Tobit-Piecewise Estimator of the Regression Coefficient

Titirut Mekbunditkul* , Pachitjanut Siripanich
School of Applied Statistics, National Institute of Development Administration.
***EMAIL: tple@hotmail.com**

Abstract: In this paper, an alternative estimator of the regression coefficient, called TP estimator, is proposed based on the idea of Tobit and piecewise regression, in order to fit a data with outliers. A suitable likelihood function is derived for desired conditions so that the TP estimator is the maximum likelihood estimator (MLE). It is found that the regression line obtained by the proposed method is preferable to the least square (LS) and others method as shown by various examples.

Keywords: Outliers, Tobit regression model, Piecewise regression model, MLE

I. Introduction

An important problem usually found in regression analysis is that data are from two or more contaminated distributions and hence outliers are appeared. These cause a regression line away from most of the data points. To cope this problem, some may fit the model by deleting the outliers or "down-weight" the outliers. Many robust regression methods were introduced by various authors. This paper introduces an alternative method called TP, abbreviated from Tobit-piecewise. Consider the piecewise regression first proposed by Quandt [4], for instance, fit one data set, by two regression lines in stead of a single one so that two distributions of error are taken into account as they should be. In addition, based on Tobit regression introduced by Tobin [9], putting limited value for outliers is one procedure of "down-weighted" value (reduce effect) of outliers at the inner fences of the data (Hyndman [1]).

Consider a scatter diagram of a data set in Figure 1, the points in a circle are obviously outliers. When this data is fitted by LS, the outliers have strong influential on regression line so that it does not represent the bulk of the data meanwhile we fit the data by either Tobit or piecewise regression that is better performance than the ordinary regression. In stead, if we combine the Tobit and piecewise modeling ideas, called the TP model described in the next section, we will get the TP regression that yields the best fitting among 4 regression lines for this data.

Hence, the TP method is an alternative robust method for situation when error are from two distributions of difference various and/or outliers exist. A data set in Figure 1, for example, supports such "belief".

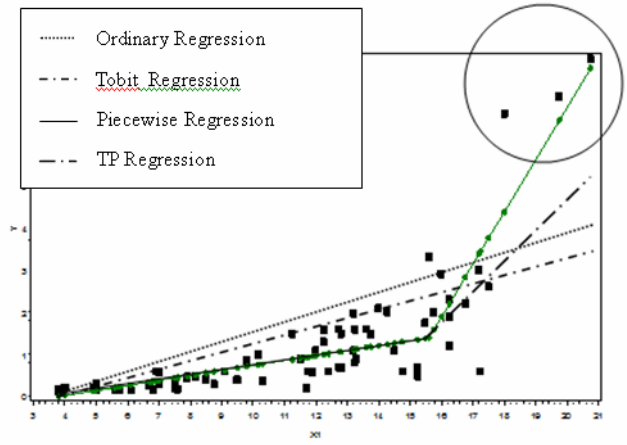

Figure 1. The comparison of 4 regressions fitting

II. Fundamental Notions

Basic Idea Related to the Tobit Regression Model

Tobin [9] introduced the Tobit regression to fit a linear relationship by putting limit on values of some variables. As explained by Tobin, there are some phenomena where the dependent variable being the total durable goods expenditure has some observed data take value as zero during time of survey. This zero value does not mean that the observation has never bought the durable goods but only during the time of survey he has not. In another case, the dependent variable, the students' monthly expenditure, has some observations of higher expenditure than the rest because they are rich persons. Consequently, this zero value or the higher expenditure is assumed to be the lower or upper limit, respectively, in Tobit regression analysis. In this paper, we assign an upper or lower value to the data point in order to limit value of outliers instead of ignoring (weighted by zero) them and then construct an alternative estimator of regression coefficient. In particularly, bounds are the locally inner fences.

Consider a two-limit Tobit model (Tobin [9], Rosett [5] and Jöreskog [2]) which is a relationship of variables Y and Y^* as shown in the model (1). The observed variable Y satisfies

$$Y_i = \begin{cases} L & ; \quad Y_i^* \le L \\ Y_i^* & ; \ L < Y_i^* < U \\ U & ; \quad Y_i^* \ge U, \end{cases} \tag{1}$$

where $Y_i^* = \alpha + \beta x_i + \varepsilon_i$, x_i is a regressor variable, and $\varepsilon_i \sim N\left(0, \sigma^2\right)$. The lower and upper limits are respectively denoted by L and U.

