Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 77–82

A BAYESIAN METHOD TO DETECT OUTLIERS IN MULTIVARIATE LINEAR REGRESSION

Ufuk Ekiz*

Received 30. 04. 2002

Abstract

In this study, a Bayesian method will be introduced to describe outlying observations in multivariate linear regression. This method was proposed by Chaloner and Brant [3]. Later on, Varbanov [7] extended this method to apply to multivariate linear regression. According to Chaloner and Brant, an observation will be accepted as an outlier if the following condition is fulfilled: the posterior probability of the occurrence of the realized error (Arnold Zelner [8]) of an observation being greater than a critical value "k", is higher than the probability an error occurred in the model, with a critical value "k" over the assumed distribution. That is, if $\text{pr}[(\epsilon_i/\Sigma, y) > k] > \text{pr}(\epsilon_i > k)$ then the i^{th} observation will be accepted as an outlier. In the second section, the method proposed by Varbanov [7] will be considered. In the application section the existence or non-existence of outlying observations over the posterior distribution of the square form of the realized error in multivariate linear regression data is discussed.

Key Words: Realized error, posterior distribution, Jeffrey's prior, multivariate linear regression

1. Introduction

When a model is defined it is assumed that the random sample $Y_1, Y_2, \ldots, Y_n$ is composed of independent variables with the same distribution. Suppose we have observed values obtained from a random sample $Y_1, Y_2, \ldots, Y_n$. We would like to know two important things. Namely, whether they are from the same distribution, and even if they have the same distribution whether they have a large error or not. Answers to these questions are important since they determine whether the model can be used or not.

If some of observations do not come from the assumed distribution, Box and Tiao [1] called the observations outliers. According to Chaloner and Brant [3],

*Gazi University Faculty of Science, Department of Statistics, Ankara, Turkey.

an outlying observation is performed under the assumed model, but has a large random error.

Bayesian methods to describe outliers in linear regression models are divided into two groups, based on the underlying model. The first group comprises methods applied under the *null model* assumption $\epsilon \sim N(0, I)$. The second includes methods used assuming models such as mean-shift, variance-inflation and mixture which are defined when there are believed to be outlier observations. There are two different methods to find out possible outlying observations under the null mode. The first method, is called the *predictive density method* (Geisser S. [4]). The predictive conditional density is defined as follows:

$$\mathrm{pr}(y_I / y_{(I)}) = \int_\theta f(y_I) p(\theta / y_{(I)})\, d\theta$$

In this equation, y_I is a vector of possible outlying observations; $f(y_1/\theta)$ is the conditional probability density function of the outlying observations; $p(\theta/y_{(I)}) = \ell(\theta/y_{(I)})p(\theta)$ is the posterior density function of θ under the condition that the observations $y(I)$ are known. Using this function we test whether the suspected values y_I and $y_{(I)}$ came from the same distribution.

The second method was proposed by Chaloner and Brant [3]. According to this method, all observations were performed under the assumed model. However some of them may have a large error. In order to accept an observation as an outlier, the magnitude of the corresponding random error must reach a critical value. Using the probability density function of the random error associated with the assumed model, we want to find the critical value k for the model under consideration. If we choose the probability that non of the observations are outliers to be large, for example $0.95 = [\mathrm{pr}(\epsilon_i < k),\ \text{for all } i]$, then for a sample of size n we obtain $\mathrm{pr}(\epsilon_i < k) = (0.95)^{1/n}$ as the probability that a given observation is not an outlier. According to this, the probability that an observation is an outlier is $\mathrm{pr}(\epsilon_i > k) = 1 - (0.95)^{1/n}$. The "$k$" satisfying this equality is defined as the critical value. Actually, Chaloner and Brant use the condition that the probability $\mathrm{pr}((\epsilon_i/\Sigma, Y) > k)$ of the realized but unobserved error over the posterior function should be larger than the probability $\mathrm{pr}(\epsilon_i > k)$ of ϵ_i over the distribution defined in the model as an indicator that the i th observation is an outlier.

Varbanov [7] extended the work of Chaloner and Brant to multivariate linear regression analysis.

2. A Bayesian Method to detect Outliers in Multivariate Linear Regression

Let Y_i be the vector with p response variables on the i th sampling unit ($i = 1, 3, \ldots, n$). Let Y be the $n \times p$ data matrix defined by,

$$Y = \begin{bmatrix} Y_1^T \\ Y_2^T \\ \vdots \\ Y_n^T \end{bmatrix} = \begin{bmatrix} Y^{(1)} & Y^{(2)} & \cdots & Y^{(p)} \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix}.$$

