The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 805-814

A Bayesian Approach to Detecting Outliers Using Variance-Inflation Model

Sangjeen Lee[1] and Younshik Chung[2]

Abstract

The problem of 'outliers', observations which look suspicious in some way, has long been one of the most concern in the statistical structure to experimenters and data analysts. We propose a model for outliers problem and also analyze it in linear regression model using a Bayesian approach with the variance-inflation model. We will use Geweke's(1996) ideas which is based on the data augmentation method for detecting outliers in linear regression model. The advantage of the proposed method is to find a subset of data which is most suspicious in the given model by the posterior probability. The sampling based approach can be used to allow the complicated Bayesian computation. Finally, our proposed methodology is applied to a simulated and a real data.

Keywords : Gibbs sampler; Latent variable; Linear regression model; Mean shift model; Outlier; Metropolis-Hastings algorithm; Variance inflation model.

1. Introduction

The problem of 'outliers', observations which look suspicious in some way, has long been one of the most concern in the statistical structure to experimenters and data analysts. An outlier is ususally defined to be an observation that does not come from the assumed model or an extreme observation that is far away from the rest of observations. Giving a precise definition to the concept of an outlier is difficult since the notion of an 'extreme observation' is subtle. We refer the reader to Pettit and Smith(1985) for a discussion.

In this paper, we propose a model for an outlier problem and also analyze it using a Bayesian approach. The Bayesian approaches for outlier detection can be classified into two procedures such as using alternative model for outliers or not. For the method without having alternative model, the predictive distribution is used in Geisser(1985) and Pettit and Smith(1985), or the posterior distribution is used in Chaloner and Brant(1988) and Guttman

1) School of Computer and Information, Ulsan College, Ulsan, 682-090 Korea
2) Department of Statistics, Pusan National University, Pusan, 609-735 Korea
E-mail : yschung@hyowon.cc.pusan.ac.kr

and Pena(1993). For alternative model, the mean-shift model and the variance-inflation model are used in Guttman(1973) and Sharples(1990).

Let y be an observation vector from $N(\mu, \sigma^2)$. The mean-shift model is that a suspicious observation is distributed as $N(\mu+m,\ \sigma^2)$. If m is not a zero, the corresponding observation is decided as an outlier, Guttman(1973) applied the mean-shift model to a linear model. Recently, Chung and Kim(1999) considered mean-shift model to a mixed linear model using MCMC method. The variance-inflation model is that an observation y_i be from $N(\mu,\ b_i\sigma^2)$. The observation, y_i, with b_i >> 1 y_i, is treated as an outlier (Box and Tiao, 1968). Sharples(1990) showed how variance inflation can be incorporated easily into general hierarchical models, retaining tractability of analysis. For detecting outliers, we use Geweke(1996)'s method, which was introduced as the method for selecting the best predictors in multiple regression model using Gibbs sampler.

The plan of this article is as follows. In section 2, we introduce the variance-inflation model and motivate the hierarchical framework to which Geweke's(1996) method is applied to detect the outliers. Finally, we apply our proposed methodology to a simulated data and a real data set (Darwin's data: Guttman, Dutter and Freeman, 1978) in section 3.

2. Variance-Inflation Model

2.1. Geweke's method

For detecting outliers in linear regression model, the variance-inflation model is used. This method is that an observation y_i is assumed to come from $N(\mu,\ c_i^2\sigma^2)$, $i=1,\cdots,n$. Then the observation, y_i, with $c_i \gg 1$, is treated as an outlier (Box and Tiao, 1968). Sharples(1990) showed how the variance inflation can be incorporated easily into general hierarchical models, retaining analytical tractibility.

