International Journal of Engineering Technology and Scientific Innovation

Volume:01,Issue:01

www.ijetsi.org

COMPARISON OF DIFFERENT METHODS OF A REGRESSION ANALYSIS OF SMALL SAMPLE

N. Abbasi[1] L. Nazari[1]

[1]Department of Statistics, Payame Noor University, I. R. Iran

ABSTRACT

There are many options for regression analysis. In this paper, several methods, including parametric and non-parametric methods were applied to a real data set. The calculations in this paper were performed by using the R software.

Keywords: Least Median of Squares; Least Square Error; Least Quantile of Squares; Maximum Likelihood Estimator; Mean Square Error; Posterior Mean Estimator.

1. INTRODUCTION

Our assumed measurement error model for the observations $\{(y_1, x_1), \dots, (y_n, x_n)\}$ is:

$$(1)\quad \begin{matrix} y = \beta_0 \mathbb{1}_n + \beta_1 \xi + \epsilon \\ x = \xi + u \end{matrix}$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$, $y = (y_1, \dots, y_n)'$, $u = (u_1, \dots, u_n)'$ and $\mathbb{1}_n$ is component n of a vector as $\mathbb{1}_n = (1, \dots, 1)'$. This is an error in variables model because ξ is not observed, and the observed data are $(y_1, \dots, y_n)'$ and $(x_1, \dots, x_n)'$.

Recently, theoretical and simulation results by Vidal and Bolfarine (2011)were prouder while x observed these model. For the least squares error method, maximum likelihood method, and the method of estimation of the mean of the posterior distribution (Bayesian analysis). Abbasi and Shokrizadeh (2013) presented a posterior distribution of the model structural parameters in the regression model with elliptical measurement error. The data was the activities of 17 faculty members of Shiraz Payame Noor University in a year (2011). It seemed more important aspects of the scientific production of each faculty member will depend on several factors. Job motivation, the desire and motivation to perform, and all internal and production of knowledge varied from one person to another. The posterior distribution of model properties can be obtained in accordance with a linear model using the least squares error. If more information is available from the prior distribution of the variable β_0 and β_1 certainly more precise results will be obtained.

It seems model (1) can be analyzed as a simple regression one, the same way as the calculation we perform for the least squares error method, As it is observed, there exists a outlier that is assumed to be processed with robust methods to obtained a model better than PME.

Linear least squares sensitive to outlier values. Robust estimators to deal with outliers are often limited by the breakdown point of samples measured (Donoho and Huber (1983), Hampel (1971)). We re-estimated the model with the robust methods. In the next section, we review three robust methods that is used in this article. We fitted least median of squares (LMS), least quantile of squares (LQS), and least squares trimmed (LST) done and the results Abbasi (2012) are compared. The basis for comparison, in addition to the model coefficients is the mean square error of each model.

2. METHODS

Let $\{u_i, \dots, u_n\}_{i:n}$ denote the i -th order statistic of n numbers $u_1, \dots, u_n$ .The estimators that are expressed in the following three methods are considered robust.

2-1. Least Median of Squares

The method of least median of squares was introduced by Rousseeuw (1984). According to the definition Rousseeuw (1984) and also Hampel (1975), we see that the median minimize the value of the absolute of residuals.

(2) $\mathrm{LMS(Z)} = \arg \min_{\hat{\beta}_1,\dots,\hat{\beta}_p} \mathrm{med}\{r_1^2, \dots, r_n^2\}.$

2-2. Least Quantile of Squares

Least quantile of squares (LQS) has been expressed by Kraks et al in 1997. The generalized method of median least squares (LMS), is defined as follows.

(3) $LQS(Z) = \underset{\hat{\beta}_1,\dots,\hat{\beta}_p}{argmin}\{r_1^2, \dots, r_n^2\}_{h_p:n}$

whereh_p with $p \le h_p \le n$ is typically set to $h_p = [(n + p + 1)/2]$ for maximum breakdown point.

2-3. Least Squares Trimmed

Trimmed least squares (LTS), or a least squares trimmed by Rousseeuw (1984) stated. Instead of the standard least squares method, which minimizes the sum of residual over n, the LTS method attempts to minimize the sum of the remaining K and $N - K$ is not affected. The ordinary least squares problem is to minimize the objective function of the square of the parameter values estimated residual, is defined as follows.

(5) $LTS(Z) = \underset{\hat{\beta}_1,\dots,\hat{\beta}_p}{argmin} \sum_{i=1}^{h_p}\{r_1^2, \dots, r_n^2\}_{i:n}$

3. SCIENTIFIC DATA ANALYISIS

Admission of graduate students (MS and PHD) is one of the factors which contribute to the increased level of scientific production at the universities. Participation score means whether the faculty has been and advisor or counsellor in a thesis committee at PHD or Master level. We have considered points for these roles. In surveying of the faculty members, we understood that the discrepancy of the variables is very high (from the zero to 160).In this article, the selected sample is the same location we work now, that is, Shiraz center. The individuals in the sample are not homogenous with respect to expertise, job background, and gender. If we think that the sample is a homogenous, it cannot be a relatively true condition. Thus Bayesian theory has been employed in order to consider the variable to have random parameters.

