G.U. Journal of Science
19(2): 113-117 (2006)

GUJS
www.gujs.org

Bayesian Approach to identify Masking and Swamping Problems in the Multivariate-Multiple Regression Analysis

Ufuk EKİZ♣

Gazi University, Faculty of Science and Letters, Department of Statistics, 06500, Besevler, Ankara, TURKEY

Received: 03. 12. 2004Accepted: 27. 01. 2006

ABSTRACT

The Bayesian method developed by Chaloner and Brant to see whether, for instance, the outlier observation is included in the linear model or not, focuses on posterior distributions of realized but unobserved errors [5]. This method has been expanded by Varbanov for the multivariate-multiple regression model in a way to provide analysis opportunities for separate observations in terms of their being an outlier [13]. This study expanded the proposed method by Varbanov to offer the opportunity for analyzing whether observations in groups are an outlier or not. Therefore, it is likely to claim the existence of masking and swamping problems.

Key Words: Realized but unobserved errors, masking and swamping problems, outlier observation.

1. INTRODUCTION

Barnett & Lewis, Hawkins, Bekman & Cook and Pettit & Smith [2], [10], [4], [12] conducted comprehensive studies on outlier observation. Most of the Bayesian approaches that are used to determine the outlier observation make use of the Freeman's definition [6]. This definition is as follows 'the observation which is not from the same distribution as the others in the sample is called the outlier observation'. The definition by Freeman requires a defining by another model that is the generator of the outlier observation. [3] and [9] serve as sample studies conducted in accordance with this definition. The method proposed by Chaloner & Brant gives a different definition of the outlier observation from Freeman as focusing on the assumed model [5]. This method was proposed by Zellner and defines the outlier observation through realized but unobserved error [14]. Another approach which does not require a whole new different definition is proposed by Geisser [7, 8].

In the proposed method by Chaloner & Brant, to determine the outlier observation in the linear model, the ith observation's realized but not observed errors' being bigger than "k" critical value, its posterior probability's being bigger than the prior probability stemmed from the assumed model are all considered as an indicator of ith 's being the outlier observation. Accordingly, it is more likely to define the observation/s farther than the others as the outlier. Namely, observations called the outlier will be one of the extreme observations. Therefore, it becomes inevitable to use a sub-ordering principles which enables to put observations in order according to their magnitude in the multivariate linear observations. Barnett examined a variety of sub-ordering principle in multivariate problems and categorized them into four (1). These categories are named as reduced, marginal, partial and conditional. Reduced sub-ordering is the one and only criterion to define multivariate outlier observations. This measurement is defined as sequencing p-dimensional $Y_1, Y_2, ..., Y_n$ observations as in the example given below in terms of single- variance R(Y) statistics. If,

♣Corresponding author, e-mail: ufukekiz@gazi.edu.tr

$$R(Y_i) = \max\left\{R(Y_j); j = 1,2,...,n\right\} \quad (1)$$

is, Y_i observation is thought to be the outlier. Many frequency methods use the squared form of the

$$R_j = R(Y_j, \mu, \Sigma) = (Y_j - \mu)\Sigma^{-1}(Y_j - \mu)' \quad (2)$$

as $R(Y_i)$ statistics to determine multivariate outlier observations. Here, μ represents the center of the distribution parameter and variance- covariance matrix in Σ. The approach by Varbanov, Chaloner & Brant, has been expanded in the multivariate linear model by using the squared form stated in (2), assuming that independent and parameters of errors as 0 and Σ has the same multivariate normal distribution [13, 5]. Nevertheless, in this study the observations can be examined one by one to check if they are the outlier or not. On the other hand, the existence of masking and swamping problems should be clarified; observations should be analyzed in groups to check the outlierness. Therefore, in section 2 the expansion will enable the analysis of the Varbanov's approach and observations in groups for multivariate-multiple regression analysis, followed by an implementation in section 3.

2. EXAMINATION OF OBSERVATIONS' OUTLIERNESS IN GROUPS IN THE MULTIVARIANCE-MULTIPLE REGRESSION ANALYSIS

When p is considered as the variable number and ith dependent variable Y_i, i=1,2,...,n, as,

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

(nxp) dimensional data matrix is given. Here, $Y^{(j)}$ is the vector formed from n numbered dependent observations in (j=1,2,...,p), jth variable.

