Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 15

11-1-2005

Estimating The Slope Of Simple Linear Regression In The Presence Of Outliers

Mohammed Al-Haj Ebrahem
Yarmouk University, Irbid, Jordan, m_hassanb@hotmail.com

Amjad D. Al-Nasser
Yarmouk University, Irbid, Jordan, amjadn@yu.edu.jo

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No.2, 509-513

1538 – 9472/05/$95.00

Estimating The Slope Of Simple Linear Regression In The Presence Of Outliers

Mohammed Al-Haj Ebrahem Amjad D. Al-Nasser
Department of Statistics, Faculty of Science, Yarmouk University
Irbid, Jordan

In this article, an estimation procedure to simple linear regression in the presence of outliers is proposed. The performance of the proposed estimator, the AM estimator, is compared with other traditional estimators: least squares, Theil type repeated median, and geometric mean. A numerical example is given to illustrate the proposed estimator. Simulation results indicate that the proposed estimator is accurate and has a high precision in the presence of outliers.

Key words: Least squares, geometric mean, Theil-type estimators, simple linear regression, outliers

Introduction

Regression analysis was first developed by Sir Francis Galton in the later part of the 19th century. Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to revert or regress to the mean of the group. Galton developed a mathematical description of this tendency, the precursor of today's regression models (Neter, et. al., 1996).

Consider the simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \text{ , i = 1,2,…,n} \qquad (1)$$

where y_i is the response variable in the ith trial, α (intercept) and β(slope) are parameters. X_i is a known constant, namely; the value of the predictor variable in the ith trial. ε_i is a random error term with mean zero and variance σ^2.

Mohammed Al-Haj is an Assistant Professor in the Department of Statistics. His research interests are is in reliability, accelerated life testing, and non-parametric regression models. E-mail: m_hassanb@hotmail.com. Email Amjad D. Al-Nasser at amjadn@yu.edu.jo.

Most of the methods used in the literature to estimate the model parameters are based on the normality assumption. However, in some situations it is unreliable to use the normality assumption to identify the model; instead one may use non-parametric estimation approach. Moreover, if the data contains outlier observations, then robust methods are needed to polish the effect of the outliers. More details can be found in Montgomery and Peck (1992), Rousseeuw and Leroy (1987), Davies (1993), Fernandez (1997), and Olive (2005). A new non-parametric procedure is proposed in order to estimate the slope of model (1).

Estimation Methods for Simple Linear Regression Model

The various estimators that have been suggested for the slope are as follows:

(1) Method of Least Squares (LS)

The least square criterion requires that one consider the sum of n squared deviations; this criterion is denoted by Q

$$Q = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2$$

According to the method of least squares, the estimates of α (intercept) and β(slope) are those values $\hat{\alpha}_{ls}$, $\hat{\beta}_{ls}$ respectively, that minimize the criterion Q for the given sample observations

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, using the analytical approach it can be shown that the estimate values of α (intercept) and β(slope) are

$$\hat{\beta}_{ls} = \frac{Sxy}{Sxx} \text{ and } \hat{\alpha}_{ls} = \bar{y} - \hat{\beta}_{ls}\bar{x}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \ ,$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Note that $\hat{\beta}_{ls}$ is unbiased estimator of β. However, regression outliers (either in x or in y) pose a serious threat to least squares analysis.

(2) The Geometric Mean Functional Relationship (GM)

This estimator was proposed by Dent (1935). This estimator has been widely used, especially in fisher's researches:

$$\hat{\beta}_{GM} = Sign(Cov(x,y)) * \left(\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2} \right)^{1/2}$$

It can be noted that this estimator is symmetric in x and y. Where Cov(x,y) is the covariance of x and y. $\hat{\beta}_T = median(B_{ij})$

(3) Repeated Median Theil-Type Method (T)

Theil (1950) proposed this method. The data are ordered either to the x variable or the y variable. Find all possible pairs of observations, assuming that all x_i's are distinct,

$$B_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}};\ i = 1,2,..., j-1\ ,\ j = 2,3,...,n$$

which yields $\binom{n}{2}$ slope values, then where m can be chosen to be the maximum divisor of n such that $m \le r$. For example, when n = 20 then m = 4 and r =5 are selected.