In this paper, we use the variance-inflation model for the linear regression problem. Specifically, we assume that for some particular $(n\times p)$ design matrix X of constants, it is intended to generate data $y=(y_1,\ldots,y_n)^t$ such that

$$y = X\beta + \varepsilon\ , \tag{2.1}$$

where $\beta=(\beta_1,\ldots,\beta_p)^t$ is a set of p unknown regression parameters, and the $(n\times 1)$ error vector ε is normally distributed with mean vector 0, and variance-covariance matrix $\sigma^2 C_n$, where σ^2 is unknown and C_n is an $n\times n$ diagonal matrix with diagonal elements

$(\frac{1}{b_1^2}, \cdots, \frac{1}{b_n^2})$ for the sake of computational convenience. That is, $c_i = \frac{1}{b_i}$ for $i = 1, \ldots, n$ Therefore if $b_i = 1$, then the i th observation is not an outlier. It is considered as an outlier when $0 < b_i < 1$, since its corresponding variance is very large compared to the common variance σ^2. For detecting outliers, we use Geweke(1996)'s method which was introduced as the method for selecting the best predictors in multiple regression model using Gibbs sampler.

By introducing the latent variable $\gamma_j = 0$ or 1, we represent our normal mixture by

$$[b_j|\gamma_j] \propto (1-\gamma_j)I(b_j=1) + \gamma_j N(b_j^0, \tau_j^2)I(0 < b_j < 1) \tag{2.2}$$

where

$$\Pr(\gamma_j = 1) = 1 - \Pr(\gamma_j = 0) = p_j \tag{2.3}$$

$[\cdot]$ denotes it density and $I(\cdot)$ is an indicate function with value 1 for including in the set or 0 otherwise.

This is based on the data augmentation idea of Tanner and Wong(1987). Geweke(1996) use similar structure as in (2.2) and (2.3) for variable selection in linear regression model.

When $\gamma_j = 0$, $b_j \sim I(b_j = 1)$ which implies that the corresponding data y_j is not an outlier and if $\gamma_j = 1$, $b_j \sim N(b_j^0, \tau_j^2)I(0 < b_j < 1)$ implies that the corresponding data y_j should probably be an outlier in the given model. The choice of $f(\gamma)$ should incorporate any available prior information about which subsets of $y_1, \cdots, y_n$ should be outliers in the given model. For example, a reasonable choice might be the γ's independent with marginal distributions, so that

$$f(\gamma|p_1, \cdots, p_n) = \prod_{i=1}^{n} p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}. \tag{2.4}$$

Again (2.4) implies that the outlier y_i is independent of the outlier of y_j for all $i \neq j$, we found it to work well in the various situations. The uniform or indifference prior $f(\gamma) = 2^{-n}$ is the special case of (2.4) where each y_i has an equal chance of being an outlier. Also, it is assumed that the prior of $\beta = (\beta_1, \ldots, \beta_p)^t$ is a normal distribution with mean vector $\mu = (\mu_1, \ldots, \mu_p)$ and variance-covariance matrix Σ^{-1}, that is, $\beta \sim N(\mu, \Sigma^{-1})$ and σ^2 is

distributed as the inverse gamma conjugate prior $\sigma^2|\gamma \sim IG(\frac{\nu_\gamma}{2}, \frac{\nu_\gamma \lambda_\gamma}{2})$.

The posterior distribution may be expressed up to a constant by combining the densities defined from (2.1) through (2.4) as

$$[\beta, b, \sigma^2, \gamma | Y]$$
$$\propto \prod_{j=1}^{n} p(b_j)(\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^t C_n^{-1}(y - X\beta)\} \exp\{-\frac{1}{2}(\beta - \mu)^t \Sigma(\beta - \mu)\}$$
$$\prod_{i=1}^{n} p_i^{\gamma_i}(1 - p_i)^{1-\gamma_i} p(\sigma^2) \tag{2.5}$$

where $p(b_j)$ is the prior density for b_j as defined in (2.2) and $p(\sigma^2) = IG(\sigma^2; \frac{V_\gamma}{2}, \frac{V_\gamma \lambda_\gamma}{2})$ and $b = (b_1, \ldots, b_n)$ and $\gamma = (\gamma_1, \ldots, \gamma_n)$.