Multivariate-multiple regression model is hypothesized as,

$$Y^{(j)} = X\theta_j + \varepsilon^{(j)} \quad ; j=1,2,...,p$$

where a X (nxq) is the dimensional design matrix. This model in the form of a matrix is defined as,

$$Y = X\Theta + E \quad , \ \Theta = (\theta_1, \theta_2, ..., \theta_p) \quad (3)$$

It covers ith line of the E matrix of (nxp) dimension, random errors of Y_i and formulated as;

$$E = \begin{bmatrix} \varepsilon_i\Sigma^{1/2} \\ \varepsilon_2\Sigma^{1/2} \\ . \\ . \\ . \\ \varepsilon_n\Sigma^{1/2} \end{bmatrix} = \begin{bmatrix} \varepsilon_{11}, \varepsilon_{12}, ..., \varepsilon_{1p} \\ \varepsilon_{21}, \varepsilon_{22}, ..., \varepsilon_{2p} \\ . \\ . \\ . \\ \varepsilon_{n1}, \varepsilon_{n2}, ..., \varepsilon_{np} \end{bmatrix} \Sigma^{1/2}$$

Here, it is assumed that there are random variables where the ε_i's are independent, mean is 0 and the variance-covariance matrix is the p dimensioned unit matrix (having the same multivariate normal distribution). Therefore, the model in (3) can also be written as,

$$Y_i = X_i\Theta + \varepsilon_i\Sigma^{1/2} \quad ; \ i=1,2,...,n \quad (4)$$

It is $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, ..., \varepsilon_{pi}) = (Y_i - X_i\Theta)\Sigma^{-1/2} \sim N(0, I)$

For the observations in the n diameter sample, 2^n numbers of different probable situations exist. The group composed of α numbers of observations, probable of being an outlier which is found in each of the probable situations is called group A. Defined as (α=0,1, 2,...,n),

$$\varepsilon^* = \left[\varepsilon_{11}, ..., \varepsilon_{1p}; \varepsilon_{21}, ..., \varepsilon_{2p}; ...; \varepsilon_{n1}, ..., \varepsilon_{np}\right] = \left[\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\right] \quad (5)$$

If in the squared form of realized but unobserved errors for the probable outlier observations that took place in A,

$$R_A = \varepsilon_A^* \varepsilon_A^{*\prime} \quad (6)$$

the posterior probability of its "k_A" critical value, being greater than the prior probability related to the observations' outlierness will result with the observations in A being defined as the outlier. The prior probability related to the observation group's outlierness in A and "k_A" critical value is defined as given below.

For example, when the prior probability of consisting no outlier observation is chosen as a large number, (e.g. 0.95) then the prior probability of any of the observations (ith observation) in the sample as the outlier,

$$0.95 = \Pr\{R_i \le k \text{ , for each i }\} = \left\{F_p(k)\right\}^n$$

is found in the form of $(0.95)^{1/n} = F_p(k)$. Here, F_p represents the distribution function of the R_i random variable having a chi-square distribution with p degrees of freedom. The prior probability of any observation's outlierness becomes $(1-F_p(k)) = [\Pr(R_i > k),\ i \text{ için}]$. The prior probability of α number of observation in A as being the outlier is defined as;

$$\Pr(R_A > k_A) = \Pr(R_i > k, R_j > k, ..., R_t > k),$$

$$\{i, j, ...t\} \in A$$

$$= \Pr(R_i > k)\Pr(R_j > k)...\Pr(R_t > k)$$

$$= [\Pr(R_i > k)]^{\alpha}$$

$$= [1 - F_p(k)]^{\alpha}$$

[5]. The critical value presenting $[1-F_p(k)]^{\alpha}$ probability is formed as

$$k_A = F_{\alpha p}^{-1}\left[[1-F_p(k)]^{\alpha}\right].$$

In another approach,

$H_{0i} : R_A \geq k_A$, if the observations in A are outlier (i=1, 2,..., n)

