Communications for Statistical Applications and Methods
2019, Vol. 26, No. 2, 149–161

https://doi.org/10.29220/CSAM.2019.26.2.149
Print ISSN 2287-7843 / Online ISSN 2383-4757

Simultaneous outlier detection and variable selection via difference-based regression model and stochastic search variable selection

Jong Suk Park[a], Chun Gun Park[b], Kyeong Eun Lee[1,a]

[a]Department of Statistics, Kyungpook National University, Korea;
[b]Department of Mathematics, Kyonggi University, Korea

Abstract

In this article, we suggest the following approaches to simultaneous variable selection and outlier detection. First, we determine possible candidates for outliers using properties of an intercept estimator in a difference-based regression model, and the information of outliers is reflected in the multiple regression model adding mean shift parameters. Second, we select the best model from the model including the outlier candidates as predictors using stochastic search variable selection. Finally, we evaluate our method using simulations and real data analysis to yield promising results. In addition, we need to develop our method to make robust estimates. We will also to the nonparametric regression model for simultaneous outlier detection and variable selection.

Keywords: Bayesian variable selection, difference-based regression model, mean-shift outlier model, stochastic search variable selection

1. Introduction

In multiple linear regression models, the separated data point on the vertical axis is called an outlier that is different from the others in the data (Weisberg, 2004). These outliers have serious effects in inference and model selection (Kahng *et al.*, 2016). The selection of predictors is a crucial problem in building a multiple linear regression model (George and McCulloch, 1993).

Many approaches for the detection of outliers or selection of variables have been proposed. Most authors have focused on these problems separately; however, some authors proposed methods to perform outlier detection and variable selection simultaneously. For example, Hoeting *et al.* (1996) considered the model with all variables and outlier candidates which are determined using the least median of squares regression (Rousseeuw, 1984) and computed posterior probabilities for all possible subset models using Markov chain Monte Carlo approach. This approach is efficient in detecting masked outliers and also selecting variables (Kim *et al.*, 2008).

After that Kim *et al.* (2008) proposed two similar steps, but the details are different from the approach proposed by Hoeting *et al.* (1996). The first step is to determine outlier candidates using a multiple outlier identification procedure, and the second step is to apply all possible subset regressions of the mean-shift outlier models to select the best model. Compared with the Frequentist model selection methods, the Bayesian model selection methods have several advantages, the major one of which exists in its ability to incorporate prior knowledge into the selection process. In order to take

[1] Corresponding author: Department of Statistics, Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566, Korea. E-mail: artlee@knu.ac.kr

Published 31 March 2019 / journal homepage: http://csam.or.kr

this advantage, we provide an alternative approach for simultaneous variable selection and outlier detection using stochastic search variable selection (SSVS) (George and McCulloch, 1993) in a multiple regression model.

We use the mean-shift outlier model for outlier candidates. To determine these outlier candidates, we use the properties of an intercept estimator in the difference-based regression model (DBRM) (Choi *et al.*, 2018; Park and Kim, 2018b). This type of model was originally used in a time series analysis to remove trends in the mean function (Park and Kim, 2018a; Park, 2018; Park *et al.*, 2012). This method has advantages of good performance and simple usage that uses only the intercept and does not need to estimate the mean function.

The remainder of this paper is organized as follows. In Section 2, after introducing the notation, we describe the mean-shift outlier model and the difference-based regression model (Park and Kim, 2018b) to determine the outlier candidates. In Section 3, we introduce the Bayesian variable selection, SSVS proposed by George and McCulloch (1993). Then, we propose a method that can simultaneously perform outlier detection and variable selection by using SSVS in regression model with outlier candidates. In Sections 4 and 5, we present simulations and a real data example, respectively. Finally, we provide conclusions and recommendations in Section 6.

2. Determining outlier candidates

In this section, we explain how to determine outlier candidates and set up a regression model that includes outlier candidates. We then introduce the difference-based regression model (Choi *et al.*, 2018; Park and Kim, 2018b) in Section 2.2.

2.1. Linear model with information about outliers

Let $\mathbf{X} = [\mathbf{1}_n : \mathbf{X}_1]$ be an $n \times (p+1)$ matrix, with rank$(\mathbf{X}) = p + 1$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ be a $(p \times 1)$ coefficient vector where β_0 is an intercept. Assume that the response vector is $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$.

We consider the mean-shift outlier model (Belsley *et al.*, 1980):

$$\mathbf{Y} = \mathbf{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\gamma} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right), \tag{2.1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)'$ is an $n \times 1$ vector and γ_j is nonzero only when the j^{th} observation is an outlier.

In this paper, we use the mean-shift outlier model described above in order to include potential q $(0 \leq q < n)$ outliers, instead of using the entire n's. To identify potential outliers, we use the properties of an intercept estimator in the DBRM (Park and Kim, 2018b). Since the variance of an outlier indicator variable is $(n-1)/n^2 \approx 1/n$, we adjust the weight of them as $\sqrt{n}$. Without loss of generality, the dataset is sorted by the order of magnitude of the absolute intercept estimators in DBRM. So we denote

$$\mathbf{Z} = \left(\mathbf{X} : \sqrt{n}\mathbf{I}_{n\times q}\right),$$

where $\mathbf{I}_{n\times q}$ is the submatrix with the first q columns of $\mathbf{I}_n$.

