	5.40	0.96	0.95(2.09)	0.95(3.88)
Logistic	OLS	0.80	0.96	-4.10	3.21	0.96(4.06)	0.94(7.04)
	TBS	-0.20	0.57	-1.20	0.24	0.93(2.88)	0.97(2.06)
	POM	0.80	0.96	-4.00	3.22	0.93(3.71)	0.93(6.79)
	VEXP	0.80	0.93	-3.60	3.13	0.94(3.78)	0.94(6.99)

The results presented in Table 4.9 and Table 4.10 where the variance is a function of the fitted values reveal that the OLS, POM, and VEXP methods provide severe bias and MSE in all distributions. TBS, although considered to be the best one, still gives small bias estimates. Moreover, TBS also performs well in terms of coverage probabilities and the average lengths of the confidence intervals. When the sample size increases as shown in Table 4.11 and Table 4.12, the results are in agreement with those in small sample size. However, as expected, the severity of bias and MSE of the OLS, POM, and VEXP estimates seem to decrease. This suggest that sample size is crucial for quality of being accurate in estimating parameters, especially under extreme heteroscedasticity.

4.2 Results and discussion of practical applications

The graphs of prediction interval in Figure 4.1 show the performance of each estimation method. Figure 4.2 shows the standardized residuals of each fitted model.

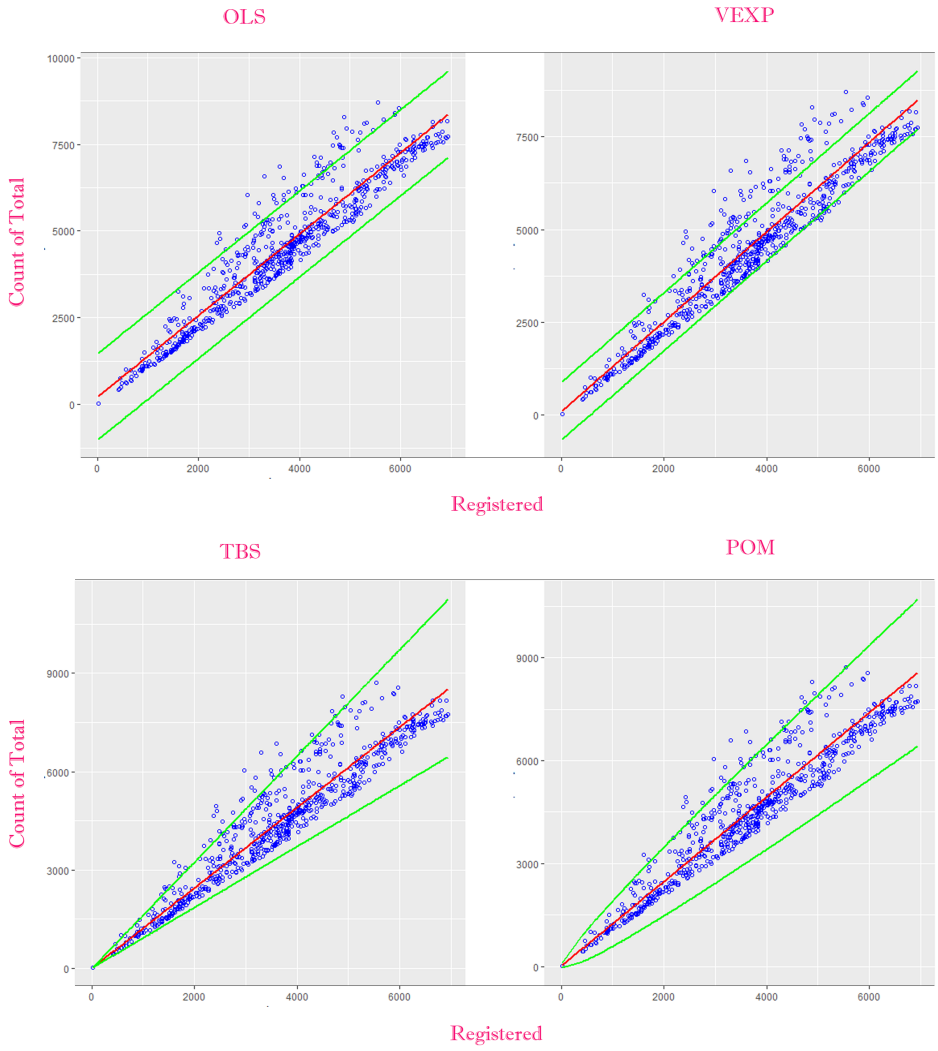

Figure 4.1: Prediction intervals of the model fits on bike rental count daily and registered users with OLS, TBS, POM, and VEXP methods.

Figure 4.1 shows the comparison of model fits with prediction intervals. The OLS and VEXP model obviously overestimate the total number count with small numbers of registered users, although, OLS acquires the accuracy of wider prediction intervals. The TBS and POM method, however, are considered as the best ones among all other fitted models in which most of the observations are within prediction intervals.