In this procedure, our main concern for embedding the normal linear model (1.1) into the hierarchical mixture model is to obtain the marginal posterior distribution $f(\gamma|Y) \propto f(Y|\gamma) f(\gamma)$, which contains the information relevant to outliers. As mentioned as before, $f(\gamma)$ may be interpreted as the statistician's prior probability that the y_i's corresponding to an non-zero components of γ should be outliers in the given model. The posterior density $f(\gamma|Y)$ updates the prior probabilities on each of the 2^n possible values of γ. Identifying each γ with a subset of data via that $\gamma_i = 1$ is equivalent to that y_i is an outlier. Those γ with higher posterior probability $f(\gamma|Y)$ identify that a subset of data is most suspicious by data and the statistician's prior information. Therefore, $f(\gamma|Y)$ provides a ranking that can be used to select a subset of the most suspicious data.

2.2. Computation

The computational procedure employed here is a Gibbs sampler. A value for each b_j is drawn in turn from its conditional on $b_l (l \neq j)$, β and σ^2 and a value for σ^2 is drawn from conditional on b and β, and the value for β is generated from the conditional on b and σ^2.

The conditional distributions involved in the algorithm are simple. Given $b_l (l \neq j)$, σ^2 and β, the likelihood function kernel of b_j is

$$p(b_j) b_j \exp\{-\frac{b_j^2 (y_j - \sum_{k=1}^{p} \beta_k x_{jk})^2}{2\sigma^2}\}. \tag{2.6}$$

Conditional on $b_j = 1$, the value of the kernel is

$$\exp\{-\frac{(y_j - \sum_{k=1}^{p} \beta_k x_{jk})^2}{2\sigma^2}\}. \tag{2.7}$$

Conditional on $b_j \neq 1$ the corresponding kernel density for b_j is

$$\begin{aligned} b_j \sim\ & b_j \exp\{-\frac{b_j^2(y_j - \sum_{k=1}^{p} \beta_k x_{jk})^2}{2\sigma^2}\}\exp\{-\frac{(b_j - b_j^0)^2}{2\tau_j^2}\} \\ & \tau_j^{-1}\{\Phi(\frac{1-b_j^0}{\tau_j}) - \Phi(\frac{-b_j^0}{\tau_j})\}^{-1} I(0 < b_j < 1) \\ =\ & b_j \exp\{-\frac{(\tau_j^2 z_j^2 + \sigma^2)(b_j - \mu_{b_j})^2}{2\sigma^2\tau_j^2}\}\exp\{-\frac{b_j^0(b_j^0 - \mu_{b_j})}{2\tau_j^2}\} \\ & \tau_j^{-1}\{\Phi(\frac{1-b_j^0}{\tau_j}) - \Phi(\frac{-b_j^0}{\tau_j})\}^{-1} I(0 < b_j < 1) \end{aligned} \tag{2.8}$$

where $z_j^2 = (y_j - \sum_{k=1}^{p} \beta_k x_{jk})^2$ and $\mu_{b_j} = \frac{\sigma^2 b_j^0}{\tau_j^2 z_j^2 + \sigma^2}$. To remove the conditioning on $b_j = 1$ or $b_j \neq 1$, it is necessary to integrate (2.8) over b_j and compare this expression to (2.7). The integration yields

$$\begin{aligned} & \int_0^1 b_j \exp\{-\frac{(\tau_j^2 z_j^2 + \sigma^2)(b_j - \mu_{b_j})^2}{2\sigma^2\tau_j^2}\} db_j \\ & \exp\{-\frac{b_j^0(b_j^0 - \mu_{b_j})}{2\tau_j^2}\}\tau_j^{-1}\{\Phi(\frac{1-b_j^0}{\tau_j}) - \Phi(\frac{-b_j^0}{\tau_j})\}^{-1} \end{aligned} \tag{2.9}$$