In this paper, we use the mean-shift outlier model described above in order to include potential q $(0 \leq q < n)$ outliers, instead of using the entire n's. To identify potential outliers, we use the properties of an intercept estimator in the DBRM (Park and Kim, 2018b). Then, we can rewrite the model under the above assumptions:

$$\begin{aligned}\mathbf{Y} &= \sum_{j=1}^{p+q+1} \mathbf{z}_j\theta_j + \boldsymbol{\epsilon} \\ &= \mathbf{Z}_{n\times(p+q+1)}\boldsymbol{\theta}_{(p+q+1)\times 1} + \boldsymbol{\epsilon},\end{aligned} \tag{2.2}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. Here, the new model includes $p + q + 1$ parameters, and $\text{rank}(\mathbf{Z}) = p + q + 1$. The first $(p + 1)$ of θ_j's come from the β_j's, $j = 0, 1, \ldots, p$, and the other q coefficients indicate the effects of the outlier candidates.

2.2. Difference-based regression model

Park and Kim (2018b) propose an outlier-detection approach that uses the properties of an intercept estimator in the difference-based regression model. This method uses only the estimated intercepts: it does not require estimating the other parameters in the DBRM. To identify whether if the observations are outliers, the DBRM uses a mean-shift outlier model.

In this paper, we describe the DBRM defined by Park and Kim (2018b) using Equation (2.1). Assume that $\mathbf{Y}_{(i)}$ and $\mathbf{X}_{1(i)}$ are $\mathbf{Y}$ and $\mathbf{X}_1$ without the i^{th} row for $i = 1, 2, \ldots, n$. And $\mathbf{D}_{(i)}\mathbf{Y}$ is the difference between $\mathbf{Y}_{(i)}$ and $y_i\mathbf{1}_{n-1}$. Then the difference-based regression model can be written as follows:

$$\begin{aligned}\mathbf{D}_{(i)}\mathbf{Y} &= \mathbf{1}_{n-1}(-\gamma_i) + \mathbf{D}_{(i)}\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{A}_{(i)}\boldsymbol{\gamma} + \mathbf{D}_{(i)}\boldsymbol{\epsilon} \\ &= [\mathbf{1}_{n-1} : \mathbf{D}_{(i)}\mathbf{X}_1]\begin{pmatrix}-\gamma_i \\ \boldsymbol{\beta}_1\end{pmatrix} + \mathbf{A}_{(i)}\boldsymbol{\gamma} + \mathbf{D}_{(i)}\boldsymbol{\epsilon}, \quad i = 1, \ldots, n,\end{aligned} \tag{2.3}$$

where $-\gamma_i$ is the intercept and $\mathbf{D}_{(i)}$ and $\mathbf{A}_{(i)}$ are the $(n-1) \times n$ matrices as follows:

$$\mathbf{D}_{(i)} = \begin{pmatrix}\mathbf{I}_{i-1} & -\mathbf{1}_{i-1} & \mathbf{0}_{(i-1),(n-i)} \\ \mathbf{0}_{(n-i),(i-1)} & -\mathbf{1}_{n-i} & \mathbf{I}_{n-i}\end{pmatrix} \quad \text{and} \quad \mathbf{A}_{(i)} = \begin{pmatrix}\mathbf{I}_{i-1} & \mathbf{0}_{i-1} & \mathbf{0}_{(i-1),(n-i)} \\ \mathbf{0}_{(n-i),(i-1)} & \mathbf{0}_{n-i} & \mathbf{I}_{n-i}\end{pmatrix},$$

where $\mathbf{I}_a$ is the $a \times a$ identity matrix and $\mathbf{0}_{a,b}$ is the $a \times b$ null matrix.

Park and Kim (2018b) estimate intercepts, $-\gamma_i$ in the DBRM (2.3) using the least square estimators. Then estimated intercepts are as follows:

$$-\hat{\gamma}_i = \begin{cases}\dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{D}_{(i)}\boldsymbol{\epsilon}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{1}_{n-1}}, & \text{no outlier,} \\ \dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{A}_{(i)}\boldsymbol{\gamma}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{1}_{n-1}} + \dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{D}_{(j)}\boldsymbol{\epsilon}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(i)})\mathbf{1}_{n-1}}, & \text{several outliers,}\end{cases} \tag{2.4}$$

where the i^{th} observation is not an outlier and $\mathbf{H}_{(i)} = \mathbf{D}_{(i)}\mathbf{X}_1(\mathbf{X}'_1\mathbf{D}'_{(i)}\mathbf{D}_{(i)}\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{D}'_{(i)}$.

$$-\hat{\gamma}_k = \begin{cases}-\gamma_k + \dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{D}_{(k)}\boldsymbol{\epsilon}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{1}_{n-1}}, & \text{one outlier,} \\ -\gamma_k + \dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{A}_{(k)}\boldsymbol{\gamma}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{1}_{n-1}} + \dfrac{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{D}_{(k)}\boldsymbol{\epsilon}}{\mathbf{1}'_{n-1}(\mathbf{I}_{n-1} - \mathbf{H}_{(k)})\mathbf{1}_{n-1}}, & \text{several outliers,}\end{cases} \tag{2.5}$$

where the k^{th} observation is an outlier and $\mathbf{H}_{(k)} = \mathbf{D}_{(k)}\mathbf{X}_1(\mathbf{X}'_1\mathbf{D}'_{(k)}\mathbf{D}_{(k)}\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{D}'_{(k)}$.