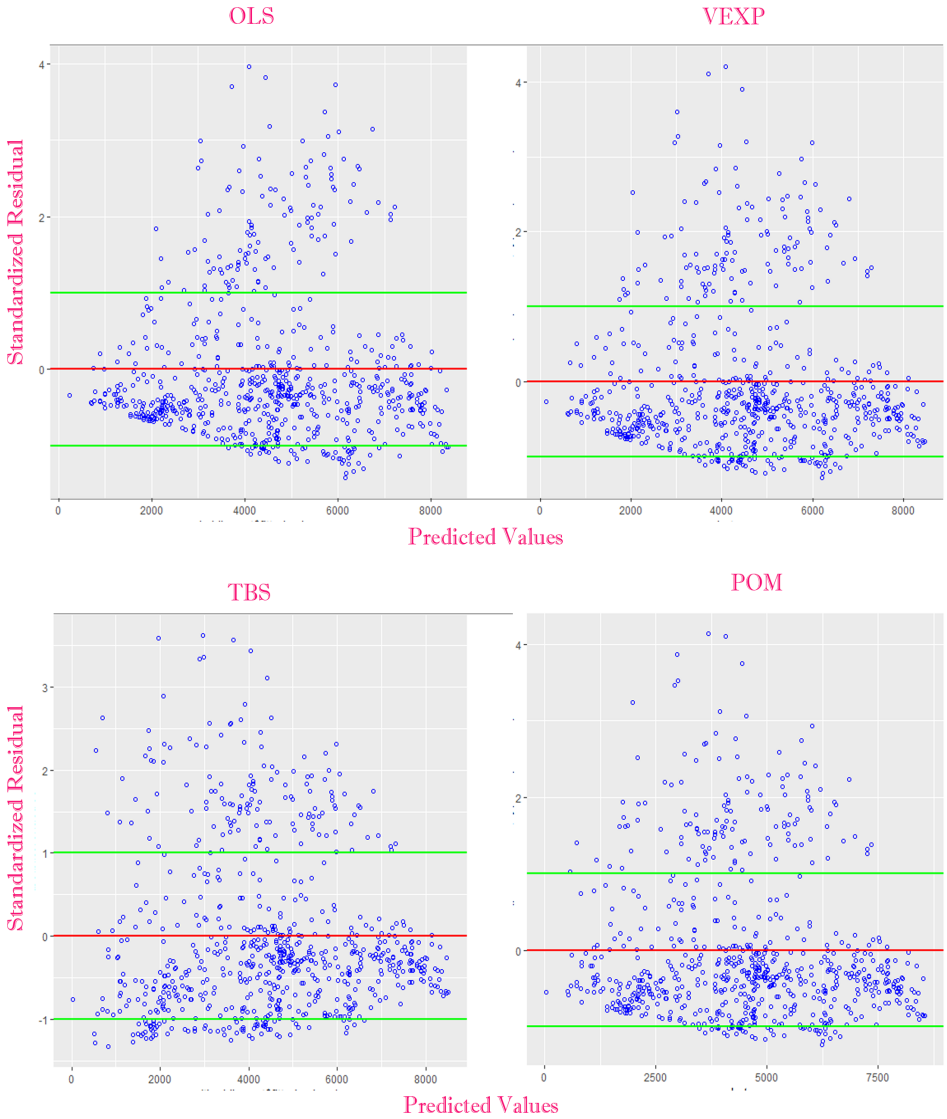

Figure 4.2: Standardized residuals of the models with OLS, TBS, POM, and VEXP.

Figure 4.2 shows that the OLS and VEXP method present the funnel effect with greater error variance corresponding to larger predicted values, which indicate that OLS and VEXP methods do not confirm the assumption of homogeneity of variance. The POM method, although is slightly better, still presents a little funnel effect whereas TBS method does not show the funnel effect and the assumption of homogeneity is verified.

Chapter 5

Conclusions

In this thesis, we reviewed and compared four estimation methods under three different forms of non-constant variances and four symmetric distributions. The assessment of the estimation methods was based on their relative bias, MSEs, coverage probabilities and average lengths. The results given by the simulations study indicate that each estimation method performed differently on different variance structures and different distributions whereas the sample size did not give much effect on each estimation method except in the case of extreme heteroscedasticity.

As expected, under the power of the predictor variance structure with normal distribution, there was not much difference between estimation methods. Under non-normal distributions, however, the POM and TBS methods performed best in terms of bias and MSE whereas the OLS method obviously overestimated the slope parameter. When the variance was in the form of exponential function of the predictor, VEXP and TBS were the best methods to estimate the intercept and slope parameters, respectively. Moreover, the TBS method also performed obviously well under extreme heteroscedasticity such as variance structure of fitted values function whereas the other methods had quite poor performances under this variance structure.

In the real studied data, all discussed estimation methods were in agreement with the performance in the simulation study. The TBS and POM method performed best in terms of model estimates, prediction intervals, and standardized residuals.

In overall, the TBS method seemed to perform best in terms of smallest bias, MSE, prediction intervals, and standardized residuals. On the other hand, the OLS method was very accurate in maintaining the nominal coverage probabilities although it had relatively poor performance in terms of bias. The large sample size had slightly improved for each estimation method when the variance was extremely heteroscedasticity, though it did not much affect on each distribution.

References

Anderson, D., Sweeney, D., Williams, T., Camm, J., & Cochran, J. (2017). *Statistics for business & economics, revised* (11th ed.). Mason, OH 45040, USA: Cengage Learning.

Ávila, J., Salvo, S., & Muñoz, C. (2013). Comparison of linear regression models considering heteroscedasticity of fruits and flower buds of highbush blueberry cultivated in chile. *Scientia horticulturae*, *151*, 57–62.