Note that the conditional probability of $\gamma_j = 1$, p_j^0, is proportional to $p_j \Pr(0 < b_j < 1 | b_{(-j)}\beta, \sigma^2, y)$ and that of $\gamma_j = 0$ is proportional to $(1-p_j)\Pr(1-\varepsilon < b_j < 1+\varepsilon | b_{(-j)}, \beta, \sigma^2, y)$ with $\varepsilon \to 0$. The notation $b_{(-j)}$ means that the b_l's except b_j. Thus p_j^0 can be expressed in terms of the Bayes factor for H_0 : $b_j = 1$ against H_1 : $0 < b_j < 1$ obtained from the full conditional distribution of b_j. The conditional Bayes factor in favor of $b_j \neq 1$, versus $b_j = 1$, is

$$\mathrm{BF} = \int_0^1 b_j \exp\{-\frac{(\tau_j^2 z_j^2+\sigma^2)(b_j-\mu_{b_j})^2}{2\sigma^2\tau_j^2}\}db_j \exp\{-\frac{b_j^0(b_j^0-\mu_{b_j})}{2\tau_j^2}+\frac{z_j^2}{2\sigma^2}\}$$

$$\tau_j^{-1}\{\Phi(\frac{1-b_j^0}{\tau_j})-\Phi(\frac{-b_j^0}{\tau_j})\}^{-1}. \tag{2.10}$$

Here, the numerical integration is used for computing the integration part of BF in (2.10). Recall that the Bayes factor is the ratio of posterior and prior odd. That is,

$$BF=\frac{(1-p_j^0)/p_j^0}{(1-p_j)/p_j}. \tag{2.11}$$

To draw b_j from its conditional distribution, the conditional posterior probability that $b_j=1$ is computed from the conditional Bayes factor (2.12):

$$p_j^0=\frac{p_j}{(1-p_j)BF+p_j} \tag{2.12}$$

This gives the full conditional distribution of γ_j as Bernuolli with a success probability p_j^0. That is,

$$[\gamma_j|\beta, b, \sigma^2, y]\sim Ber(p_j^0) \tag{2.13}$$

Based on the comparison of this probability with a drawing from the uniform distribution on [0,1], the choice $b_j=1$ or $b_j\neq 1$ is made. If $b_j\neq 1$ then b_j is drawn from (2.8). And then since the form of (2.8) is not standard, we can use Metropolis-Hastings algorithm (Chib and Greenberg, 1995) which need the derived function π as follows;

$$\pi(b_j)=\exp\{-\frac{(\tau_j^2 z_j^2+\sigma^2)(b_j-\mu_{b_j})^2}{2\sigma^2\tau_j^2}\}I(0<b_j<1), \tag{2.14}$$

which is the truncated normal distribution with mean μ_{b_j} and variance $\frac{\sigma^2\tau_j^2}{\tau_j^2 z_j^2+\sigma^2}$ on $I(0<b_j<1)$. Then the acceptance probability α is $\alpha=\min\{\frac{b_j^*}{b_j^{(n)}},1\}$ where b_j^* and $b_j^{(n)}$ are the candidate at the current (n+1) stage and the variate at the previous (n) stage of b_j, respectively. We further have

$$[\beta|b,\gamma,\sigma^2,y] \propto \exp\{-\frac{1}{2\sigma^2}(y-X\beta)^tC_n^{-1}(y-X\beta)\}\exp\{-\frac{1}{2}(\beta-\mu)^t\Sigma(\beta-\mu)\} \quad (2.15)$$

and

$$[\sigma^2|\beta,m,\gamma,y] = IG(\frac{n+\nu_\gamma}{2}, \frac{(y-X\beta)^tC_n^{-1}(y-X\beta)+\nu_\gamma\lambda_\gamma}{2})$$