The data in following, has two variables: Numbers of the articles and score of participation in thesis committee. The data are 17 members of scientific board in one center of Payame Noor University in 2011. The members scientific productions depend on several factors, but their professional and internal motives to conduct and create science varies. It should be emphasized that if such activities are always supported, related participation and scientific productions will accordingly.

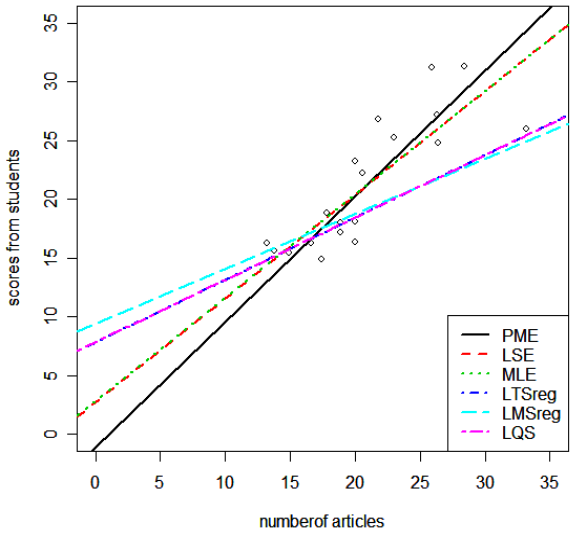

Figure 1: Six models fitted to the data

Table 2: coefficient and intercept of the regression model and the MSE under six different methods

	MLE	LSE	PME	LMSreg	LQS	LTSreg
$\hat{\beta}$	*0.8805*	*0.8805*	*0.8833*	*0.4675*	*0.5294*	*0.5294*
$\hat{\beta}_0$	*2.7864*	*2.7864*	*2.7603*	*9.4208*	*7.8594*	*7.8594*
MSE	*8.493621*	*8.493621*	*8.494876*	*16.62*	*16.60*	*16.60*

An important assumption in Bayesian analysis is to determine the distribution of the random variables. Normally this hypothesis for small data is not easily

established. Here is a simple relationship between two variables assumption, and the three methods which have been proposed based on the analysis done. It does not include observations of the outlier and some of the rest of the great distances. However, it will be interesting to see how accurate methods are robust in comparing probable ways. The calculations were done in software R and six linear models we have plotted on a graph simultaneously. As shown in Figure 1,a model of PME and LSE can be seen to have equal value. Also, in robust methods, the results are the LQS by ltsreg. As expected, Table 2 shows that the PME methods has the lowest MSE. The results show that the method based on the LSE and MLE methods does not work better. Certainly, by knowing the distributions of observations and the actual parameters, the model is better than the other models.

4. CONCLUSION

Six methods, Least Median of Squares; Least Square Error; Least Quantile of Squares; Maximum Likelihood Estimator; Mean Square Error; Posterior Mean Estimator, were used to analyze the data. Since the number of observations was small, putting hypotheses based on Bayesian theory could have achieved better results. Data analysis showed that Bayesian method(or PMS)than other methods of estimation, will lead to better results.

REFERENCES

[1] Abbasi, N. and Shokrizadah, M. (2012) An Application of Bayes Regression for a Small Sample, Journal of Novel Applied Sciences, 1-1/30-34

[2] Donoho, D., Huber, P., 1983. The notion of breakdown point. In: Bickel, P., Doksum, K., Hodges Jr., J.(Eds.), A Festschrift for Erich L. Lehmann. Wadsworth,Belmont, CA, pp. 157_184.

[3] Croux, C., Rousseeuw, P.J., 1992. Time-efficient algorithms for two highly robust estimators of scale.Computational Statistics 1, 411_428.

[4] Croux, C., Rousseeuw, P.J., Hössjer, O., 1994. Generalized S-estimators. Journal of the AmericanStatistical Association 89, 1271_1281.

[5] Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American StatisticalAssociation 79, 871_880.

[6] Rousseeuw, P.J., 1997. Introduction to positive-breakdown methods. In: Maddala, G.S., Rao, C.R. (Eds.),Handbook of Statistics, Vol. 15. Elsevier, North-Holland, Inc., Amsterdam, The Netherlands, pp. 101_121.

[7] Vidal, I. and Bolfarini, H. (2011) Bayesian estimation of regression parameters in elliptical measurement error models, Statistics and Probability Letters, 81, 1398–1406.