The i^{th} OLS intercept estimator is highly affected by the i^{th} outlier effect (Park and Kim, 2018b). This means that the intercept estimators corresponding to the observations including the outlier effect are large values. Subsequently we can roughly discriminate outliers among the observations. To do this, we estimate the absolute magnitude of the OLS intercept estimators and regard some observations corresponding to the large absolute intercept estimates as possible outliers using a triangular outlier-detection approach proposed by Park and Kim (2018b). This process is determined as follows:

(1) $\mathbf{D}_{(i)}\mathbf{Y} = \mathbf{Y}_{(i)} - y_i\mathbf{1}_{n-1} = \mathbf{1}_{n-1}(-\gamma_i) + \mathbf{D}_{(i)}\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{A}_{(i)}\boldsymbol{\gamma} + \mathbf{D}_{(i)}\boldsymbol{\epsilon}$, for $i = 1, \ldots, n$;

(2) Estimate intercepts, $-\hat{\gamma}_i$, $i = 1, \ldots, n$, and rewrite $\gamma_i^* = abs(-\hat{\gamma}_i)$;

(3) Ascend them, γ_i^*, $\gamma_{(1)}^* \leq \gamma_{(2)}^* \leq \cdots \leq \gamma_{(n)}^*$.

We assume that q is the greatest integer less than one fifth of the number of data and the value of q can be adjusted by looking at the results of the DBRM. That is, the candidate outliers are set by a rapid change point of the estimated intercepts trend.

3. Simultaneous outlier detection and variable selection

In this section, we briefly describe the Bayesian variable selection and introduce stochastic search variable selection (George and McCulloch, 1993). Finally, we explain how to simultaneously detect outliers and select variables for the regression model (2.2) which includes independent variables as the outlier effect.

3.1. Stochastic search variable selection

George and McCulloch (1993, 1997) assumed that the prior distributions of the regression coefficients are independent and expressed a prior distribution of each coefficient to be a mixture of two normal distributions. Both normal distributions are centered at 0 with one being a very small variance and the other being a very large variance.

Then we can describe stochastic search variable selection using equation (2.2). The prior distribution of the coefficient θ_j given the indicator δ_j is

$$\theta_j|\delta_j \sim (1-\delta_j)N\left(0, \tau_j^2\right) + \delta_j N\left(0, c_j^2\tau_j^2\right), \quad j = 1, 2, \ldots, (p+q+1). \tag{3.1}$$

The value of τ_j^2 is set to be small, and $N(0, \tau_j^2)$ is the prior distribution of the coefficient θ_j, if the variable $\mathbf{z}_j$ is not selected. The value of c_j is set to be large ($c_j > 1$) so that if $\delta_j = 1$, then a non-zero estimate of θ_j should be included in the final model. The prior distribution of δ_j is $P(\delta_j = 1) = 1 - P(\delta_j = 0) = p_j$. The prior distributions of (δ_j, θ_j) is independent from the prior distribution of σ^2, which is an inverse gamma (IG) distribution, IG$(\nu_\delta/2, \nu_\delta\lambda_\delta/2)$.

3.2. Simultaneous outlier detection and variable selection

Our approach for simultaneous variable selection and outlier detection consists of two steps: the first step is to determine a set of outlier candidates using properties of an intercept in the DBRM described in Section 2.2. The second step is to perform SSVS of the mean-shift outlier model with the outlier candidates detected in step 1. Our approach is similar to that of Hoeting *et al.* (1996) and Kim *et al.* (2008), but different in determination of outlier candidates and the variable selection method.

To simultaneously perform outlier detection and variable selection, we use hierarchical model of equation (2.2), and as the prior for θ_j conditional on δ_j and τ_j^2 use mixture of two normal densities, which can be written as

$$\theta_j|\delta_j, \tau_j^2 \sim (1-\delta_j)N\left(0, \tau_j^2\right) + \delta_j N\left(0, c^2\tau_j^2\right), \quad j = 1, 2, \ldots, (p+q+1), \tag{3.2}$$

where c^2 and τ_j^2 are variance components. We assume an inverse gamma prior for σ^2 and τ_j^2 and that δ_j is distributed as Bernoulli with inclusion probability p_j, $j = 1, 2, \ldots, (p+q+1)$. Thus, we have the following multilevel model:

$$\mathbf{Y}|\boldsymbol{\theta}, \sigma^2 \sim N\left(\mathbf{Z}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n\right), \tag{3.3}$$

$$\boldsymbol{\theta}|\boldsymbol{\delta}, \boldsymbol{\tau}^2 \sim N_{p+q+1}\left(\mathbf{0}, \mathbf{D}_{\delta\tau}\mathbf{R}\mathbf{D}_{\delta\tau}\right), \tag{3.4}$$

$$\delta_j \overset{\text{ind}}{\sim} \text{Bernoulli}(p_j), \tag{3.5}$$

$$\sigma^2 \sim \text{IG}\left(\frac{a_1}{2}, \frac{b_1}{2}\right), \tag{3.6}$$

$$\tau_j^2 \sim \text{IG}\left(\frac{a_2}{2}, \frac{b_2}{2}\right), \tag{3.7}$$

where $\mathbf{R}$ is the prior correlation matrix, and we assume $\mathbf{R} = \mathbf{I}$. Also, $\mathbf{D}_{\delta\tau} = \text{diag}[d_1\tau_1, \ldots, d_{p+q+1}\tau_{p+q+1}]$ where $d_j^2 = 1$ if $\delta_j = 0$ and $d_j^2 = c^2$ if $\delta_j = 1$.