$H_{1i} : R_A < k_A$

to test this hypothesis, the Bayesian factor is referred. The bayes factor used to test H_{0i} against H_{1i},

$$B_A = \frac{\frac{p_A}{1-p_A}}{\frac{[1-F_p(k)]^{\alpha}}{\{1-[1-F_p(k)]^{\alpha}\}}} = \frac{p_A\{1-[1-F_p(k)]^{\alpha}\}}{(1-p_A)[1-F_p(k)]^{\alpha}}$$

is the proportion of posterior odd to prior odd. Kass and Raftery state that B_i 's being bigger than 10 is a strong and that of 100 is a much stronger proof for the validity of H_{0i} [11].

To get $p_A = \Pr(R_A > k_A)$ posterior probability, realized but unobserved error vector symbolized as ε^* and marginal posterior distribution of ε_A should be accessed.

In the multivariate-multiple regression analysis, joint non-informative prior function of $p(\Theta, \Sigma) = p(\Theta)p(\Sigma) = |\Sigma|^{-(p+1/2)}$ and likelihood function of $\ell(\Theta, \Sigma / Y)$ multiplied, Θ and Σ parameters' joint posterior distribution becomes,

$$p(\Theta, \Sigma / Y) \alpha |\Sigma|^{-\frac{1}{2}(n+p+1)} \exp\left\{-\frac{1}{2} iz\Sigma^{-1}\left[A + (\Theta - \hat{\Theta})'X'X(\Theta - \hat{\Theta})\right]\right\}$$

, $-\infty < \Theta < \infty$, $\Sigma > 0$

$A = \{a_{mj}\}$, $a_{mj} = (Y_m - X\hat{\theta}_m)'(Y_j - X\hat{\theta}_j)$ covariance matrix and proportional (pxp) dimension symmetrical matrix related to the example]. When Σ and Y is evidently known; the marginal posterior distribution of Θ,

$$(\Theta / \Sigma, Y) \sim N(\hat{\Theta}, \Sigma \otimes (X'X)^{-1}) \quad (7)$$

is pq dimension normal distribution. When Y is known; the marginal posterior distribution of Σ,

$$(\Sigma / Y) \sim W^{-1}(A, n-q-p+1) \quad (8)$$

is (n+p-q+1) degrees of freedom inverted wishart. The marginal posterior distribution of (Σ^{-1} / Y), where the transformation of Σ into Σ^{-1} jacobian as $|\Sigma|^{p+1}$,

$$(\Sigma^{-1} / Y) \sim W(A^{-1}, n-p-q+1) \quad (9)$$

(n-p-q+1) degrees of freedom wishart distribution is achieved.

In regards with (7) and (8), the posterior distribution of $(\varepsilon / \Sigma, Y)$ as $H = X(X'X)^{-1}X'$ is formed as,

$$p(\varepsilon / \Sigma, Y) \propto \exp\left\{-\frac{1}{2} iz\left\{I_{pxp}\left[\hat{\varepsilon} - \varepsilon\right]' H\left[\hat{\varepsilon} - \varepsilon\right]\right\}\right\}$$

The posterior distribution of ε^* is formulated as,

$$p(\varepsilon^* / \Sigma, Y) \propto \exp\left\{-\frac{1}{2}(\varepsilon^* - \hat{\varepsilon}^*)(H \otimes I_{pxp})(\varepsilon^* - \hat{\varepsilon}^*)'\right\} \quad (10)$$

To get the posterior probability of $\Pr(R_A = \varepsilon_A^* \varepsilon_A^{*\prime} > k_A)$, a transformation enabling the $(\varepsilon^* / \Sigma, Y)$ variance covariance matrix $(H \otimes I_{pxp})$ into I_{npxnp} form should be exercised.