Basic Idea Related to the Piecewise Regression Model

The piecewise regression proposed by Quandt [4] was continuously developed by many researchers. In 1978, Suits [7] introduced the piecewise regression model written as in the multiple regression model (2) with a dummy variable, D_i,

$$Y_i = \alpha_1 + \beta_1 x_i + \beta_2 \left(x_i - \delta\right) D_i + \varepsilon_i \tag{2}$$

where $x_i^* = \left(x_i - \delta\right) D_i$, $D_i = \begin{cases} 0 & ; x_i \le \delta \\ 1 & ; x_i > \delta \end{cases}$, δ is an unknown joined point of two regression lines, and ε_i 's are i.i.d. $N\left(0, \sigma^2\right)$. The coefficient parameters in the model (2) can be estimated by the traditional way, e.g. the LS or ML method.

TP Estimator or Regression Coefficient

The combination of the Tobit (1) and piecewise (2) regression model to be the TP model is shown in the model (3):

$$Y_i = \begin{cases} L_0 & ; \quad Y_i^* \le L_0 \\ Y_i^* & ; \ L_0 < Y_i^* < U_0 \\ U_0 & ; \quad Y_i^* \ge U_0, \end{cases} \tag{3}$$

where $Y_i^* = \alpha_1 + \beta_1 x_i + \beta_2 x_i^* + \varepsilon$, the regressor variables are x_i and x , $x = (x - \delta)D$, δ is an unknown joined point of two regression lines, and ε_i 's are i.i.d. $N\left(0, \sigma_i^2\right)$. Note $\sigma_i^2 = \begin{cases} \sigma_a^2 & \text{if } x_i \le \delta \\ \sigma_b^2 & \text{if } x_i > \delta \end{cases}$. The locally lower and upper limits are $L_0 = \begin{cases} L_a & ; x_i \le \delta \\ L_b & ; x_i > \delta \end{cases}$, and $U_0 = \begin{cases} U_a & ; x_i \le \delta \\ U_b & ; x_i > \delta \end{cases}$. The probability density function (p.d.f.) of Y is determined by

$$f_Y\left(y_i\right) = \Phi\left(\frac{L - \alpha - \underset{\sim}{x}_i \underset{\sim}{\beta}}{\sigma}\right) \text{ if } \quad y_i = L, \quad f_Y\left(y_i\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - \alpha - \underset{\sim}{x}_i \underset{\sim}{\beta}}{\sigma}\right) \text{ if } y_i > L, \text{ and } f_Y\left(y_i\right) = 0 \text{ otherwise}$$

Note $n = \sum_{j=1}^{3} n_j$. The TP estimator of $\underset{\sim}{\theta}$ can be can be achieved by the ML method when the log-likelihood function of $\underset{\sim}{\theta} = \left(\alpha_1, \beta_1, \beta_2; L_0, U_0, \delta_0, \sigma\right)'$ given $\underset{\sim}{Y}$ for some fixed values of L_0, U_0, $\delta = \delta_0$, and σ^2 known can be written as

$$\ln L\left(\underset{\sim}{\theta}; \underset{\sim}{Y}\right) = \sum_{i=1}^{n_1}\left\{\ln \Phi\left(\frac{L_{0i} - \alpha_1 - \beta_1 x_i - \beta_2 x_i^*}{\sigma_i}\right)\right\} + \sum_{j=1}^{n_2}\left\{\ln\left(\frac{1}{\sigma_j}\phi\left(\frac{y_j - \alpha_1 - \beta_1 x_j - \beta_2 x_j^*}{\sigma_j}\right)\right)\right\} + \sum_{k=1}^{n_3}\left\{\ln\left(1 - \Phi\left(\frac{U_{0k} - \alpha_1 - \beta_1 x_k - \beta_2 x_k^*}{\sigma_k}\right)\right)\right\}$$