Using conjugate priors, it is easy to obtain posterior distributions. Thus Gibbs sampling procedures are easily implemented for calculating the posterior distributions. Accordingly, the posterior distributions as follows:

$$\begin{aligned}
\boldsymbol{\theta}|\mathbf{Y}, \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\delta} &\sim N\left(\frac{1}{\sigma^2}\mathbf{A}_{\delta\tau}\mathbf{Z}'\mathbf{Y}, \mathbf{A}_{\delta\tau}\right),\\
\sigma^2|\mathbf{Y}, \boldsymbol{\tau}^2, \boldsymbol{\delta}, \boldsymbol{\theta} &\sim \text{IG}\left(\frac{n+a_1}{2}, \frac{1}{2}[(\mathbf{Y}-\mathbf{Z}\boldsymbol{\theta})'(\mathbf{Y}-\mathbf{Z}\boldsymbol{\theta}) + b_1]\right),\\
\tau_j^2|\mathbf{Y}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\delta} &\sim \text{IG}\left(\frac{a_2+1}{2}, \frac{\theta_j^2/d_j^2 + b_2}{2}\right),\\
\delta_j|\boldsymbol{\delta}_{(j)}, \boldsymbol{\theta}, \boldsymbol{\tau}^2 &\sim \text{Bernoulli}\left(\hat{\delta}_j\right)
\end{aligned}$$

where $\mathbf{A}_{\delta\tau} = (\sigma^{-2}\mathbf{Z}'\mathbf{Z} + (\mathbf{D}_{\delta\tau}\mathbf{R}\mathbf{D}_{\delta\tau})^{-1})^{-1}$, $\boldsymbol{\delta}_{(j)} = (\delta_1, \delta_2, \ldots, \delta_{j-1}, \delta_{j+1}, \ldots, \delta_{p+q+1})'$, $d_j^2 = 1$ if $\delta_j = 0$ and $d_j^2 = c^2$ if $\delta_j = 1$, and

$$\hat{\delta}_j = \frac{P(\theta_j|\delta_j = 1)p_j}{P(\theta_j|\delta_j = 1)p_j + P(\theta_j|\delta_j = 0)(1-p_j)}.$$

Therefore the best subset of variables is selected according to the information contained in the $\boldsymbol{\delta}$. The posterior probability $p(\delta_j = 1|\mathbf{Y})$ for the regressor $\mathbf{z}_j$ with $j = 1, \ldots, p+q+1$ to be included in the model can be estimated by the mean of $\hat{\delta}_j$ or alternatively by the mean of $p(\delta_j^{(m)} = 1|\mathbf{Y})$.

4. Simulation studies

We conduct simulations to evaluate the performance of our approach compared to that of another existing approach. We compare our approach with BayesVarSel. Therefore, we introduce BayesVarSel before simulation.

Table 1: Priors used in BayesVarSel package

(a) Prior probabilities for models	- prior.models = "ScottBerger" (default) - prior.models = "Constant" - prior.models = "User", priorprobs
(b) Prior probabilities for the coefficients	- prior.betas = "Robust" (default) - prior.betas = "ZellnerSiow" - prior.betas = "gZellner" - prior.betas = "FLS" - prior.betas = "Liangetal"
(c) Null model contains just the intercept	- fixed.cov = c("Intercept") (default) - fixed.cov = NULL

Donato and Forte (2017) introduce the R package, BayesVarSel which implements Bayesian methodology for hypothesis testing and variable selection in linear models. To perform the simulation compared to our method, we will use the variable selection in this package. The variable selection functions in this package are Bvs, PBvs, and GibbsBvs. Except for a few arguments the usage of the three functions is very similar.

This package implements the criteria-based priors of the regression coefficients proposed by Bayarri *et al.* (2012), but the advanced user has the possibility of using several other popular priors in the literature. These priors are shown in Table 1.

4.1. Simulation setting

We consider two cases of multiple regression: $p = 4$ and $p = 9$ to demonstrate the performance of the proposed method. To study the behavior of selection of regressors, the coefficient vectors are sets to different $\boldsymbol{\beta}_1$s in (2.1), $\boldsymbol{\beta}_1 = (1, 1, 0, 0)'$ and $\boldsymbol{\beta}_1 = (1, 1, 1, 0, 0, 0, 0, 0, 0)'$. Also, we consider three different samples sizes ($n = 30, 50, 100$), each with 10% randomly assigned outliers whose size is randomly determined to be 7 or 10 and whose sign is + or −. We set model matrix $\mathbf{X} = \{x_{ij}\} \sim N(\mathbf{0}, 3^2\mathbf{C})$, where $\mathbf{C} = \{\rho^{|i-j|}\}$, $\rho = 0.5$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. In addition, we generate errors using $N(0, 1)$, and 100 data sets for each case.

We apply SSVS using the hyperparameter settings: prior correlation matrix $\mathbf{R} \equiv \mathbf{I}$, and $\boldsymbol{\delta}$ prior (3.5) with $p_j = 0.5$ which yields $\pi(\boldsymbol{\delta}) \equiv 1/2^{(p+q+1)}$. We use three different values of $c_j^2 = c^2$, $c^2 = 10, 25, 100$. For τ_j^2, we consider two cases: $\tau_{\boldsymbol{\beta}}^2 = \tau_{\gamma}^2 (= \tau^2)$ or $\tau_{\boldsymbol{\beta}}^2 \neq \tau_{\gamma}^2$ where $\tau_{\boldsymbol{\beta}}^2$ denotes common prior variance for regression coefficients and τ_{γ}^2 denotes common prior variance for outliers. For each setting, initial values are randomly generated from their prior distributions discussed in Section 3.2.

In all cases, 30,000 samples from the MCMC simulation are used to estimate the parameters, where the first 10,000 samples are discarded as burn-in. In addition, we confirm the convergence of the Markov chain by using Gelman-Rubin diagnostic (Gelman and Rubin, 1992); all the values are close to one.

For each case, we compare our approach with Bvs ($p + q + 1 < 20$) and GibbsBvs ($p + q + 1 \geq 20$) function in BayesVarSel. For choosing prior probabilities of models and $\boldsymbol{\theta}$, we use two prior probabilities of models (ScottBerger, Constant) and five prior probabilities of $\boldsymbol{\theta}$ in Table 1. We use default values in the package for other conditions.