U; the matrix of $(H \otimes I_{pxp})$ eigen vectors matrix of (npxnp)

Λ ; diagonal matrix of (npxnp) consisting $(H \otimes I_{pxp})$ matrix eigen value, the equivalence no (10);

$$\varepsilon^* = Z\Lambda^{-1/2}U'$$

when the transformation is realized, as jacobian transformation is fixed, $Z\Lambda^{-1/2}U'$ is written instead of ε^*, under the condition that Σ is known where posterior function of Z is,

$$p(Z/\Sigma, Y) = c.\exp\left\{-\frac{1}{2}\left\{\left[Z-\hat{\varepsilon}^*\Lambda^{-1/2}U'\right]\left[Z-\hat{\varepsilon}^*\Lambda^{-1/2}U'\right]'\right\}\right\}$$

This is the multivariate normal distribution with np dimension ($(Z/\Sigma, Y) \sim N_{np}(\hat{\varepsilon}^*\Lambda^{-1/2}U'; I_{npxnp})$).

The posterior distribution, αp degree of freedom and non-central parameter of $R_A = Z_A Z_A'$'s amount, with the condition that Σ is known, is the chi square distribution:

$$\lambda_A = \hat{\varepsilon}_A^*(H_A \otimes I)^{-1}(\hat{\varepsilon}_A^*)'$$

At the end of the transformation, the critical value considered as

$$K = \left(\sqrt{\frac{k_A}{\alpha p}}; \sqrt{\frac{k_A}{\alpha p}}; \ldots; \sqrt{\frac{k_A}{\alpha p}}\right)_{1x\alpha p},$$

$$k_Z = K(H_{ij} \otimes I)^{-1}K'$$

this formula is formed. Via this non-central posterior distribution,

$$p_A = E_{\Sigma/Y}\left\{Pr(R_A > k_Z/\Sigma, Y)\right\}$$

or

$$p_A = E_{\Sigma^{-1}/Y}\left\{Pr(R_A > k_Z/\Sigma^{-1}, Y)\right\}$$

posterior distribution is achieved. In practice, for all the probable values of α, the p_A posterior probabilities are obtained through the Monte-Carlo simulation technique written in the MATLAB program.

3. IMPLEMENTATION

Simulation data were preferred to be used to see whether the proposed Bayesian approach is successful in determining the observations that cause the problems of masking and swamping; another reason for this preference was the difficulty to find the multivariate regression data on which outlier observations have been worked on. The data simulated appropriately to the model in (3) are displayed in Table 1.

Table 1: Simulated multivariate- multiple regression data

	Y_1	Y_2	X_1	X_2	X_3	X_4
1	201.9410	291.4064	3.7146	7.8149	20.0148	50.1134
2	201.6816	292.7837	4.0972	8.6551	20.1234	49.5626
3	200.7539	289.8407	3.7034	7.7940	20.1384	49.6584
4	200.4720	289.9292	3.6196	8.1439	19.8738	49.6624
5	200.6772	289.6901	3.8804	7.7620	19.8304	49.7849
6	202.4464	293.0461	3.7959	8.4305	19.5707	50.3636
7	202.4720	292.5687	4.2398	7.8751	19.9736	50.1404
8	202.6303	292.5790	3.7222	7.9704	19.9132	50.3735
9	201.3104	291.7532	4.4491	8.1899	20.1471	49.4528
10	201.0433	290.0800	3.4208	7.8784	19.8206	50.0376
11	203.3517	293.0813	4.0029	7.5594	20.0365	50.5701
12	201.6263	291.9232	4.3148	8.2653	19.3091	50.1461
13	202.4640	292.1624	3.9079	7.7772	19.9119	50.3202
14	200.3946	290.0285	3.8040	8.1740	19.9795	49.4893
15	202.0619	292.5877	3.5244	8.5310	19.8110	50.1296
16	202.2683	292.2683	4.0000	8.0000	20.0000	50.0000
17	205.2317	297.7317	4.5000	8.5000	20.5000	50.5000

In the data where the formulations are ; n=17, p=2, q=4 and

$$S = \begin{bmatrix} 0.1323 & 0.1330 \\ 0.1330 & 0.1344 \end{bmatrix}$$

the 16th and 17th observations are generated out of a different normal distribution with different parameters from the normal distribution where the other observations are generated. Therefore, the simulation data displayed in Table 1 are designed as data because they were different from the 16th and 17th observations. In order to reduce the operational complexity and for easier interpretation, the number of observations is restricted to 17.