Äyräväinen, L., Heikkinen, A. M., Kuuliala, A., Ahola, K., Koivuniemi, R., Peltola, J., … Leirisalo-Repo, M. (2018). Activity of rheumatoid arthritis correlates with oral inflammatory burden. *Rheumatology International*, *38*(9), 1661–1669. doi: 10.1007/s00296-018-4108-z

Beal, S., & Sheiner, L. (1988). Heteroscedastic nonlinear regression. *Technometrics*, *30*(3), 327–338. doi: 10.1080/00401706.1988.10488406

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367.

Carroll, R. J., & Ruppert, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, *77*(380), 878–882.

Carroll, R. J., & Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, *79*(386), 321–328.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY 10001 USA: Taylor & Francis.

Clark, K. E., Hartley, S. E., Brennan, R. M., Jennings, S. N., McMenemy, L. S., McNicol, J. W., ... Johnson, S. N. (2012). Effects of cultivar and egg density on a colonizing vine weevil (otiorhynchus sulcatus) population and its impacts on red raspberry growth and yield. *Crop protection*, *32*, 76–82.

Davidian, M., & Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, *82*(400), 1079–1091.

Fanaee T, H., & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 1-15. doi: 10.1007/s13748-013-0040-3

Gebhard, C., Rhainds, D., He, G., Rodés-Cabau, J., Lavi, S., Spence, J. D., ... Grégoire, J. (2018). Elevated level of lecithin: Cholesterol acyltransferase (lcat) is associated with reduced coronary atheroma burden. *Atherosclerosis*, *276*, 131–139.

Giltinan, D. M., Carroll, R. J., & Ruppert, D. (1986). Some new estimation methods for weighted regression when there are possible outliers. *Technometrics*, *28*(3), 219–230.

Giltinan, D. M., & Ruppert, D. (1989). Fitting heteroscedastic regression models to individual pharmacokinetic data using standard statistical software. *Journal of pharmacokinetics and biopharmaceutics*, *17*(5), 601–614.

Hall, I., Aalders, L., & McRae, K. (1982). Lowbush blueberry production in eastern canada as related to certain weather data. *Canadian Journal of Plant Science*, *62*(3), 809–812.

Hao, X., Yujun, S., Xinjie, W., Jin, W., & Yao, F. (2015). Linear mixed-effects models to describe individual tree crown width for China-Fir in Fujian province, southeast China. *PloS one*, *10*(4), e0122257.

Ihalainen, M., & Pukkala, T. (2001). Modelling cowberry (vaccinium vitis-idaea) and bilberry (vaccinium myrtillus) yields from mineral soils and peatlands on the basis of visual field estimates. *Silva Fennica*, *35*(3), 329-340.

Ihalainen, M., Salo, K., & Pukkala, T. (2003). Empirical prediction models for Vaccinium myrtillus and V. vitis-idaea berry yields in north Karelia, Finland. *Silva Fennica*, *37*(1), 95-108.

Jiménez, C. M., & Díaz, J. B. R. (2003a). Statistical model estimates potential yields in pear cultivars 'blanquilla'and 'conference' before bloom. *Journal of the American Society for horticultural science*, *128*(4), 452–457.

Jiménez, C. M., & Díaz, J. B. R. (2003b). A statistical model to estimate potential yields in peach before bloom. *Journal of the American Society for Horticultural Science*, *128*(3), 297–301.

Jiménez, C. M., & Díaz, J. B. R. (2004). Statistical model estimates potential yields in 'golden delicious' and 'royal gala' apples before bloom. *Journal of the American Society for Horticultural Science*, *129*(1), 20–25.

Ketelaere, B. D., Stulens, J., Lammertyn, J., Cuong, N., & Baerdemaeker, J. D. (2006). A methodological approach for the identification and quantification of sources of biological variance in postharvest research. *Postharvest Biology and Technology*, *39*(1), 1 - 9. doi: https://doi.org/10.1016/j.postharvbio.2005.09.004

Kutner, M. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw-Hill Irwin.

Layard, M. (1973). Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, *68*(341), 195–198.

Le Cren, E. (1951). The length-weight relationship and seasonal cycle in gonad weight and condition in the perch (perca fluviatilis). *The Journal of Animal Ecology*, *20*(2), 201–219.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, 278–292.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American statistical Association*, *62*(318), 399–402.

Mendes, M., & Pala, A. (2003). Type I error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, *2*(2), 135–139.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician*, *38*(2), 124–126.

Moore, D., MacCabe, G., & Craig, B. (2009). *Introduction to the practice of statistics*. New York, NY: W.H. Freeman and Company.

Newman, M. C. (1993). Regression analysis of log-transformed data: Statistical bias and its correction. *Environmental Toxicology and Chemistry*, *12*(6), 1129–1133.

Nicholls, S. J., Tuzcu, E. M., Crowe, T., Sipahi, I., Schoenhagen, P., Kapadia, S., … Ntanios, F. (2006). Relationship between cardiovascular risk factors and atherosclerotic disease burden measured by intravascular ultrasound. *Journal of the American College of Cardiology*, *47*(10), 1967–1975.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus* (illustrated, reprint ed.). Springer New York.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Salvo, S., Muñoz, C., Ávila, J., Bustos, J., Ramirez-Valdivia, M., Silva, C., & Vivallo, G. (2012). An estimate of potential blueberry yield using regression models that relate the number of fruits to the number of flower buds and to climatic variables. *Scientia horticulturae*, *133*, 56–63.

Seber, G., & Lee, A. (2012). *Linear regression analysis* (2nd ed.). New Jersey, USA: Wiley.

Shaphiro, S., & Wilk, M. (1965). An analysis of variance test for normality. *Biometrika*, *52*(3), 591–611.