(2.16)

3. Illustrative Example

3.1. Simulated Data for Variance-Inflation Model

Consider the simple linear regression model with the intercept term and one slope, that is,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1,\cdots,n \quad (3.1)$$

For notational convenience, let

$$\mu_1^0 = \frac{\sum_{i=1}^n b_i^2(y_i-\beta_1 x_i) + \frac{1}{2}\sigma^2(2\mu_1\sigma_{11}+\mu_2\sigma_{21}+\mu_2\sigma_{12}-\beta_1\sigma_{21}-\beta_1\sigma_{12})}{\sum_{i=1}^n b_i^2+\sigma_{11}\sigma^2}$$

and

$$\mu_2^0 = \frac{\sum_{i=1}^n b_i^2(y_i x_i-\beta_0 x_i) - \frac{1}{2}\sigma^2(\beta_0\sigma_{21}+\beta_0\sigma_{12}-\mu_1\sigma_{21}-\mu_1\sigma_{12}-2\mu_2\sigma_{22})}{\sum_{i=1}^n b_i^2 x_i^2+\sigma^2\sigma_{22}}$$

where $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$. To apply Gibbs sampler, the following full conditional densities are needed;

$$[\beta_1|\beta_2,b,\gamma,\sigma^2,y_1,\ldots,y_n] = N(\mu_1^0, \frac{\sigma^2}{\sum_{i=1}^n b_i^2+\sigma_{11}\sigma^2})$$

and

$$[\beta_2|\beta_1,b,\gamma,\sigma^2,y_1,\ldots,y_n] = N(\mu_2^0, \frac{\sigma^2}{\sum_{i=1}^n b_i^2 x_i^2+\sigma_{22}\sigma^2})$$

First we generate the data from (3.1) with $n=10$, $\varepsilon_i \sim N(0,1)$ for $i=1,\ldots,10$ and

$(\beta_0, \beta_1) = (0.5, 1)$. Furthermore, let $b_i = 1$ for $i \neq 1$ and $b_1 = 0.1$. Therefore we assume that observation 1 is outlier in this data set. We applied Geweke's method to our model with the indifference prior $f(\gamma) = (\frac{1}{2})^{10}$ $b_i^0 = 0.5$ $\tau_i^2 = 0.01$ for $i = 1, \ldots, 10$ and $\mu_1 = \mu_2 = 1$, $\nu_\gamma = \lambda_\gamma = 0$. Also, set $\sigma_{11} = \sigma_{22} = 10^{-3}$ for the diffuse prior for β. The Gibbs sampler generates 4000 iterations and Metropolis_Hastings algorithm for generating b_j is repeated 10,000 times. After discarding first 2,000 iterations, we use only the variates of remaining iterations. Convergence of the Gibbs sampler was assessed via Geweke (1992) method, using the CODA (Best, Cowles and Vines, 1995) suitable of diagnostics in S-plus. In Table 3.1, observation 1 is considered as an outlier since its corresponding posterior probability is 0.34 which is highest among these posterior probabilties of all data. Also, observations 1 and 4 may be considered as outliers since its corresponding posterior is second highest which is 0.22.

Table 3.1.

outlier numbers	observation	Posterior probability
0		0.00
1	1	0.34
2	1, 2	0.03
	1, 3	0.11
	1, 4	0.22
	1, 5	0.01
	1, 9	0.03
3	1, 2, 3	0.03
	1, 2, 4	0.03
	1, 3, 4	0.04
	1, 3, 7	0.02
4	1, 2, 3, 4	0.01
	1, 2, 4, 7	0.01

3.2. Darwin's Data

Consider the analysis of Darwin's data on the difference in heights of self- and cross-fertilized plants quoted by Box and Tiao (1973). The data consists of measurements on 15 pairs of plants. Each pair contained a self-fertilized and cross-fertilized plant grown in the same pot and from the same seed. Arranged for convenience in order of magnitude, the n=15 observations (on differences in heights in eighths of an inch of self-fertilized and cross-fertilized plants) are: -67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75. Guttman,