Let $\lambda_j = \dfrac{y_j - \alpha_1 - \beta_1 x_j - \beta_2 x_j^*}{\sigma_j}$; $L_{0j} < y_j < U_{0j}$

$$\lambda_i = \frac{L_{0i} - \alpha_1 - \beta_1 x_i - \beta_2 x_i^*}{\sigma_i} \ ; \ y_i = L_{0i} \text{ and,}$$

$$\lambda_k = \frac{U_{0k} - \alpha_1 - \beta_1 x_k - \beta_2 x_k^*}{\sigma_k} \ ; \ y_k = U_{0k}.$$

We can get the TP estimator by solving the solution of equations (4a) to (4c)

$$\frac{\partial \ln L\left(\underset{\sim}{\theta}; \underset{\sim}{y}\right)}{\partial \alpha_1} = \sum_{i=1}^{n_1}\left\{-\frac{\phi\left(\hat{\lambda}_i\right)}{\sigma_i \Phi\left(\hat{\lambda}_i\right)}\right\} + \sum_{j=1}^{n_2}\left\{\frac{\hat{\lambda}_j}{\sigma_j}\right\} + \sum_{k=1}^{n_3}\left\{\frac{\phi\left(\hat{\lambda}_k\right)}{\sigma_k\left(1 - \Phi\left(\hat{\lambda}_k\right)\right)}\right\} = 0 \tag{4a}$$

$$\frac{\partial \ln L\left(\underset{\sim}{\theta}; \underset{\sim}{y}\right)}{\partial \beta_1} = \sum_{i=1}^{n_1}\left\{-\frac{x_i \phi\left(\hat{\lambda}_i\right)}{\sigma_i \Phi\left(\hat{\lambda}_i\right)}\right\} + \sum_{j=1}^{n_2}\left\{\frac{x_j \hat{\lambda}_j}{\sigma_j}\right\} + \sum_{k=1}^{n_3}\left\{\frac{x_k \phi\left(\hat{\lambda}_k\right)}{\sigma_k\left(1 - \Phi\left(\hat{\lambda}_k\right)\right)}\right\} = 0 \tag{4b}$$

$$\frac{\partial \ln L\left(\underset{\sim}{\theta};\underset{\sim}{y}\right)}{\partial \beta_2}=\sum_{i=1}^{n_1}\left\{-\frac{x_i^{*}\phi\left(\hat{\lambda}_i\right)}{\sigma_i\,\Phi\left(\hat{\lambda}_i\right)}\right\}+\sum_{j=1}^{n_2}\left\{\frac{x_j^{*}\hat{\lambda}_j}{\sigma_j}\right\}+\sum_{k=1}^{n_3}\left\{\frac{x_k^{*}\phi\left(\hat{\lambda}_k\right)}{\sigma_k\left(1-\Phi\left(\hat{\lambda}_k\right)\right)}\right\}=0 \quad (4c)$$

Where $\hat{\lambda}$ is an estimator of λ. And now we let $\underset{\sim}{Y}_1=\left(Y_{11},...,Y_{1n_1}\right)'$ where $Y_{1i}=L_{0i}$, $\underset{\sim}{Y}_2=\left(Y_{21},...,Y_{2n_2}\right)'$ where $L_{0j}<Y_{2j}<U_{0j}$, and $\underset{\sim}{Y}_3=\left(Y_{31},...,Y_{3n_3}\right)'$ where $Y_{3k}=U_{0k}$. The regressor matrices are X_1, X_2 and X_3 corresponding to $\underset{\sim}{Y}_1$, $\underset{\sim}{Y}_2$ and $\underset{\sim}{Y}_3$. And Σ_1, Σ_2, Σ_3 are denoted the covariance matrices of $\underset{\sim}{Y}_1$, $\underset{\sim}{Y}_2$ and $\underset{\sim}{Y}_3$., respectively. We can obtain the vector of TP estimator with the ordinary least squares estimator of non-limit observations served as the initial estimator by

$$\begin{aligned}\hat{\underset{\sim}{\theta}}_{TP}=&\left(X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_2'\Sigma_2^{-1}\underset{\sim}{Y}_2\right)\\&-\left(X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_1'\Sigma_1^{-1/2}\left[\underset{\sim}{\bar{H}}_1\left(\hat{\lambda}\right)\right]\right)\\&+\left(X_2'\Sigma_2^{-1}X_2\right)^{-1}\left(X_3'\Sigma_3^{-1/2}\left[\underset{\sim}{H}_3\left(\hat{\lambda}\right)\right]\right),\end{aligned} \quad (6)$$