Sidik, K., & Jonkman, J. N. (2016). A comparison of the variance estimation methods for heteroscedastic nonlinear models. *Statistics in medicine*, *35*(26), 4856–4874.

Wackerly, D., Mendenhall, W., & Scheaffer, R. (2014). *Mathematical statistics with applications* (7th ed.). Belmont, CA 94002-3098 USA: Cengage Learning.

Weiers, R. (2010). *Introduction to business statistics* (7th ed.). Mason, OH 45040 USA: Cengage Learning.

Appendix

R simulation codes

7.1 Simulation process of non-constant variance of $0.5 + \hat{y}$ under Logistic distribution with sample size n=200. We just change the sample sizes, variances, and distributions.

```
library(hydroGOF)
library(nlme)
set.seed(3)
tbs.yhat.rlogisb0 <- 0
tbs.yhat.rlogisb1 <- 0
yhat.ols.rlogisb0<- 0
yhat.ols.rlogisb1<- 0
yhat.pom.rlogisb0<- 0
yhat.pom.rlogisb1<- 0
yhat.exp.rlogisb0<-0
yhat.exp.rlogisb1<-0
confb0.tbsrlogis.explow<-0
confb0.tbsrlogis.expup<-0
confb1.tbsrlogis.explow<-0
confb1.tbsrlogis.expup<-0
```

```
confb0.olsrlogis.explow<-0
confb0.olsrlogis.expup<-0
confb1.olsrlogis.explow<-0
confb1.olsrlogis.expup<-0
confb0.powrlogis.explow<-0
confb0.powrlogis.expup<-0
confb1.powrlogis.explow<-0
confb1.powrlogis.expup<-0
confb0.exprlogis.explow<-0
confb0.exprlogis.expup<-0
confb1.exprlogis.explow<-0
confb1.exprlogis.expup<-0
#Loop 1000 rounds to simulate data generate
from Logistic distribution
for(i in 1:1000)
{
x<-runif(200)
a<-3
b<-1
########################################
epshat<- rnorm(200,0,1)
yhat= a +b*x + epshat
fit.yhat<- lm(yhat ~ x)
exvar<- 0.5 + fit.yhat$fitted.values
```

```
#Other non-constant variances
exeps<- rlogis(200,0,exvar) #Switch to other distribution
y<- a + b*x + exeps
#Normal, Laplace, Location scale of t-distributions
########################################
#TBS ================================================
lox<-log(x)
ytbs.yhat.rlogis<- exp(a + b*log(x) + exeps)
loy<- log(ytbs.yhat.rlogis)
tbs.yhat.rlogis<-lm(loy~lox)
tbs.yhat.rlogisb0[i]<- tbs.yhat.rlogis$coefficients[1]
tbs.yhat.rlogisb1[i]<- tbs.yhat.rlogis$coefficients[2]
#CI for TBS==========================================
conf.tbsrnorm.exp<-confint(tbs.yhat.rlogis, level = 0.95)
confb0.tbsrlogis.explow[i]<-conf.tbsrnorm.exp[1,1]
confb0.tbsrlogis.expup[i]<-conf.tbsrnorm.exp[1,2]
confb1.tbsrlogis.explow[i]<-conf.tbsrnorm.exp[2,1]
confb1.tbsrlogis.expup[i]<-conf.tbsrnorm.exp[2,2]
#Fit OLS=============================================
yhat.ols.rlogis<-lm(y~x)
yhat.ols.rlogisb0[i]<-yhat.ols.rlogis$coefficients[1]
yhat.ols.rlogisb1[i]<-yhat.ols.rlogis$coefficients[2]
```

```
#CI for OLS=============================================
conf.olsrlogis.exp<-confint(yhat.ols.rlogis,level = 0.95)
confb0.olsrlogis.explow[i]<-conf.olsrlogis.exp[1,1]
confb0.olsrlogis.expup[i]<-conf.olsrlogis.exp[1,2]
confb1.olsrlogis.explow[i]<-conf.olsrlogis.exp[2,1]
confb1.olsrlogis.expup[i]<-conf.olsrlogis.exp[2,2]
#Fit Power==============================================
yhat.pom.rlogis<-  gls(y~x,weights = varPower(form = ~x))
yhat.pom.rlogisb0[i]<- yhat.pom.rlogis$coefficients[1]
yhat.pom.rlogisb1[i]<- yhat.pom.rlogis$coefficients[2]
#CI for Power===========================================
conf.powrlogis.exp<-intervals(yhat.pom.rlogis,level = 0.95)
confb0.powrlogis.explow[i]<-conf.powrlogis.exp$coef[1,1]
confb0.powrlogis.expup[i]<-conf.powrlogis.exp$coef[1,3]
confb1.powrlogis.explow[i]<-conf.powrlogis.exp$coef[2,1]
confb1.powrlogis.expup[i]<-conf.powrlogis.exp$coef[2,3]
#Fit VarExp=============================================
yhat.exp.rlogis<- gls(y~x, weights = varExp(form = ~x))
yhat.exp.rlogisb0[i]<- yhat.exp.rlogis$coefficients[1]
yhat.exp.rlogisb1[i]<- yhat.exp.rlogis$coefficients[2]
#CI for VarExp==========================================
conf.exprlogis.exp<-intervals(yhat.exp.rlogis,level = 0.95)
confb0.exprlogis.explow[i]<-conf.exprlogis.exp$coef[1,1]
confb0.exprlogis.expup[i]<-conf.exprlogis.exp$coef[1,3]
confb1.exprlogis.explow[i]<-conf.exprlogis.exp$coef[2,1]
confb1.exprlogis.expup[i]<-conf.exprlogis.exp$coef[2,3]
}
```

```
#Coverage for OLS=======================================
Coverageb0.olsrlogis.exp<-ifelse(confb0.olsrlogis.explow<=3
&confb0.olsrlogis.expup>=3,1,0)
t<-table(Coverageb0.olsrlogis.exp)
covb0.olsrlogis.exp<-1-t[1]/1000
COverageb1.olsrlogis.exp<-ifelse(confb1.olsrlogis.explow<=1
&confb1.olsrlogis.expup>=1,1,0)
t<-table(COverageb1.olsrlogis.exp)
covb1.olsrlogis.exp<-1-t[1]/1000
c("CovB0"=covb0.olsrlogis.exp,"CovB1"=covb1.olsrlogis.exp)
#Length CI of OLS=======================================
