CHAPTER 1

INTRODUCTION

1.1 Background

A statistical analysis used for analyzing and indicating the relationship between a dependent/response and independent/predictor variables is called regression analysis. It is now presentable a change on one of the variables in correspondence with a change in the other. In addition, the regression model can estimate or predict the value of response variable when knowing the value of predictor variables. A linear or straight line relationship can be written as follow:

$$\underset{\sim}{Y} = X\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}, \tag{1.1}$$

where $\underset{\sim}{Y}$ is an $n \times 1$ vector of dependent/response variable, X is an $n \times (k+1)$ matrix of independent/predictor variables, $\underset{\sim}{\theta}$ is a (k+1)x1 vector of unknown parameters, $\underset{\sim}{\varepsilon}$ is an $n \times 1$ vector of errors where it element, ε_i, is independent identically distributed as normal with zero mean and constant variance. It is also assumed that matrix X is of full rank, i.e., rank(X) is p = k+1 and it is less than the sample size n. If $\underset{\sim}{\varepsilon}$ are $\text{NID}\left(\underset{\sim}{0}, \sigma^2 I_n\right)$, then the least square (LS) estimator of $\underset{\sim}{\theta}$ is the same as the maximum likelihood (ML) estimator and it is $\underset{\sim}{\hat{\theta}} = \left(X'X\right)^{-1}\left(X'\underset{\sim}{Y}\right)$. In addition, under Gauss-Markov assumptions, LS method yields the best linear unbiased estimator (BLUE) of $\underset{\sim}{\theta}$.

Whenever the assumptions are violated, e.g. non-normality, heteroscedasticity (variances of the errors are not constant), and non-linearity, the LS estimator would not be preferable. Another important problem is that the observed data contain outliers. These outliers may have a large effect on the estimated value,

Heteroscedasticity problem in regression analysis might be caused by outliers in $y-direction$ and/or $x-direction$ (Rousseeuw and Leroy, 1987: 3-59). Hence, a better way to estimate the parameters $\underset{\sim}{\theta}$ in the regression model (1.1) are needed.

1.2 Statement of the Problem

The least square method is a mathematical way to make the "magnitude" of random errors as small as possible. Magnitude of errors are measured as the square of n terms of $e_i = y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\theta}}$ and add them up. Then the $\hat{\underset{\sim}{\theta}}$ minimizing the resulting sum of squared errors is called LS estimator. LS approach, a traditional approach to regression analysis, attains the BLUE estimator when Guass-Markov assumptions hold. In practice, it often found that the normal distribution might not be in line with the assumption caused by some observations being quite far from the bulk of the data. These observations are called outliers. These are data points not typically located close to the usual data (Montgomery and Peck, 1982:70). The outliers in this research are considered in the sense of regression outliers. They are the observed data that are distinct from the linear relationship representing most of the data and they can draw a regression line away from the usual data. Nevertheless, they exclude unusual incidents. It is often found in practice that outliers can have a large distorting influence on LS estimate, for example as shown in figure 1.1.

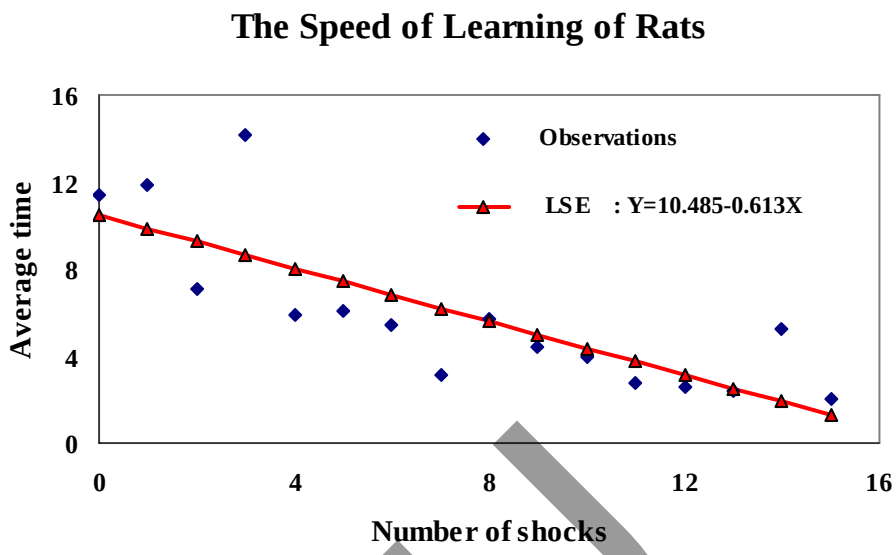

Figure 1.1 LSE fitting with Shock data
Source of Data: Maronna, Martin and Yohai, 2006: 87-88.