Here $Y^{(j)}$ is a vector composed of n independent observations of the j th variable. The Multivariate Linear Regression model can be written as,

$$Y^{(j)} = X\theta_j + \epsilon^{(j)},\ j = 1, 2, \ldots, p$$

where X is the $n \times p$ dimensional design matrix. In matrix form, the model is expressed as

$$Y = X\Theta + E,\ \Theta = (\theta_1, \theta_2, \ldots, \theta_p).$$

The i th row of the matrix E contains the random error of Y_i, where E is defined as follows. Here the ϵ_i's are assumed to be random variables of he same Normal distribution having independent "0" and "I" parameters with dimension p. In this case the model $Y = X\Theta + E$ can be written as,

$$Y_i = \Theta^T X_i + \Sigma^{1/2}\epsilon_i,\ i = 1, 2, \ldots, n$$

$$(\epsilon_i = (\epsilon_{1i}, \epsilon_{2i}, \ldots, \epsilon_{pi}) = \Sigma^{-1/2}(Y_i - X_i\Theta) \sim N(0, I)).$$

Varbanov used the following equation to test whether the i th observation in

$$\delta_i = \epsilon_i^T \epsilon_i = (Y_i - \Theta' X_i)' \Sigma^{-1} (Y_i - \Theta' X_i)$$

is an outlier or not. If the ϵ_i's have p variables having mean 0 and the variance of the covarians matrix of I and the same distribution, δ_i will have central chi-square distribution of p degrees of freedom.

If for the appropriate model distribution (namely a chi-square distribution with p degrees of freedom) we choose the probability of no outliers to be large, say 0.95, then for a sample of size n

$$\begin{aligned} 0.95 &= \mathrm{pr}\{\delta_i \le k \text{ for all } i\} \\ &= \mathrm{pr}\{\delta_1 \le k\} \cdot \mathrm{pr}\{\delta_2 \le k\} \cdots \mathrm{pr}\{\delta_n \le k\} = \{F_p(k)\} \end{aligned}$$

Hence the critical value k may be found from the equality $\mathrm{pr}\{\delta_i \le k\} = \{F_p(k)\}^{1/n}$ for the probability of any observation not being an outlier.

The fact that $\mathrm{pr}\{(\delta_i/\Sigma, Y) > k\}$, the posterior probability that the square form δ_i of the realized but unobserved error exceeds the critical value k, is larger than $\mathrm{pr}(\delta_i > k)$, the probability that the statistic δ_i exceeds k under the hypotheses of the model, indicates that the i th observation may be an outlier.

Another approach is to use Bayes Factor to test the hypothesis,

$H_{0i} : \delta_i > k$, if Y_i is an outlying observation,

$H_{1i} : \delta_i \le k$.

The Bayes Factor used to test H_{0i} against H_{1i} is the ratio of posterior odds to prior odds. Bayes Factor B_i is expressed as

$$B_i = \frac{p_i F_p(k)}{(1 - p_i)\{1 - F_p(k)\}},$$

where the posterior probability p_i will be defined in section 3.

Kass and Raftery [6] say that if $B_i > 10$, H_{0i} is valid; if $B_i > 100$, H_{0i} is certainly valid.

3. The Posterior Distribution of the square form of the Realizable Error

In order to obtain the posterior distribution of $\delta_i = \epsilon_i^T \epsilon_i$, which is the square form of the realized error, the posterior distribution of $(\epsilon_i/\Sigma, Y)$ has to be obtained. Since $\epsilon_i = \Sigma^{-1/2}(Y_i - \Theta^T X_i)$ is a linear function of Θ and Σ, the joint posterior distribution of Θ and Σ,

$$\text{pr}(\Theta, \Sigma/Y) = \ell(\Theta, \Sigma/Y)\text{pr}(\Theta, \Sigma)$$

must be obtained. Here $\ell(\Theta, \Sigma/Y)$ denotes the likelihood function of Θ and Σ, where Y is given and $\text{pr}(\Theta, \Sigma)$ denotes the joint prior distribution. If Σ is regarded as a nuisance parameter, that is its only importance is in making inferences from Θ, the posterior distribution of Θ and Σ can be expressed as

$$\text{pr}(\Theta, \Sigma/Y) = \text{pr}(\Theta/\Sigma, Y) \cdot \text{pr}(\Sigma/Y).$$

In multivariate linear regression analysis, Jeffrey's prior $\text{pr}(\Theta, \Sigma) = \text{pr}(\Theta) \cdot \text{pr}(\Sigma) = |\Sigma|^{-\frac{1}{2}(p+1)}$ is used to find the joint posterior distribution of Θ and Σ as follows (Box and Tiao [2]):

$$\text{pr}(\Theta, \Sigma/Y) = |\Sigma|^{-\frac{1}{2}(n+p+1)} \exp\{-\tfrac{1}{2}\text{tr}\, \Sigma^{-1}[A + (\Theta - \hat{\Theta})' X' X (\Theta - \hat{\Theta})]\},$$
$$-\infty < \Theta < \infty, \ \Sigma > 0.$$

Here, $A = \{a_{ij}\}$, $a_{ij} = (Y_i - X\hat{\Theta}_i)'(Y_j - X\hat{\Theta}_j)$, $i, j = 1, 2, \dots p$.