where

$$\underset{\sim}{\bar{H}}_1\left(\hat{\lambda}\right)=\left(\frac{\phi\left(\frac{L_{01}-\underset{\sim}{x}_{11}\hat{\underset{\sim}{\theta}}_{TP}}{\sigma_{11}}\right)}{\Phi\left(\frac{L_{01}-\underset{\sim}{x}_{11}\hat{\underset{\sim}{\theta}}_{TP}}{\sigma_{11}}\right)}\quad\cdots\quad\frac{\phi\left(\frac{L_{0n_1}-\underset{\sim}{x}_{1n_1}\hat{\underset{\sim}{\theta}}_{TP}}{\sigma_{1n_1}}\right)}{\Phi\left(\frac{L_{0n_1}-\underset{\sim}{x}_{1n_1}\hat{\underset{\sim}{\theta}}_{TP}}{\sigma_{1n_1}}\right)}\right)'_{1\times n_1}$$

, $\underset{\sim}{H}_3\left(\hat{\lambda}\right)=\left(\frac{\phi\left(\hat{\lambda}_1\right)}{1-\Phi\left(\hat{\lambda}_1\right)}\quad\cdots\quad\frac{\phi\left(\hat{\lambda}_{n_3}\right)}{1-\Phi\left(\hat{\lambda}_{n_3}\right)}\right)'_{1\times n_3}$, Φ and ϕ are the cumulative density function (c.d.f.) and the probability density function (p.d.f.) of the standard normal distribution, respectively.

This estimation process is required the Newton's method to solve the solution of the equation (6). There is a proof of the biasedness of TP estimator (Titirut [8]), although, the regression line obtained by the TP estimator may fit the data with outliers better than the traditional estimator illustrated in the next section.

III. Numerical Examples

The performance of TP regression line was considered in sense of 1) the relative efficiency (R.E.), the ratio of MSE of regression line obtained by the LSE to Tobit, piecewise and the TP estimator, 2) the fitting of regression exhibited in Figure 2(a) to 2(d)**.** Data set 1, the data for the empirical test of Quandt [4], is no outlier. The LS and Tobit regression lines are the same while the piecewise and TP regression lines are the same and slightly different from the LS or Tobit with R.E. as only 1.05.

In case of there are outliers in y such as data sets 2 to 4, are the original stock exchanges from Mcgee [3], brain and body weights of 28 animals from Weiskerg [10] and discharge and bedload transport data from Ryan [6], respectively. They can be seen that the TP regression line is less affected by outliers than others with R.E. is between 1.66 and 2.

Table 1 The MSE and Relative Efficiency (R.E.) of 4 regression lines.

Data set		Estimator			
		LS	**Tobit**	**Piecewise**	**TP**
1	MSE	1.507	1.507	1.431	1.431
	R.E.	1.00	1.00	**1.05**	**1.05**
2	MSE	1320.58	1030.56	1318.76	493.47
	R.E.	1.00	1.28	1.00	**1.66**
3	MSE	0.411	0.411	0.354	0.207
	R.E.	1.00	1.00	1.16	**1.98**
4	MSE	0.0301	0.0185	0.0234	0.0147
	R.E.	1.00	1.63	1.29	**2.05**

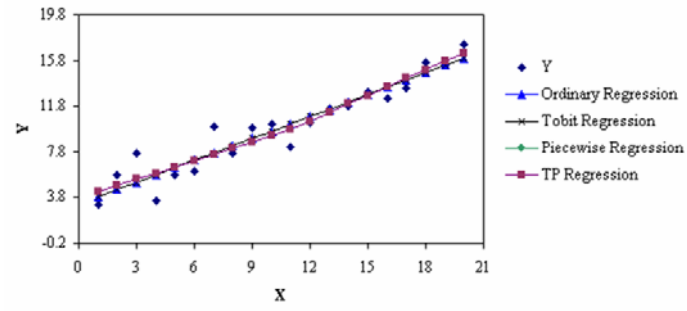

Figure 2(a). The comparison of 4 regression lines for the data set 1.