lengthb0.ols= confb0.olsrlogis.expup-confb0.olsrlogis.explow
lengthb0.ols.logis=sum(lengthb0.ols)/length(lengthb0.ols)
lengthb1.ols=confb1.olsrlogis.expup-confb1.olsrlogis.explow
lengthb1.ols.logis=sum(lengthb1.ols)/length(lengthb1.ols)
c("Length b0"=lengthb0.ols.logis,"Length b1"=lengthb1.ols.logis)
#Coverage for TBS=======================================
coverageb0.tbsrlogis.exp<-ifelse(confb0.tbsrlogis.explow<=3
&confb0.tbsrlogis.expup>=3,1,0)
t<-table(coverageb0.tbsrlogis.exp)
Covb0.tbsrlogis.exp<-1-t[1]/1000
coverageb1.tbsrlogis.exp<-ifelse(confb1.tbsrlogis.explow<=1
&confb1.tbsrlogis.expup>=1,1,0)
t<-table(coverageb1.tbsrlogis.exp)
covb1.tbsrlogis.exp<-1-t[1]/1000
c("CovB0"=Covb0.tbsrlogis.exp,"CovB1"=covb1.tbsrlogis.exp)
```

```
#Length CI of TBS=======================================
lengthb0.tbs=confb0.tbsrlogis.expup-confb0.tbsrlogis.explow
lengthb0.tbs.logis=sum(lengthb0.tbs)/length(lengthb0.tbs)
lengthb1.tbs= confb1.tbsrlogis.expup-confb1.tbsrlogis.explow
lengthb1.tbs.logis=sum(lengthb1.tbs)/length(lengthb1.tbs)
c("Length b0"=lengthb0.tbs.logis,"Length b1 "=lengthb1.tbs.logis)
#Coverage for VarPower==================================
Coverageb0.powrlogis.exp<-ifelse(confb0.powrlogis.explow<=3&
confb0.powrlogis.expup>=3,1,0)
t<-table(Coverageb0.powrlogis.exp)
Covb0.powrlogis.exp<-1-t[1]/1000
Coverageb1.powrlogis.exp<-ifelse(confb1.powrlogis.explow<=1&
confb1.powrlogis.expup>=1,1,0)
t<-table(Coverageb1.powrlogis.exp)
Covb1.powrlogis.exp<-1-t[1]/1000
c("CovB0"=Covb0.powrlogis.exp,"CovB1"=Covb1.powrlogis.exp)
#Length CI of VarPower==================================
lengthb0.varpower= confb0.powrlogis.expup-confb0.powrlogis.explow
lengthb0.varpower.logis=sum(lengthb0.varpower)/
length(lengthb0.varpower)
lengthb1.varpower=confb1.powrlogis.expup-confb1.powrlogis.explow
lengthb1.varpower.logis=sum(lengthb1.varpower)/
length(lengthb1.varpower)
c("Length b0"=lengthb0.varpower.logis,
"Length b1"=lengthb1.varpower.logis)
```

```
#Coverage for VarExp=====================================
Coverageb0.exprlogis.exp<-ifelse(confb0.exprlogis.explow<=3&
confb0.exprlogis.expup>=3,1,0)
t<-table(Coverageb0.exprlogis.exp)
Covb0.exprlogis.exp<-1-t[1]/1000
Coverageb1.exprlogis.exp<-ifelse(confb1.exprlogis.explow<=1&
confb1.exprlogis.expup>=1,1,0)
t<-table(Coverageb1.exprlogis.exp)
Covb1.exprlogis.exp<-1-t[1]/1000
c("COVB0"=Covb0.exprlogis.exp,"COvB1"=Covb1.exprlogis.exp)
#Length CI of varExp=====================================
lengthb0.varexp=confb0.exprlogis.expup-confb0.exprlogis.explow
lengthb0.varexp.logis=sum(lengthb0.varexp)/length(lengthb0.varexp)
lengthb1.varexp=confb1.exprlogis.expup-confb1.exprlogis.explow
lengthb1.varexp.logis= sum(lengthb1.varexp)/
length(lengthb1.varexp)
c("Length b0"=lengthb0.varexp.logis,
"Length b1"=lengthb1.varexp.logis)
#Bias and MSE for TBS====================================
biastbs.yhat.rlogisb0<- pbias(tbs.yhat.rlogisb0,rep(a,1000))
msetbs.yhat.rlogisb0<-mse(tbs.yhat.rlogisb0,rep(a,1000))
biastbs.yhat.rlogisb1<-pbias(tbs.yhat.rlogisb1,rep(b,1000))
msetbs.yhat.rlogisb1<-mse(tbs.yhat.rlogisb1,rep(b,1000))
c("Bias B0"=biastbs.yhat.rlogisb0,"MSE B0"=msetbs.yhat.rlogisb0,
"Bias B1"=biastbs.yhat.rlogisb1,"MSE B1"=msetbs.yhat.rlogisb1)
```

```
#Bias and MSE for OLS==================================
biasols.yhat.rlogisb0<- pbias(yhat.ols.rlogisb0,rep(a,1000))
mseols.yhat.rlogisb0<- mse(yhat.ols.rlogisb0,rep(a,1000))
biasols.yhat.rlogisb1<- pbias(yhat.ols.rlogisb1,rep(b,1000))
mseols.yhat.rlogisb1<- mse(yhat.ols.rlogisb1,rep(b,1000))
c("Bias B0"=biasols.yhat.rlogisb0,"MSE B0"=mseols.yhat.rlogisb0,
"Bias B1"=biasols.yhat.rlogisb1,"MSE B1"=mseols.yhat.rlogisb1)
#Bias and MSE for VarPower==============================
biaspom.yhat.rlogisb0<- pbias(yhat.pom.rlogisb0,rep(a,1000))
msepom.yhat.rlogisb0<- mse(yhat.pom.rlogisb0,rep(a,1000))
biaspom.yhat.rlogisb1<-pbias(yhat.pom.rlogisb1,rep(b,1000))
msepom.yhat.rlogisb1<- mse(yhat.pom.rlogisb1,rep(b,1000))
c("Bias B0"=biaspom.yhat.rlogisb0,"MSE B0"=msepom.yhat.rlogisb0,
"Bias B1"=biaspom.yhat.rlogisb1,"MSE B1"=msepom.yhat.rlogisb1)
#Bias and MSE for VarExp================================
biasexp.yhat.rlogisb0<- pbias(yhat.exp.rlogisb0,rep(a,1000))
mseexp.yhat.rlogisb0<- mse(yhat.exp.rlogisb0,rep(a,1000))
biasexp.yhat.rlogisb1<- pbias(yhat.exp.rlogisb1,rep(b,1000))
mseexp.yhat.rlogisb1<- mse(yhat.exp.rlogisb1,rep(b,1000))
c("Bias B0"=biasexp.yhat.rlogisb0,"MSE B0"=mseexp.yhat.rlogisb0,
"Bias B1"=biasexp.yhat.rlogisb1,"MSE B1"=mseexp.yhat.rlogisb1)
```