4.2. Criteria

The performances of our proposed procedure are evaluated in two parts: outlier detection and variable selection. In the first part, we use three criteria proposed by Choi *et al.* (2018) to detect outliers.

Table 2: Criteria of outlier detection

		True		Sum
		Outlier	Non-outlier	
Detection	Outlier	n_{CD}	n_{ID}	n_D
	Non-outlier	n_{IU}	n_{CU}	n_{UD}
	Sum	n_O	n_{NO}	n

Let n_O be the number of true outliers, n_D be the number of detected outliers, n_{CD} be the number of correctly detected outliers, n_{ID} be the number of incorrectly detected outliers, n_{IU} be the number of incorrectly undetected outliers, and n_{CU} be the number of correctly undetected non-outliers (Table 2). Then, we use the relative frequency of perfect detection (PD), the relative frequency of only-swamping with detection (overdetection) (OS), and the average number of detected outliers (AN) to compare performance.

$$\mathrm{PD} = \frac{1}{n_{\mathrm{sim}}} \sum_{s=1}^{n_{\mathrm{sim}}} I\left(n_{\mathrm{CD}(s)} = n_{O(s)} = n_{D(s)}\right),$$

$$\mathrm{OS} = \frac{1}{n_{\mathrm{sim}}} \sum_{s=1}^{n_{\mathrm{sim}}} I\left(n_{\mathrm{IU}(s)} = 0, n_{\mathrm{ID}(s)} > 0, n_{\mathrm{CD}(s)} > 0\right),$$

$$\mathrm{AN} = \frac{1}{n_{\mathrm{sim}}} \sum_{s=1}^{n_{\mathrm{sim}}} \left(\#\left\{j : \hat{\gamma}_{j,s} \neq 0\right\}\right).$$

In the second part, to select variables, we use two criteria (Choi, *et al.*, 2018). Let CS be the relative frequency of correct selection, the AN be the average number of selected variables.

$$\mathrm{CS} = \frac{1}{n_{\mathrm{sim}}} \sum_{s=1}^{n_{\mathrm{sim}}} I\left(\left\{j : \hat{\beta}_{j,s} \neq 0\right\} = \left\{j : \beta_j \neq 0\right\}\right),$$

$$\mathrm{AN} = \frac{1}{n_{\mathrm{sim}}} \sum_{s=1}^{n_{\mathrm{sim}}} \left(\#\left\{j : \hat{\beta}_{j,s} \neq 0\right\}\right).$$

Also, to compare the performance of our approach and BayesVarSel, we consider the following two models: the model with the highest probability model (HPM) and the model consisting of variables with inclusion probability greater than 0.5 (MPM) (Barbieri and Berger, 2004).

4.3. Simulation results

The assignment of prior probabilities $\pi(\boldsymbol{\delta})$ of BayesVarSel, is "Constant", which stands for $\pi(\boldsymbol{\delta}) = 1/2^{p+q}$ and intercept (β_0) presents in all models. Because the results of MPM similar to the results of HPM, we present the results of HPM in a tabular form. The results over the simulated 100 data sets are summarized in Table 3 and Table 4. Except for few cases discussed below, simulation results show that the SSVS performs better than the BayesVarSel regardless of sample sizes. Our method using SSVS shows the best results at $c^2 = 10$ and $\tau^2_\beta = \tau^2_\gamma$. In contrast, the method using BayesVarSel shows different results depending on the prior probability of $\boldsymbol{\theta}$. For prior probabilities for models in Table 1, "contrast" is better than "ScottBerger". For prior probabilities for the coefficients, gZellner and FLS are better than the other three priors of $\boldsymbol{\theta}$.

With regard to variable selection, Table 3 ($p = 4$) indicates that the CS of our method is better than that of BayesVarSel, and the AN of our method tends to be closer to the number of true variables.

Table 3: Results on simulated data with $p = 4$ and HPM on 100 replicates

n	Method			Variable selection				Outlier detection				
				CS	AN	Max	Mode	PD	PD + OS	AN	Max	Mode
30	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.90	2.13	4	2	0.88	0.99	3.12	5	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.88	2.15	4	2	0.76	0.98	3.26	5	3
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.91	2.07	4	2	0.93	0.99	3.08	6	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.86	2.20	4	2	0.88	0.99	3.16	6	3
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.93	2.07	4	2	0.98	0.99	3.00	4	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.90	2.13	4	2	0.93	0.99	3.08	6	3
	BayesVarSel	FLS		0.91	2.10	4	2	0.79	0.99	3.22	5	3
		gZellner		0.93	2.08	4	2	0.82	0.99	3.18	5	3
		Liangetal		0.91	2.10	4	2	0.80	0.99	3.21	5	3
		Robust (default)		0.91	2.10	4	2	0.76	0.99	3.26	5	3
		ZellnerSiow		0.91	2.10	4	2	0.80	0.99	3.21	5	3
50	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.94	2.07	4	2	0.69	0.95	5.30	8	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.90	2.13	4	2	0.53	0.95	5.58	9	5
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.95	2.05	3	2	0.96	1.00	5.04	6	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.93	2.08	4	2	0.91	0.99	5.10	7	5
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.97	2.02	4	2	0.95	0.95	4.95	5	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.95	2.05	3	2	0.92	0.95	4.98	6	5
	BayesVarSel	FLS		0.95	2.05	3	2	0.49	0.95	5.64	9	5
		gZellner		0.94	2.07	4	2	0.49	0.95	5.64	9	5
		Liangetal		0.94	2.07	4	2	0.48	0.95	5.66	9	5
		Robust (default)		0.94	2.07	4	2	0.46	0.95	5.70	9	5
		ZellnerSiow		0.94	2.07	4	2	0.47	0.95	5.67	9	5
100	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.93	2.08	4	2	0.70	0.99	10.31	12	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.96	2.04	3	2	0.65	0.99	10.41	12	10
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.97	2.03	3	2	0.93	0.97	10.02	11	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.99	2.01	3	2	0.92	0.97	10.03	11	10
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.96	2.04	3	2	0.99	0.99	9.99	10	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.97	2.04	4	2	0.98	0.99	10.09	20	10
	BayesVarSel	FLS		0.97	2.03	3	2	0.41	0.99	10.95	14	10
		gZellner		0.89	2.11	3	2	0.21	0.99	11.60	16	11
		Liangetal		0.90	2.10	3	2	0.27	0.99	11.42	16	10
		Robust (default)		0.90	2.10	3	2	0.25	0.99	11.51	16	11
		ZellnerSiow		0.90	2.10	3	2	0.27	0.99	11.42	16	10