To cope with the existent problem of outliers, some may fix the model by deleting (weighted by zero) the outliers but this method can be dangerous as it can give the user a false sense of precision in estimation and prediction. Another two ideas with the different benefit first considered in Mekbunditkul's earlier research (Mekbunditkul, 2010) can be concluded as following: First, Tobit regression is a tool used to investigate the linear relationship when the dependent variable in a regression model is limited. This concept is taken into account for this study in the sense that putting limited value at some desired variable can reduce effect value of outliers in *y*- and *xy-directions*. However, the existence of other types of outliers has not been manipulated. Second, piecewise regression is a regression analysis properly applied when structural change in regression occurs. Hence, in this regression analysis, outliers in *x*- and *xy-directions* are taken into account. However, piecewise is rather not suitable for data consisting of outliers in *y-directions*.

Another approach, termed TP (abbreviated from Tobit-piecewise) regression, employs a fitting criterion to unusual data that is not as contained as LS. An alternative approach, in addition, is used to reduce or down-weigh value of

outliers. This research applied Tobit and piecewise regression to TP regression model (Mekbunditkul, 2010). She constructed the TP regression model by the combination of the Tobit and piecewise regression models. Moreover, there was first found the evidence that the Tobit model (Tobin, 1958, Rosett, 1975 and Jöreskog, 2002), limited by some desired variables could reduce the effect of outliers in some situations in the study of Mekbunditkul. Nevertheless, censoring the data set with one value of either lower or upper limit might not be suitable, thus we need to filter outliers with more than one limiting value. It is dependent on the structure of the whole data. According to the piecewise regression model (Quandt, 1958: 874, Hudson, 1966: 1097-1129, Goldfeld, Kelejian and Quandt, 1971, Suits, Mason and Chan 1978: 132-133), for instance, one data set is fit with two regression regimes when a single regression is inadequate; the structural change is then taken into account as they should be. This structural change in the meaning of regression analysis is a change in one or more of the parameters in a regression model (Poon, et.al. 2008).

Moreover, according to the evidence in simulation results of Mekbunditkul's dissertation, we found that: TP regression model can reduce the effect value of outliers and can utilize the data with structural change more effectively than piecewise, Tobit and LS. Nevertheless, there was not any study regarding an estimation of joined point in TP regression model. Therefore, in this research, the matter is studied. Considering figure 1.1, three obvious outlier data affect the LS regression drawn away from the bulk of the data. This means that the LS regression might not be preferable for the particular case. Whilst these data are analyzed by both Tobit and piecewise regression models, they yield better results than LS regression. Moreover, we can see that TP regression model yields the best among all four different regression models. Therefore, instead of using LS, Tobit or piecewise, we use TP regression model to fit the data consisting of outliers in the sense that TP regression model can reduce the effect of outliers better than Tobit and piecewise.