7.2 Function for implementing a scatter plot of original data from simulations.

```
het.plot<-function(x,y){
lsfit<- lm(y~x)
data.dat<- data.frame(x,y)
plot.xy<- ggplot(data.dat,aes(x,y))+
geom_point(shape=1, size= 2, color="blue")
plot.xy+geom_line(aes(y=lsfit$fitted.values),
 color = "red",linetype = "solid", size = 1)
}
```

7.3 Code of each fitted model in real life data set.

```
library(MASS)
library(olsrr)
library(nlme)
library(AICcmodavg)
library("ggplot2")
library(Rcmdr)
bikerent<- read.csv("day.csv", header=TRUE)
head(bikerent)
attach(bikerent)
#Test assumption of response data===============
shapiro.test(cnt)
#Bartlett's test for homogeneity
cnt.check<- cnt
summary(cnt.check)
cnt.check1<-ifelse(cnt.check<=summary(cnt.check)[2],1,cnt.check)
cnt.check1<-ifelse(cnt.check>summary(cnt.check)[2]
&cnt.check<=summary(cnt.check)[3],2,cnt.check1)
cnt.check1<-ifelse(cnt.check>summary(cnt.check)[3]
&cnt.check<=summary(cnt.check)[5],3,cnt.check1)
cnt.check1<-ifelse(cnt.check>summary(cnt.check)[5],4,cnt.check1)
bartlett.test(cnt,cnt.check1)
```

```
#==============Start modelling=================
#Fit OLS=======================================
olsx<-registered
olsy<-cnt
ols.bikerent<-lm(olsy~olsx)
new.dat<-data.frame(x=olsx)
head(ols.bikerent$fitted.values)
#Plot Confidence interval======================
conf_interval<- predict(ols.bikerent,new.dat,+
interval="prediction",level = 0.95)
#ggplot PI for OLS=============================
windows()
ols.data<- cbind(bikerent,conf_interval)
ols.ggp<-ggplot(ols.data, aes(olsx, olsy))
+geom_point(shape=1,color="blue")
#Add prediction intervals
ols.ggp+geom_line(aes(y=ols.bikerent$fitted.values), +
color="red", linetype="solid",size=1)+
geom_line(aes(y = conf_interval[,2]), color = "green",
 linetype = "solid",size=1)+
geom_line(aes(y = conf_interval[,3]), color = "green",
linetype = "solid",size=1)
#Ggplot Residual of OLS=========================
ols.bikerent.stres<-rstandard(ols.bikerent)#studres(ols.bikerent)
windows()
olsresid.ggp<-ggplot(ols.data, aes(ols.bikerent$fitted.values,+
ols.bikerent.stres))+geom_point(shape=1,color="blue")
```

```
#Add line to residual
olsresid.ggp+geom_hline(yintercept= -1, color="green",
linetype = "solid", size=1)+
geom_hline(yintercept = 0, color="red",
linetype = "solid",size=1)+
geom_hline(yintercept = 1, color="green",
linetype = "solid",size=1)
#Lavene's Test for Homogeneity OLS================
sv.ols<-ols.bikerent$residuals
summary(sv.ols)
sv.ols1<-ifelse(sv.ols<=summary(sv.ols)[2],1,sv.ols)
sv.ols1<-ifelse(sv.ols>summary(sv.ols)[2]&
summary(sv.ols)[5],2,sv.ols1)
sv.ols1<-ifelse(sv.ols>summary(sv.ols)[3]&
sv.ols<=summary(sv.ols)[5],3,sv.ols1)
sv.ols1<-ifelse(sv.ols>summary(sv.ols)[5],4,sv.ols1)
sv.ols.group<-factor(sv.ols1)
leveneTest(sv.ols,sv.ols.group)
bartlett.test(sv.ols,sv.ols.group)
leveneTest(sv.ols,season)
#Lavene's Test for Homogeneity OLS================
sv.ols.fi<-ols.bikerent$fitted.values
summary(sv.ols.fi)
sv.ols.fi1<-ifelse(sv.ols.fi<=summary(sv.ols.fi)[2],1,sv.ols.fi)
sv.ols.fi1<-ifelse(sv.ols.fi>summary(sv.ols.fi)[2]&
summary(sv.ols.fi)[5],2,sv.ols.fi1)
```

```
sv.ols.fi1<-ifelse(sv.ols.fi>summary(sv.ols.fi)[3]&
sv.ols.fi<=summary(sv.ols.fi)[5],3,sv.ols.fi1)
sv.ols.fi1<-ifelse(sv.ols.fi>summary(sv.ols.fi)[5],4,sv.ols.fi1)
sv.fi.ols.group<-factor(sv.ols.fi1)