When the probability of none of the observations' outlierness is chosen as 0.95, k_A=16.3148 is achieved in the process of analyzing the observations' outlierness separately. In this stage, the p_i posterior probabilities obtained from the Monte Carlo simulation program which has 20000 circles and written in MATLAB package program and the 16th as well as 17th observations out of the B_i Bayesian factor vales are found significant. For the 16th and 17th observations, these values are (0.0052, 18.1231), (0.0027, 9.5100) respectively.

When analyzing the observations' outliernesses in pairs k_A=28.6987 is found. According to the p_{ij} and B_{ij} values, all double sub-groups where the 4th

and 16^{th} observations take place and all pairs including (2,4,7,9,12,16).th and the 17^{th} observation are found as outlier. As the 4^{th} observation did not have the p_i value in single analysis, it is concluded that it has been hidden by other observations. It can be said that the 4^{th} observation swamps all other observations except for the 16^{th} and 17^{th} as all the pairs where the 4^{th} observation take place are found as outlier. All paired sub groups which have been found as outlier in the 16^{th} and 17^{th} observations, are swamped by the 16^{th} and 17^{th} observations. When all triple probable sub -groups are examined, it is seen that all single and double sub-group observations which have been previously found as outlier leads to the problems of masking and swamping (k_A = 46.1733 is found during the analysis of triple observation groups).

The evaluation based on the whole results show that the 4^{th}, 16^{th} and 17^{th} observations have significant impact on the model parameter estimations. If separate difference- contradiction- disagreement outliernesses of the observations were analyzed, the effect of the 4^{th} observation would not be realized.

4. CONCLUSION

Determining the outlier observations in multivariate linear models and using the posterior distribution of realized but unobserved error was first proposed by Chaloner & Brant [5]. In line with that, Varbanov suggested the use of the posterior distribution of realized but unobserved error's squared form to determine the outlier observations of multivariate-multiple linear models [13]. In this study, the purpose was to prove the existence of masking and swamping problems in the multivariate-multiple linear models. Therefore, the method proposed by Varbanov has been expanded in order to provide analysis of the observation's outlierness in groups. For example, with this expansion enabling the analysis of all the probable sub-groups' outliernesses, the results to be interpreted can be achieved easily. Moreover, this method does not require defining a new model except for the proposed one for the outlier observations.

This method used for a non-informative priority function can be expanded in cases where the prior information related to the model parameters exists. Nevertheless, the prior function to be used, except for the conjugate prior function, may hinder gathering the posterior function of the statistics in [2] in an evidently known form.

REFERENCES

[1] Barnett, V., "The ordering of multivariate data (with Discussion)", ***J. Roy. Statist. Soc. Ser. A***, vol. 139, pp. 318-54, 1976.

[2] Barnett, V., Lewis, T., "Outliers in Statistical Data", 3rd ed. ***Chichester***: John Wiley & Sons, 1994.

[3] Box, G. E. P., Tiao, G. C., "A bayesian approach to some outlier problems", ***Biometrika***, vol.55, pp. 119-29, 1968.

[4] Beckman, R. J. Cook, R. D. "Outliers (with discussion)", ***Technometrics***, vol. 25, 119-63, 1983.

[5] Chaloner, K., Brant, R., "A Bayesian approach to outlier detection and residual analysis", ***Biometrika***, vol. 75, 651-9, 1988.

[6] Freeman, P. R., "On the number of outliers in data from a linear model", ***In Bayesian Statistics***, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 349-65, Valencia: University Press, 1980.

[7] Geisser, S., "Discussion of a paper by G. E. P. Box." ***J. R. Statist. Soc. A***, vol. 143, 416-7, 1980.

[8] Geisser, S., "Influential observations, diagnostics and discordancy tests" ***J. Appl. Statist.***, vol. 14, 133-42, 1987.

[9] Guttman, I., Dutter, R., Freeman, P. R., "Care and handling of univariate outliers in the general linear model to detect spuriosity-a bayesian approach", ***Technometrics***, vol. 20, 187-93, 1978.

[10] Hawkins, D., "Identification of Outliers", ***London: Chapman and Hall***., 1980.

[11] Kass, R., Raftery, A., "Bayes factors", ***JASA***, vol. 90, pp. 773-95, 1995.