(4) Proposed Method (AM)

This method consists of ordering the observed pairs (x_i, y_i)'s, i = 1,2,…,n; by the magnitude of x_i's , assuming that all x_i's are distinct, then divide the observation into some groups and find all possible paired slopes. The procedure can be described as follows:

a) Arrange the observations in ascending order on the basis of the values of x_i; i.e., $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ and the associated $y_{[1]}, y_{[2]}, ..., y_{[n]}$ of the original data are taken; then the new pairs will be $(x_{(i)}, y_{[i]})$

b) Divide the data into m-subgroup each of size r such that m*r = n; then the sample can be rewritten in the form in Figure 1 on the following page.

c) Find all possible paired slopes

$$\left\{ b(k)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}};\ i = 1,2..., j-1;\ j = 2,3,..., r \right\};$$
$$k = 1,2,..., m$$

d) Then the estimated value of the slope can be defined as follows:

$$\hat{\beta}_{AM} = \underset{k}{Median}\left\{ b(k)_{ij},\ i = 1,2..., j-1;\ j = 2,3,..., r \right\};$$
$$k = 1,2,..., m$$

Note that the suggested estimator is in the form of Theil's estimator with $m\binom{r}{2}$ paired slopes to be evaluated. If the sample size n is a prime number, then the estimates leads exactly to the repeated median Theil type estimator.

$$\begin{matrix} (x_{(1)}, y_{[1]}) & (x_{(2)}, y_{[2]}) & \cdots & (x_{(r)}, y_{[r]}) \\ (x_{(r+1)}, y_{[r+1]}) & (x_{(r+2)}, y_{[r+2]}) & \cdots & (x_{(2r)}, y_{[2r]}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{((m-1)*r+1)}, y_{[(m-1)*r+1]}) & \cdots & \cdots & (x_{(mr)}, y_{[mr]}) \end{matrix}$$

Figure 1

However the advantage of the proposed one is in abstracting the number of paired slopes to be evaluated, for example when n = 100, 4950 paired slopes are needed to be evaluated by using T method. By using the suggested method (AM), where r = m = 10, only 450 paired slopes are needed, which is a good advantage for this method.

Numerical Example

In order to compare various estimation methods, the so-called Pilot-Plant data from Daniel and Wood (1971) is considered. The observed (y) corresponds to acid content determined by titration and the observed (x) is the organic acid content determined by extraction and weighing. Moreover, Rousseeuw and Leroy (1987) analyzed this data further by assuming that one of the observations is wrongly recorded, i.e. the x-value of the sixth observation might have been wrongly recorded as 370 instead of 37. Based on the data which consist of 20 observations, and for the fact the x's data point should be distinct, x_{20} is substituted to be 168 instead of 167. The various estimated slopes yielded the results as shown in Table.1.

In this example, for the proposed method, the original sample is divided into 4 sub-samples, each of size 5. The results showed that traditional LS and GM methods have been strongly affected by the single outliers. On the other hand, AM and T are hardly affected by the wild observation.

Simulation Study

To illustrate the performance of the proposed method in the presence of outliers, a simulation study was carried out as follows: it begins by generating 100 observations according to the model; $y_i = 1 + x_i + \varepsilon_i$, where $x_i = 10\frac{i}{n}$ and $\varepsilon_i \sim N(0,1)$. Then, the data is contaminated; at each step a certain percentage of the observations are deleted and replaced with outliers' observations. The contaminated data point was generated according to the given relationship where $\varepsilon_i \sim N(20,25)$. Table.2 presents the values of the estimated slopes:

The properties of these methods were investigated further by looking at the mean square of error (MSE) in 10000 trials. For each 10000 trials, samples of size 20 and 50 were generated, the simulation results are represented in Table.3.