leveneTest(sv.ols.fi,sv.fi.ols.group)
bartlett.test(sv.ols.fi,sv.fi.ols.group)
leveneTest(sv.ols,season)
#Shapiro's Test for Normality OLS==================
shapiro.test(sv.ols)
#Fit TBS=========================================
tbsx<-log(registered)
tbsy<-log(cnt)
new.tbs<- data.frame(x=tbsx)
tbs.bikerent<-lm(tbsy~tbsx)
#Plot TBS=========================================
conf_tbs <- predict(tbs.bikerent,new.tbs,interval="prediction",
level=0.95)
#ggplot PI for TBS================================
windows()
tbs.data<- cbind(bikerent,conf_tbs)
tbs.ggp<-ggplot(tbs.data, aes(olsx, olsy))+geom_point(shape=1,
color="blue")
```

```
#Add prediction intervals
tbs.ggp +geom_line(aes(y = exp(conf_tbs[,1])), color= "red",
linetype = "solid",size=1)+
geom_line(aes(y = exp(conf_tbs[,2])), color= "green",
linetype = "solid",size=1)+
geom_line(aes(y = exp(conf_tbs[,3])), color= "green",
linetype = "solid",size=1)
#Ggplot Residual of TBS============================
windows()
tbs.bikerent.stres<-rstandard(tbs.bikerent)#studres(tbs.bikerent)
tbsresid.ggp<-ggplot(tbs.data, aes(exp(tbs.bikerent$fitted.values),
tbs.bikerent.stres))+geom_point(shape=1,color="blue")
#Add line to residual
tbsresid.ggp+ geom_hline(yintercept = -1, color="green",
linetype="solid", size=1)+
geom_hline(yintercept = 0 , color="red",linetype="solid",size=1)+
geom_hline(yintercept = 1 , color="green",linetype="solid", size=1)

#Bartlett's test for Homogeneity TBS==================
sv.tbs<-tbs.bikerent$residuals
summary(sv.tbs)
sv.tbs1<-ifelse(sv.tbs<=summary(sv.tbs)[2],1,sv.tbs)
sv.tbs1<-ifelse(sv.tbs>summary(sv.tbs)[2]&
summary(sv.tbs)[5],2,sv.tbs1)
```

```
sv.tbs1<-ifelse(sv.tbs>summary(sv.tbs)[3]&
sv.tbs<=summary(sv.tbs)[5],3,sv.tbs1)
sv.tbs1<-ifelse(sv.tbs>summary(sv.tbs)[5],4,sv.tbs1)
sv.tbs.group<-factor(sv.tbs1)
bartlett.test(sv.tbs,sv.tbs.group)
#Shapiro's Test for Normality TBS====================
shapiro.test(sv.tbs)
qqnorm(tbs.bikerent.stres)
qqline(tbs.bikerent.stres)
#Fit POM=============================================
pomx<-registered
pomy<-cnt
pomframx<-data.frame(pomx)
n<-length(pomy)
new.pom.dat<-data.frame(pomx)
pom.bikerent<-gls(pomy~pomx,weights=varPower(form = ~pomx))
#Plot PI Davidian and Carroll:=======================
y.hat<- pom.bikerent$fitted
t<- abs(qt(1-0.05/2,n-2))
sigma<- pom.bikerent$sigma
sigma.yhat<-sigma*y.hat^0.6094444
yhat.pom.lower<- y.hat-t*sigma.yhat
yhat.pom.upper<- y.hat+t*sigma.yhat
data.new.fram <- data.frame(y.hat,yhat.pom.lower,yhat.pom.upper)
head(data.new.fram)
```

```
#ggplot For POM=======================================
windows()
pom.mydata<-cbind(bikerent,data.new.fram)
pom.ggp<- ggplot(pom.mydata, aes(pomx,pomy))+
geom_point(shape=1,color="blue")
pom.ggp+  geom_line(aes(y =y.hat), color="red",
linetype= "solid",size=1)+
geom_line(aes(y =yhat.pom.lower), color = "green",
linetype = "solid",size=1)+
geom_line(aes(y =yhat.pom.upper), color = "green",
linetype = "solid",size=1)
#Ggplot Residual of POM================================
windows()
pomresid.ggp<-ggplot(pom.mydata,
aes(y.hat, residuals(pom.bikerent,type="pearson")))+
geom_point(shape=1,color="blue")
#Add line to residual
pomresid.ggp+ geom_hline(yintercept = -1, color="green",
linetype="solid",size=1)+
geom_hline(yintercept =  0, color="red",
linetype="solid",size=1)+