With regard to outlier detection, when $c^2 = 10$ and $\tau_\beta^2 = \tau_\gamma^2$, our method performs best (PD is over 0.95 and PD + OS $\approx$ 1). However, BayesVarSel tends to overdetect outliers more than our method. The results of $p = 9$ (Table 4) are similar to the results of $p = 4$ (Table 3).

We now examine one data set in order to show the detailed procedures which lead into the detailed results. Consider the following model: $n = 50$, $p = 4$, and $\boldsymbol{\beta}_1 = (1, 1, 0, 0)'$. To identify outlier candidates, we calculate the absolute values of the intercept estimators (γ_i^*) and sort them in order of their magnitude. Up to 20% of the total data is determined to be outlier candidates. Figure 1 displays the estimated intercepts.

Accordingly, the dataset ($\mathbf{Y}$ and $\mathbf{Z}$) is sorted by the order of γ_i^* and we find the data of the numbers 13, 25, 33, 2, 15, 24, 44, 39, 32, and 50 to be outlier candidates in the final model. We then simultaneously perform outlier detection and variable selection by using SSVS with $c^2 = 10$ and $\tau_\beta^2 = \tau_\gamma^2$. Table 5 and Table 6 summarize high frequency models and the estimation results on simulated data. In conclusion, we can select the model that includes variable β_1, β_2 and determine that observations $(2, 13, 15, 25, 33)$ are multiple outliers.

Table 4: Results on simulated data with $p = 9$ and HPM on 100 replicates

n	Method			Variable selection				Outlier detection				
				CS	AN	Max	Mode	PD	PD + OS	AN	Max	Mode
30	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.73	3.44	7	3	0.73	0.91	3.14	6	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.71	3.47	6	3	0.55	0.91	3.40	6	3
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.70	3.27	7	3	0.81	0.87	2.94	4	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.66	3.47	6	3	0.76	0.87	3.09	6	3
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.78	3.20	6	3	0.89	0.91	2.94	4	3
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.70	3.48	9	3	0.85	0.90	3.00	4	3
	BayesVarSel	FLS		0.82	3.12	7	3	0.66	0.91	3.20	6	3
		gZellner		0.88	3.01	4	3	0.82	0.91	3.02	5	3
		Liangetal		0.84	3.06	5	3	0.68	0.91	3.18	6	3
		Robust (default)		0.81	3.14	7	3	0.66	0.91	3.22	6	3
		ZellnerSiow		0.84	3.06	5	3	0.68	0.91	3.19	6	3
50	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.87	3.20	6	3	0.79	0.94	5.10	7	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.77	3.37	8	3	0.57	0.94	5.43	7	5
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.87	3.17	6	3	0.92	0.93	4.95	6	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.79	3.28	6	3	0.90	0.93	5.00	7	5
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.89	3.16	6	3	0.94	0.94	4.93	5	5
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.87	3.23	7	3	0.92	0.94	4.95	6	5
	BayesVarSel	FLS		0.88	3.10	5	3	0.64	0.94	5.27	7	5
		gZellner		0.90	3.10	6	3	0.66	0.94	5.26	7	5
		Liangetal		0.88	3.12	6	3	0.62	0.94	5.31	7	5
		Robust (default)		0.85	3.15	6	3	0.62	0.94	5.32	7	5
		ZellnerSiow		0.88	3.12	6	3	0.62	0.94	5.31	7	5
100	SSVS	$c^2 = 100$	$\tau_\beta^2 = \tau_\gamma^2$	0.89	3.14	5	3	0.73	0.96	10.21	13	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.82	3.29	9	3	0.58	0.95	10.35	12	10
		$c^2 = 25$	$\tau_\beta^2 = \tau_\gamma^2$	0.87	3.20	6	3	0.95	0.97	9.99	11	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.91	3.13	6	3	0.95	0.97	9.99	11	10
		$c^2 = 10$	$\tau_\beta^2 = \tau_\gamma^2$	0.92	3.09	5	3	0.96	0.96	9.96	10	10
			$\tau_\beta^2 \neq \tau_\gamma^2$	0.94	3.07	5	3	0.95	0.96	9.97	11	10
	BayesVarSel	FLS		0.92	3.08	4	3	0.45	0.96	10.73	15	10
		gZellner		0.85	3.17	5	3	0.31	0.96	11.08	15	11
		Liangetal		0.85	3.17	5	3	0.31	0.96	11.09	15	11
		Robust (default)		0.85	3.17	5	3	0.31	0.96	11.10	15	11
		ZellnerSiow		0.85	3.17	5	3	0.31	0.96	11.09	15	11

5. Real data analysis

This section illustrate the performance of our method on Scottish Hill Racing data (Atkinson, 1986). This data set is used by Hoeting *et al.* (1996), Kim *et al.* (2008) and Menjoge and Welsch (2010) to evaluate their respective methods. This data contains the record-winning times for 35 hill races in Scotland and two independent variables (distance and climb).