Table.1 The slope estimates using different methods for Pilot-Plant data

Slope	$x_6 = 370$	$x_6 = 37$
Least Squares (LS)	0.0808	0.3211
Geometric Mean (GM)	0.2148	0.3220
Theil (T)	0.3170	0.3194
Proposed method (AM)	0.3273	0.3480

Table.2. Slope Estimates with n= 100 and β =1

Contamination (%)	LS	GM	T	AM
0	0.9977	1.0590	0.9906	0.8491
10	-0.1176	-1.9339	0.8585	0.7911
20	-0.9760	-2.4261	0.6003	0.7675
30	-1.6041	-2.7429	-.05473	0.7574
40	-1.9215	-2.7781	-1.4783	0.5783
50	-2.0421	-2.8190	-1.7236	0.5214

Table.3. MSE of the Slope in the presence of outliers

Contamination (%)	Sample Size / Slope	20	50
0	LS	6.0016E-03	2.3847E-03
	GM	8.4800E-03	5.4053E-03
	T	6.5697E-03	2.5118E-03
	AM	1.2690E-01	7.1048E-02
10	LS	1.2115E+00	1.1850E+00
	GM	6.1172E+00	6.5467E+00
	T	2.7433E-02	2.1701E-02
	AM	2.7372E-01	1.9499E-01
20	LS	3.7599E+00	3.7167E+00
	GM	1.1129E+01	1.1212E+01
	T	1.8782E-01	1.7369E-01
	AM	2.3882E-01	1.0105E-01
30	LS	6.4511E+00	6.3880E+00
	GM	1.3218E+01	1.3285E+01
	T	2.4676E+00	2.2527E+00
	AM	3.2630E-01	3.0625E-01
40	LS	8.4146E+00	8.3348E+00
	GM	1.4609E+01	1.4647E+01
	T	5.8036E+00	5.6501E+00
	AM	2.1543E-01	1.5468E-01
50	LS	9.12418E+00	9.04105E+00
	GM	1.52952E+01	1.53539E+01
	T	7.13609E+00	7.00981E+00
	AM	**5.62811E-01**	3.85401E-01

Conclusion

Our simulation results from Table.3 indicate that, in terms of MSE the performance of the four estimators in the absences of outliers are comparable. However, as the degree of contamination increases LS and GM methods became very sensitive to the presence of outliers. Theil-Type estimator (T), clearly affected with the outliers when the contamination became 30% or more. It is very clear that the proposed estimator (AM) is very robust in the presence of outliers. As a conclusion, the AM estimator can be consider as a good alternative to the traditional methods because it is able to produce satisfactory results even in the presence of a large amount of outliers.

References

Daniel, C. & Wood, F. S. (1975). *Fitting equation to data*. John Wiley: New York.

Davies, P. L. (1993). Aspects of robust linear regression. *The Annals of Statistics, 21*, 1843-1899.

Dent, B. (1935). On observations of points connected by a linear relation. *Proceedings of the Physical Society, 47*, 92-106.

Fernandez, G. C. J. (1997). *Detection of model specification, outlier and multicollinearity in multiple regression using Partial regression/Residual plots.* SAS Users Group International Conference. San Diego, CA.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. McGraw-Hill Inc.

Montgomery, D. C. & Peck, E. A. (1992). *Introduction to linear regression analysis (2nd ed.)*. John Wiley: New York.

Olive, D. J. (2005). Two simple resistant regression estimators. *Computational Statistics and Data Analysis*, To Appear.

Rousseeuw, P. J. & Leroy, A. (1987). *Robust regression outlier detection*. John Wiley: New York.

Theil, H. (1950). A rank-invariant methods of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*(85), 173.