Source: Quandt [4]

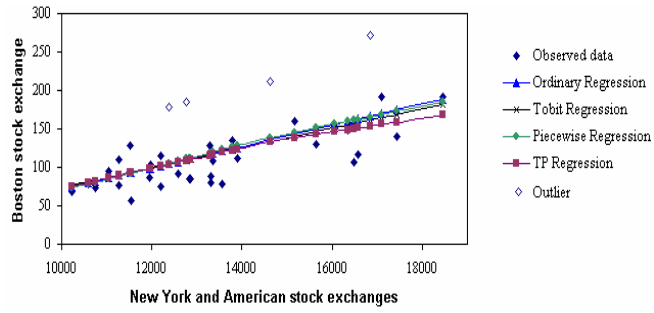

Figure 2(b). The 4 regression lines for the data set 2.

Source : Mcgee [3]

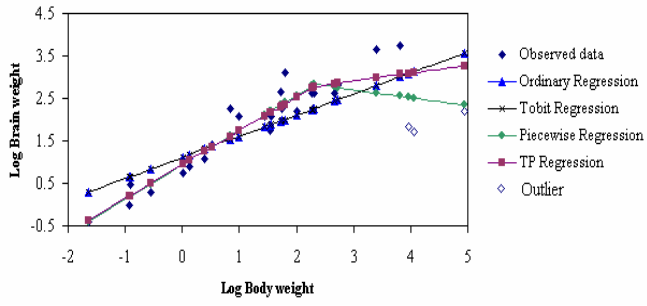

Figure 2(c). The 4 regression lines for the data set 3. Source : Weiskerg [10]

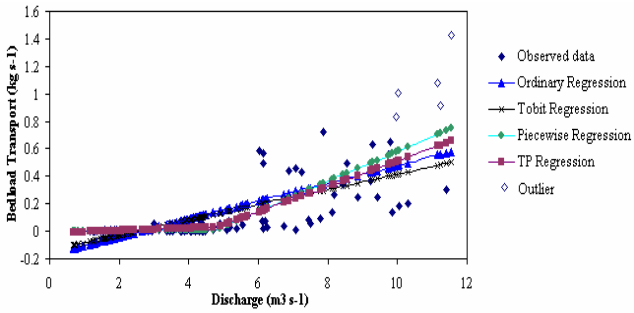

Figure 2(d). The 4 regression lines for the data set 4. Source : Ryan [6]

IV. Conclusions

In this paper, the likelihood function is verified by the combination of Tobit and piecewise models to be the TP model which is an alternative robust method for situation when error are from two distributions of difference various and/or outliers exist. The TP estimator is one of MLE although it is a biased estimator of a regression coefficient it is a consistent estimator and keeps all properties of MLE. The applying of TP regression to cope with the data consisting of outliers has not been found in literatures. The performances of estimator should be thoroughly studied in the future. Moreover, this finding can also be applied in the supply chain research when data have outliers.

V. References

[1] Hyndman, R.J. et al. 1996. Sample Quantiles in Statistical Packages. **The American Statistician**. 50 : 361-365.

[2] Jöreskog, K.G. 2002. Censored Variables and Censored Regression. Available at http://www.ssicentral.com/lisrel/techdocs/censor.pdf.

[3] Mcgee, V.E. et al. 1970. Piecewise Regression. **Journal of the American Statistical Association.** 65 : 1109-1124.

[4] Quandt, R.E. 1958. The estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. **Journal of the American Statistical Association**. 53 : 873-880.

[5] Rosett, R. et al. 1975. Estimation of the two-limit Probit regression model. **Econometrica.** 43 : 141–146.

[6] Ryan, S.E. et al. 2007. **A tutorial on the piecewise regression approach applied to bedload transport data.** Gen. Tech. Rep. RMRS-GTR-189.

[7] Suits, D. et al. 1978. Spline Functions Fitted by Standard Regression Methods. **Review of Economics and Statistics.** 132 – 139.

[8] Titirut, M. 2010. **An Alternative Estimator of Regression Coefficient with Outliers.** In processing.

[9] Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. **Econometrica.** 26 : 24-36.

[10] Weiskerg, S. 1980. **Applied Linear Regression.** New York : John Wiley&Sons,Inc.