To identify outlier candidates, we calculate the absolute values of the intercept estimators (γ_i^*) and sort them in order of their magnitude. Up to 20% of the data is determined to be the outlier candidates. Figure 2 displays the estimated intercepts in our example data.

Accordingly, the dataset is sorted by the order of γ_i^* and we find the data of the numbers 11, 18, 31, 26, 33, 17, and 5 to be outlier candidates in the final model. Then, we simultaneously perform outlier detection and variable selection by using SSVS with $c^2 = 10$ and $\tau_\beta^2 = \tau_\gamma^2$.

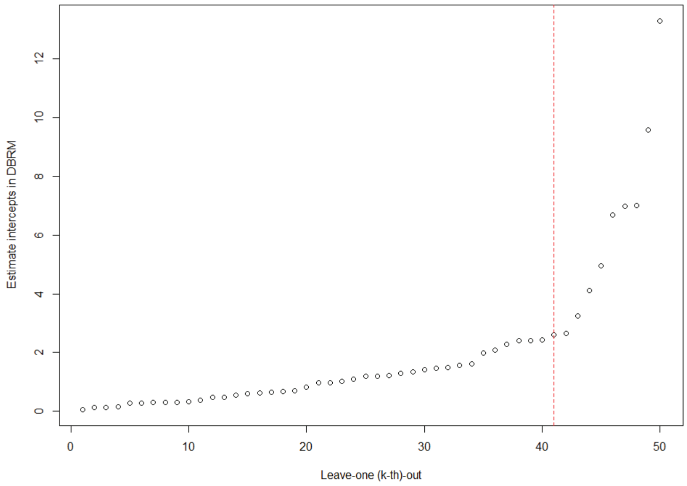

Figure 1: *Intercept estimates in DBRM on simulated data with* $n = 50$, $p = 4$, $\boldsymbol{\beta}_1 = (1, 1, 0, 0)'$, *true outlier location* $= (2, 13, 15, 25, 33)$, *and size of true outlier* $= (10, 7, 10, 7, 7)$.

Table 5: High frequency models simulated data using SSVS

Model	Index set		Probability
	Model selection	Outlier detection	
1	$\{\beta_1, \beta_2\}$	$\{\gamma_{13}, \gamma_{25}, \gamma_{33}, \gamma_2, \gamma_{15}\}$	0.99965
2	$\{\beta_1\}$	$\{\gamma_{13}, \gamma_{25}, \gamma_{33}, \gamma_2, \gamma_{15}\}$	0.00005
3	$\{\beta_1, \beta_2\}$	$\{\gamma_{13}, \gamma_{25}, \gamma_2, \gamma_{15}\}$	0.00010
4	$\{\beta_2\}$	$\{\gamma_{13}, \gamma_{25}, \gamma_{33}, \gamma_2, \gamma_{15}\}$	0.00010
5	$\{\beta_1, \beta_2\}$	$\{\gamma_{25}, \gamma_{33}, \gamma_2, \gamma_{15}\}$	0.00005
6	$\{\beta_1, \beta_2\}$	$\{\gamma_{13}, \gamma_{25}, \gamma_{33}, \gamma_2, \}$	0.00005

Table 6: Estimation results simulated data using SSVS

Parameter		Quantile (95%)		Median	Mean	sd	Inclusion probability
		2.5%	97.5%				
Variable selection	β_1	0.816	1.589	1.201	1.202	0.196	0.9999
	β_2	1.175	2.067	1.623	1.623	0.227	0.99995
	β_3	−0.071	0.820	0.354	0.359	0.227	0
	β_4	−0.487	0.227	−0.121	−0.122	0.182	0
Outlier detection	γ_{13}	−1.149	−0.639	−0.896	−0.895	0.129	0.99995
	γ_{25}	−1.180	−0.681	−0.932	−0.932	0.125	1
	γ_{33}	−1.173	−0.682	−0.927	−0.928	0.125	0.9999
	γ_2	−1.612	−1.123	−1.368	−1.368	0.124	1
	γ_{15}	1.140	1.691	1.418	1.418	0.140	0.99995
	γ_{24}	−0.151	0.316	0.085	0.084	0.119	0
	γ_{44}	−0.097	0.397	0.154	0.153	0.126	0
	γ_{39}	−0.433	0.068	−0.185	−0.183	0.127	0
	γ_{32}	0.000	0.480	0.239	0.240	0.122	0
	γ_{50}	0.090	0.614	0.353	0.352	0.134	0

High frequency models and estimation results of our example data are summarized in Table 7 and Table 8. The result of BayesVarSel is the same results; therefore, β_1 and β_2 are included in the final model, and observation 18 is determined as an outlier.