geom_hline(yintercept =  1, color="green",
linetype="solid",size=1)
```

```
#Lavene's Test for Homogeneity POM=======================
sv.pom<-pom.bikerent$residuals
summary(sv.pom)
sv.pom1<-ifelse(sv.pom<=summary(sv.pom)[2],1,sv.pom)
sv.pom1<-ifelse(sv.pom>summary(sv.pom)[2]&summary(sv.pom)[5],
2,sv.pom1)
sv.pom1<-ifelse(sv.pom>summary(sv.pom)[3]&
sv.pom<=summary(sv.pom)[5],3,sv.pom1)
sv.pom1<-ifelse(sv.pom>summary(sv.pom)[5],4,sv.pom1)
sv.pom.group<-factor(sv.pom1)
leveneTest(sv.ols,sv.pom.group)
bartlett.test(sv.pom,sv.pom.group)
#Shapiro's Test for Normality POM=======================
shapiro.test(sv.pom)
qqnorm(pom.bikerent$residuals)
qqline(pom.bikerent$residuals)
#Fit VEXP================================================
vexpx<-registered
vexpy<-cnt
new.vexp.dat<- data.frame(vexpx)
vexp.bikerent<-gls(vexpy~vexpx,weights=varExp(form = ~vexpx))
```

```
#Plot prediction interval================================
y.hat.vexp<-vexp.bikerent$fitted
sigmavexp<- vexp.bikerent$sigma
sigmavexp.yhat<- sigmavexp*(y.hat.vexp^0.000122235)
yhat.vexp.lower<- y.hat.vexp- t*sigmavexp.yhat
yhat.vexp.upper<- y.hat.vexp+ t*sigmavexp.yhat
data.new.vexp<- data.frame(y.hat.vexp,yhat.vexp.lower,
yhat.vexp.upper)
head(data.new.vexp)
#ggplot For VEXP=========================================
windows()
vexp.mydata<-cbind(bikerent,data.new.vexp)
vexp.ggp<- ggplot(vexp.mydata, aes(vexpx,vexpy))
+geom_point(shape=1,color="blue")
pom.ggp+  geom_line(aes(y =y.hat.vexp), color="red",
linetype= "solid",size=1)+
geom_line(aes(y =yhat.vexp.lower), color = "green",
linetype = "solid",size=1)+
geom_line(aes(y =yhat.vexp.upper), color = "green",
linetype = "solid",size=1)
#Ggplot Residual of VEXP==================================
windows()
vexpresid.ggp<-ggplot(vexp.mydata,
aes(y.hat.vexp, residuals(vexp.bikerent,type="pearson")))+
geom_point(shape=1,color="blue")
```

```
#Add line to residual
vexpresid.ggp+  geom_hline(yintercept = -1, color="green",
linetype="solid",size=1)+
geom_hline(yintercept =  0, color="red",
linetype="solid",size=1)+
geom_hline(yintercept =  1, color="green",
linetype="solid",size=1)
#Lavene's Test for Homogeneity VEXP=========================
sv.vexp<-vexp.bikerent$residuals
summary(sv.vexp)
sv.vexp1<-ifelse(sv.vexp<=summary(sv.vexp)[2],1,sv.vexp)
sv.vexp1<-ifelse(sv.vexp>summary(sv.vexp)[2]&
summary(sv.vexp)[5],2,sv.vexp1)
sv.vexp1<-ifelse(sv.vexp>summary(sv.vexp)[3]&
sv.vexp<=summary(sv.vexp)[5],3,sv.vexp1)
sv.vexp1<-ifelse(sv.vexp>summary(sv.vexp)[5],4,sv.vexp1)
sv.vexp.group<-factor(sv.vexp1)
leveneTest(sv.vexp,sv.vexp.group)
bartlett.test(sv.vexp,sv.vexp.group)
#Shapiro's Test for Normality VEXP==========================
shapiro.test(sv.vexp)
```

VITAE

Name Mr. Raksmey Thinh

Student ID 5910220104

Educational Attainment

Degree	Name of Institution	Year of Graduation
Bachelor of Science (Mathematics)	Royal University of Phnom Penh	2015

Scholarship Awards during Enrollment

The Higher Education Research Promotion and Thailand's Education Hub for Southern Region of ASEAN Countries Project of Higher Education Commission, contract no. TEH-AC 099/2016.

List of Conference and Publication

Thinh, R., Samart, K., & Jansakul, N. Linear regression models for heteroscedastic and non-normal of data. *Communications in Statistics-Simulation and Computation*. (*Submitted*).