Dutter and Freeman(1978) re-examine the Darwin's data to detect outlier(s) using Bayesian approach with the model as follows;

$$y = \beta \mathbf{1} + \varepsilon \tag{3.2}$$

where $\mathbf{1} = (1, \ldots, 1)^t$. Like simulated data, convergence of the Gibbs sampler was checked via Geweke (1992) method, using the CODA (Best, Cowles and Vines, 1995) suitable of diagnostics in S-plus. Guttman et al (1978) mentioned that observations 1 and 2, having values -67 and -48, are identified as spurious observations since they have the highest posterior probability. Table 3.2 show that observation 1 and 2 having values -67 and -48 are considered as outliers since they have the highest posterior probability 0.54. Also, observations 1, 2 and 7 may be considered as outliers.

Table 3.2.

outlier numbers	observation	Posterior probability
0		0.00
1		0.00
2	1,2	0.54
3	1,2,7	0.39
	1,2,8	0.02
4	1,2,7,8	0.05

Reference

[1] Best, N.G., Cowles, M.K. and Vines, S.K.(1995), *Convergence Diagnosis and Output Analysis Software for Gibbs Sampler, Version 0.3,* Cambridge University, MRC Boistatistics Unit.

[2] Box, G. E. P. and Tiao, G. C. (1968), A Bayesian Approach to Some Outlier Problems, *Biometrika*, 55, 119-129.

[3] Box, G. E. P. and Tiao, G. C.(1973), *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, M.A.

[4] Chaloner, K. and Brant, R. (1988), A Bayesian Approach to Outlier Detection and Residual Analysis, *Biometrika*, 75, 651-659.

[5] Chib, S. and Greenberg, E.(1995), Understanding the Metropolis-Hasting algorithm, *American Statistician*, 49, 327-335.

[6] Chung, Y. and Kim, H. (1999), Bayesian Outlier Detection in Regression Model, *Journal of Korean Statistical Society,* 28, 311-324.

[7] Geisser, S. (1985), On the Predicting of Observables: A Selective Update, in *Bayesian*

Statistics 2, Ed. Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., 203-230, Amsterdam: North Holland.

[8] Geweke, J.(1992), Evaluating the Accuracy of sampling-Based Approaches to calculating Posterior Moments. In *Bayesian Statistics 4*, ed. J. M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford, UK; Oxford University Press, 169-193.

[9] Geweke, J.(1996), Variable selection and model comparison in regression, *Bayesian Statistics 5*, Ed. by Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M., Oxford University Press, 609-620.

[10] Guttman, I. (1973), Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity - A Bayesian Approach, *Technometrics*, 15, 4, 723-738.

[11] Guttman, I., Dutter, R. and Freeman, P. R. (1978), Care and Handling of Univariate Outliers in the General Linear Model to Detect Spuriosity - A Bayesian Approach, *Technometrics*, 20, 2, 187-193.

[12] Guttman, I. and Pena, D. (1993), A Bayesian Look at Diagnostics in the Univariate Linear Model, *Statistical Sinica*, 3, 367-390.

[13] Pettit, L. I. and Smith, A. F. M. (1985), Outliers and Influential Observations in Linear Models, in *Bayesian Statistics 2*, Ed. Bernardo, J. M., DeGroot, M. H., Lindley, D. V. and Smith, A. F. M., 473-494, Amsterdam: Elsevier.

[14] Sharples, L. D. (1990) Identification and Accommodation of Outliers in General Hierarchical Models, *Biometrika*, 77, 3, 445-453.

[15] Tanner, M. and Wong, W. (1987), The Calculation of Posterior Distributions by Data Augmentation (with discussion), *Journal of the American Statistical Association*, 82, 528-550.