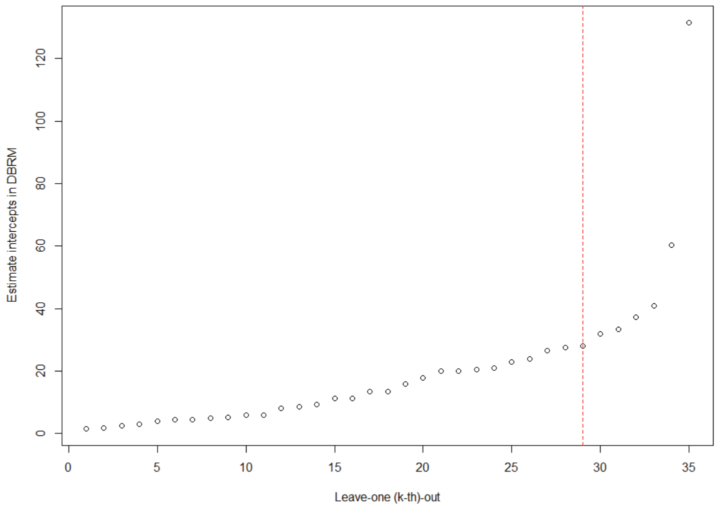

Figure 2: *Intercept estimates in DBRM in our example data.*

Table 7: High frequency models in our example data using SSVS

Model	Index set		Probability
	Model selection	Outlier detection	
1	$\{\beta_1,\beta_2\}$	$\{\gamma_{18}\}$	0.9996
2	$\{\beta_1,\beta_2\}$	$\{\ \}$	0.0003
3	$\{\beta_2\}$	$\{\gamma_{18}\}$	0.0001

Table 8: Estimation Results in our example data using SSVS

Parameter		Quantile (95%)		Median	Mean	Sd	Inclusion probability
		2.5%	97.5%				
Variable selection	β_1	21.045	35.307	28.155	28.177	3.621	0.9999
	β_2	18.938	29.414	24.186	24.178	2.643	1
Outlier detection	γ_5	−5.043	0.273	−2.429	−2.414	1.354	0
	γ_{17}	−2.277	3.543	0.620	0.609	1.474	
	γ_{33}	−1.299	4.622	1.697	1.684	1.502	0
	γ_{26}	−5.038	0.324	−2.361	−2.362	1.365	0
	γ_{31}	−6.267	−0.638	−3.527	−3.512	1.419	0
	γ_{18}	8.725	14.084	11.377	11.386	1.358	0.9997
	γ_{11}	−0.476	9.577	4.590	4.574	2.558	0

6. Discussion

In this paper, we have adopted the mean-shift outlier model in order to include information on outliers. This approach to modeling outliers is used by Kim *et al.* (2008) and Menjoge and Welsch (2010). The first step in these methods determines outlier candidates and the second step classifies outliers among them.

Accordingly, we suggest an alternative approach for simultaneous outlier detection and variable selection. First, by using properties of an intercept estimator in the DBRM (Park and Kim, 2018b), outlier candidates are determined and the information on outliers is reflected in the multiple regression model. Second, we select the best model from the model containing all variables including outlier candidates by using SSVS.

As shown in the simulation results and real data analysis, we have found that the performance of the proposed method is good under proper conditions. Furthermore, there is an advantage that the relative sizes of outliers can be confirmed from statistics of results. However, our method is affected by the constant value such as c^2 and prior variance τ^2. Therefore, we need to develop our method to make robust estimates. We will also to the nonparametric regression model for simultaneous outlier detection and variable selection.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2018R1D1A1B070).

References

Atkinson AC (1986). [Influential observations, high leverage points, and outliers in linear regression]: comment: aspects of diagnostic regression analysis, *Statistical Science*, **1**, 397–402.

Barbieri MM and Berger JO (2004). Optimal predictive model selection, *The Annals of Statistics*, **32**, 870–897.

Bayarri MJ, Berger JO, Forte A, and Donato GG (2012). Criteria for Bayesian model choice with application to variable selection, *The Annals of Statistics*, **40**, 1550–1577.

Belsley DA, Kuh E, and Welsch RE (1980). *Regression Diagnostics*, Wiley, New York.

Choi IH, Park CG, and Lee KE (2018). Outlier detection and variable selection via difference based regression model and penalized regression, *Journal of the Korean Data & Information Science Society*, **29**, 815–825.

Donato GG and Forte A (2017). BayesVarSel : Bayes factors, model choice and variable selection in linear models, R package version 1.8.0 Available on line access from https://cran.r-project.org/web/packages/BayesVarSel/BayesVarSel.pdf

George EI and McCulloch RE (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881–889.

George EI and McCulloch RE (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, **7**, 339–373.

Gelman A and Rubin DB (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–511.

Hoeting J, Raftery AE, and Madigan D (1996). A method for simultaneous variable selection and outlier identification in linear regression, *Computational Statistics and Data Analysis*, **22**, 251–270.

Kahng MW, Kim YI, Ahn CH, and Lee YG (2016). *Regression Analysis* (2nd ed), Yulgok, Seoul.

Kim S, Park SH, and Krzanowski WJ (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model, *Journal of Applied Statistics*, **35**, 283–291.

Menjoge RS and Welsch RE (2010). A diagnostic method for simultaneous feature selection and outlier identification in linear regression, *Computational Statistics and Data Analysis*, **54**, 3181–3193.

Park CG (2018). A study on robust regression estimators in heteroscedastic error models, *Journal of the Korean Data & Information Science Society*, **29**, 339–350.

Park CG and Kim I (2018a). Outlier detection using difference-based variance estimators in multiple regression, *Communications in Statistics - Theory and Methods*, **47**, 5986–6001.

Park CG and Kim I (2018b). Outlier detection using difference based regression Model, *Communications in Statistics - Theory Methods*, under review.

Park CG, Kim I, and Lee Y (2012). Error variance estimation in nonparametric regression under Lipschitz condition and small sample size, *Journal of Statistical Planning and Inference*, **142**, 2369–2385.

Rousseeuw PJ (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 871–888.

Weisberg S (2004). *Applied Linear Regression* (3rd ed.), Wiley, New York.

Received August 14, 2018; Revised November 19, 2018; Accepted February 8, 2019