If Σ and Y are known, the posterior distribution of Θ; and if Y is known, the marginal posterior distribution of Σ, are obtained respectively as follows:

$$\begin{aligned} (\Theta/\Sigma, Y) &\sim N(\hat{\Theta}, \Sigma \otimes (X'X)^{-1}), \\ (\Sigma/Y) &\sim W^{-1}(S, n - q - p + 1). \end{aligned}$$

Here $\hat{\Theta} = (X'X)^{-1} X' Y$ and $S = (Y - X\hat{\Theta})'(Y - X\hat{\Theta})$.

Using these posterior distributions the posterior distribution of $(\epsilon_i/\Sigma, Y)$ can be obtained as:

$$(\epsilon_i/\Sigma, Y) \sim N(\hat{\epsilon}_i = \Sigma^{-1/2}(Y_i - X_i\hat{\Theta}), (X_i'(X'X)^{-1}X_i)I).$$

Let $\sigma_{(i)} = X_i'(X'X)^{-1}X_i, \lambda_i = \sigma_{(i)}^{-1}(Y_i - \hat{\Theta}'X_i)'\Sigma^{-1}(Y_i - \hat{\Theta}'X_i)$. Then given Σ and Y, the posterior distribution of $T_i = \dfrac{\delta_i}{\sigma_{(i)}}$ is non-central chi-square with p degrees of freedom and parameter of non-centrality λ_i. So, we can write p_i, $i = 1, 2, \dots, n$, as

$$p_i = E_{\Sigma/Y}\{\text{pr}(T_i > \frac{k}{\sigma_{(i)}}/\Sigma, Y\}$$

or, in terms of Σ^{-1}, as

$$p_i = E_{\Sigma^{-1}/Y}\{\text{pr}(T_i > \frac{k}{\sigma_{(i)}}/\Sigma^{-1}, Y\}.$$

In practice p_i can be calculated using the Monte-carlo simulation technique.

4. An Application

The Bayesian method was applied to some data in “A Handbook of Small Data Sets” by Hand, Daly, Lunn, Conway and Ostrowski [5]. The data sets cover the following.

Independent variables X.

X_1; GNP implicit price deflator (1954 = 100.0)

X_2; GNP

X_3; Unemployment

X_4; Size of armed forces

X_5; Non-institutional population aged 14 and over

Dependent variables Y.

Y_1; Census: Agricultural employment

Y_2; Census: Self-employment

Y_3; Census: Unpaid family workers.

There are 16 observations ($n = 16$). The values of p_i and B_i were calculated using the MAT-LAB12 computer program. If the probability of the existence of no outlier was chosen to be 0.95, k was calculated as 13.7931. The corresponding values of p_i and B_i are tabulated in Table 1.

Table 1

I	$\mathbf{p_i}$	$\mathbf{B_i}$
1	0.0039	1.2171
2	7.4338E-004	0.2313
3	0.0043	1.3244
4	0.0074	2.2888
5	1.1680E-004	3.6374E-004
6	1.0759E-006	3.3507E-004
7	3.7271E-005	0.0116
8	3.0506E-0.006	9.500E-004
9	3.3058E-0.05	0.0103
10	1.1580E-004	0.0361
11	2.8354E-005	0.0088
12	7.2055E-004	0.2242
13	6.1890E-005	0.0193
14	3.1934E-007	9.9454E-005
15	1.4678E-006	4.5712E-004
16	0.0025	0.7716

Non of the B_i's were found to be greater than 10. However, for the 1 st, 3 th and 4 th observations, the limiting values of p_i were calculated to be $1 - \{F_p(k)\}^{1/n} = 0.0032006977101885$, greater than the prior probabilities. For this reason, these observations need more attention.

5. Conclusion

The logic behind the Bayesian method is quiet simple and the method is easily applicable. In addition to the non-informative prior function, other prior functions can be used to obtain the posterior distribution of the square form of the realized error. But in this case the posterior distribution may not be of a known form. This must certainly be taken into account. The non-informative prior, who's use is foreseen when no prior information is available, or when such information varies from person to person, has been used to obtain the posterior distribution of the realized error.

References

[1] Box, G. E. P. and Tiao, G. C. A Bayesian approach to some outlier problems, Biometrika **55 (1)**, 119, 1968.

[2] Box, G. E. P. and Tiao, G. C. Bayesian inference in statistical analysis, Addison-Wesley, Reading, MA, 1973.

[3] Chaloner, K. and Brant, R. A Bayesian approach to outlier detection and residual analysis, Biometrika **75**, 651–9, 1988.

[4] Geisser, S. (1965). Bayesian estimation in multivariate analysis, Ann. Math. Statist. **36**, 150–9, 1965.

[5] Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. A Handbook of small data sets, Chapman & Hall, 1994.

[6] Kass, R. and Raftery, A. Bayes factors, JASA **90**, 773–95, 1995.

[7] Varbanov, A. Bayesian approach to outlier detection in multivariate normal samples and linear models, Commun. Statist.-Theory Meth. **27 (3)**, 547–557, 1998.

[8] Zellner, A. Bayesian Analysis of Regression Error Terms (Section on Theory and Methods), Volume 70, Number 349, 1975.