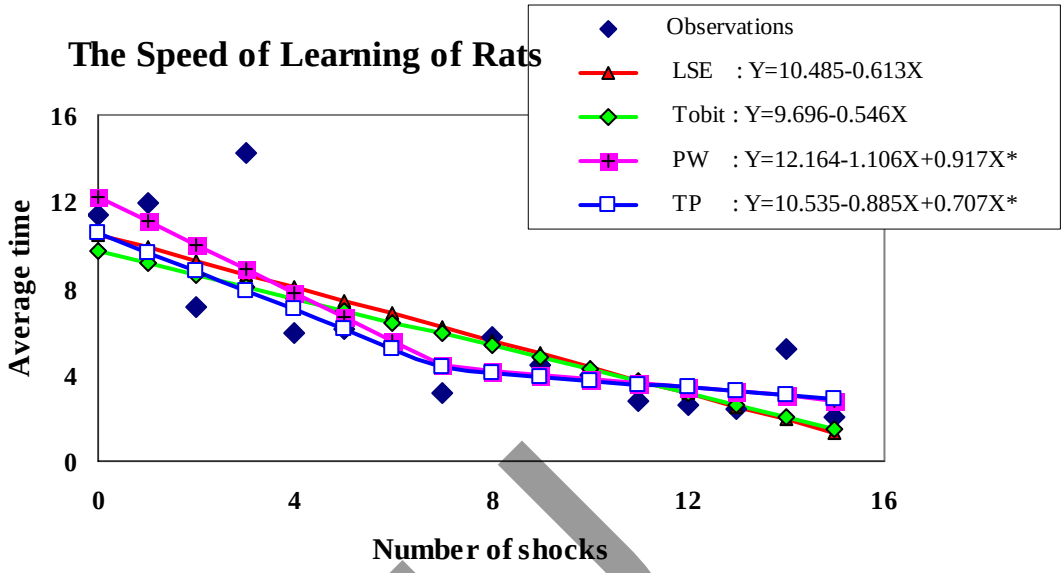

Figure 1.2 Four different regression models fitting with Shock data
Source of Data: Maronna, Martin and Yohai, 2006: 87-88.

Figure 1.2, in this particular case, shows that TP regression model seemingly results better than any other. Therefore, TP regression model is an alternative robust method for those situations where outliers exist. Such "belief" is strongly supported by the data set in the figure above. Nevertheless, there has been little to none literature that describes the estimation method for the joined point in TP regression model. As a result, this point is now first being focused throughout the research.

1.3 Objectives of the Study

To summarize, objectives of the study are as followed:

1) To estimate the joined point in TP regression model.

2) To apply the TP regression model with the real data, for example, socio-economic survey data. This research on the joined point in TP regression model can be estimated by two approaches such as ML based, Quandt's method, and LS based, Levenberg-Marquardt method,

3) To compare two estimation methods of joined point such as ML based and nonlinear LS based. The comparison will be done by both simulation and empirical data analysis.

1.4 Scope of the Study

The TP regression model achieved from the combination of two principal ideas, namely 1) Tobit regression, and 2) piecewise regression. Two estimation methods for the joined point in TP regression model are compared by means of the average sum of square residual (ASSR).

This study is restricted by the following statements:

1. The regression model (1.1)
2. $\underset{\sim}{\varepsilon}$ is normally distributed with the zero mean vector and covariance matrix Σ, where Σ is a diagonal matrix which is positive definite
3. The estimators of regression coefficients in TP regression model are derived by using maximum likelihood estimation (Mekbunditkul, 2010)
4. Two estimation methods for the joined point in TP regression model can be compared the performance by ASSR of regression model. Performances of four different estimators are compared by using the Monte-Carlo simulation. Moreover, these method will also applied to the survey data from the socio-economic survey in Thailand. The four different estimators are LS, Tobit, piecewise and TP.
5. ASSR of regression model is defined as follow.

$$\text{ASSR}_j = \left(d_j + \sqrt{\frac{1}{2} d_j^2 + 2f_j} \right)^2, \qquad j = a, b,$$

where $d_j = \dfrac{\underset{\sim}{1}'\left(\hat{\underset{\sim}{Y}}_{j1} - \underset{\sim}{L}_{j1}\right)}{n_{j2}}$, and $f_j = \dfrac{\left(\hat{\underset{\sim}{Y}}_{j1} - \underset{\sim}{L}_{j1}\right)'\left(\hat{\underset{\sim}{Y}}_{j1} - \underset{\sim}{L}_{j1}\right)}{n_{j2}} + \dfrac{\left(\underset{\sim}{Y}_{j2} - \hat{\underset{\sim}{Y}}_{j2}\right)'\left(\underset{\sim}{Y}_{j2} - \hat{\underset{\sim}{Y}}_{j2}\right)}{n_{j2}}$

(Mekbunditkul, 2010).