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

Jong Suk Park\textsuperscript{a}, Chun Gun Park\textsuperscript{b}, Kyeong Eun Lee\textsuperscript{1,a}

\textsuperscript{a}Department of Statistics, Kyungpook National University, Korea; \textsuperscript{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 \textit{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 \textit{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 \textit{et al.}, 2008).

After that Kim \textit{et al.} (2008) proposed two similar steps, but the details are different from the approach proposed by Hoeting \textit{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
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 $X = [1_n : X_1]$ be an $n \times (p + 1)$ matrix, with $\text{rank}(X) = p + 1$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ be a $(p \times 1)$ coefficient vector where $\beta_0$ is an intercept. Assume that the response vector is $Y = (y_1, y_2, \ldots, y_n)'$.

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

$$Y = 1_n \beta_0 + X_1 \beta_1 + \gamma + \epsilon = X \beta + \gamma + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n),$$

(2.1)

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)'$ is an $n \times 1$ vector and $\gamma_j$ 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

$$Z = \{X : \sqrt{n} L_{\text{sub}}\},$$

where $L_{\text{sub}}$ is the submatrix with the first $q$ columns of $L_n$.

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{equation}
Y = \sum_{j=1}^{p+q+1} z_j \beta_j + \epsilon \tag{2.2}
\end{equation}

\begin{equation}
= z_{n \times (p+q+1)} \theta (p+q+1) + \epsilon.
\end{equation}

where \( \epsilon \sim N(0, \sigma^2 I_n) \) and \( \theta = (\beta', y')' \). Here, the new model includes \( p + q + 1 \) parameters, and \( \text{rank}(Z) = p + q + 1 \). The first \( (p + 1) \) of \( \theta_j \)’s come from the \( \beta_j \)’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 \( Y_{i(0)} \) and \( X_{i(0)} \) are \( Y \) and \( X \) without the \( i^{th} \) row for \( i = 1, 2, \ldots, n \). And \( D_{i(0)}Y \) is the difference between \( Y_{i(0)} \) and \( y_{i(n-1)} \). Then the difference-based regression model can be written as follows:

\begin{equation}
D_{i(0)}Y = I_{n-1}(-\gamma_i) + D_{i(0)}X_{i(0)} \beta_{i(0)} + A_{i(0)}\gamma + D_{i(0)}\epsilon
\end{equation}

\begin{equation}
= [I_{n-1}: D_{i(0)}X_1]'\gamma_i + A_{i(0)}\gamma + D_{i(0)}\epsilon, \quad i = 1, \ldots, n, \tag{2.3}
\end{equation}

where \( -\gamma_i \) is the intercept and \( D_{i(0)} \) and \( A_{i(0)} \) are the \( (n-1) \times n \) matrices as follows:

\begin{equation}
D_{i(0)} = \begin{pmatrix}
I_{n-1} & -I_{n-1} & 0_{(n-1),(n-i)} & \vdots & 0_{(n-1),(n-o)} \\
0_{(n-i),(i-1)} & -I_{n-i} & I_{n-i} \\
\end{pmatrix}
\quad \text{and} \quad A_{i(0)} = \begin{pmatrix}
I_{n-1} & 0_{n-1} & 0_{(n-1),(n-o)} \\
0_{(n-i),(i-1)} & I_{n-i} & I_{n-i} \\
\end{pmatrix}.
\end{equation}

where \( I_a \) is the \( a \times a \) identity matrix and \( 0_{a,b} \) 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:

\begin{equation}
-\hat{\gamma}_i = \begin{cases}
\Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})D_{i(0)}\epsilon, & \text{no outlier}, \\
\Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})A_{i(0)}\epsilon, & \text{several outliers}, \\
\end{cases}
\end{equation}

\begin{equation}
= \begin{cases}
\Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})D_{i(0)}\epsilon, & \text{one outlier}, \\
\Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})A_{i(0)}\epsilon, & \text{several outliers}, \\
\end{cases}
\end{equation}

where the \( i^{th} \) observation is not an outlier and \( H_{i(0)} = D_{i(0)}X_{i(0)}'D_{i(0)}X_{i(0)}^{-1}X_{i(0)}'D_{i(0)} \).

\begin{equation}
-\hat{\gamma}_k = \begin{cases}
-\gamma_k + \Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})D_{i(0)}\epsilon, & \text{one outlier}, \\
-\gamma_k + \Gamma_{n-1}^{-1}(I_{n-1} - H_{i(0)})A_{i(0)}\epsilon, & \text{several outliers}, \\
\end{cases}
\end{equation}

where the \( k^{th} \) observation is an outlier and \( H_{i(k)} = D_{i(k)}X_{i(k)}'D_{i(k)}X_{i(k)}^{-1}X_{i(k)}'D_{i(k)} \).
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. $D_i(Y) = Y_i - y_i(l_n - 1) + A_iX_1\beta_1 + D_i\epsilon$, for $i = 1, \ldots, n$;
2. Estimate intercepts, $-\hat{\gamma}_i$, $i = 1, \ldots, n$, and rewrite $\gamma'_i = \text{abs}(-\hat{\gamma}_i)$;
3. Ascend them, $\gamma'_1, \gamma'_2, \ldots, \gamma'_n$.

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 $\theta_j$ given the indicator $\delta_j$ is

$$\theta_j | \delta_j \sim (1 - \delta_j)N(0, \tau^2_j) + \delta_j N(0, c_j^2 \tau^2_j), \quad j = 1, 2, \ldots, (p + q + 1).$$

The value of $\tau^2_j$ is set to be small, and $N(0, \tau^2_j)$ is the prior distribution of the coefficient $\theta_j$, if the variable $z_j$ 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 $\theta_j$ should be included in the final model. The prior distribution of $\delta_j$ is $P(\delta_j = 1) = 1 - P(\delta_j = 0) = p_j$. The prior distributions of $(\delta_j, \theta_j)$ is independent from the prior distribution of $\sigma^2$, which is an inverse gamma (IG) distribution, $\text{IG}(\nu_0/2, \nu_0 \lambda_0/2)$.

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 DBRMs 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 $\theta_j$ conditional on $\delta_j$ and $\tau_j^2$ use mixture of two densities, which can be written as

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

(3.2)

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

$$Y|\theta, \sigma^2 \sim N\left(Z\theta, \sigma^2 I\right),$$  

(3.3)

$$\theta|\delta, \tau^2 \sim \text{N}_{p+q+1}(0, D_{\delta\tau}RD_{\delta\tau}),$$  

(3.4)

$$\delta_j \sim \text{Bernoulli}(p_j),$$  

(3.5)

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

(3.6)

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

(3.7)

where $R$ is the prior correlation matrix, and we assume $R = I$. Also, $D_{\delta\tau} = \text{diag}[d_1, \ldots, d_{p+q+1}]$ 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:

$$\theta|Y, \sigma^2, \tau^2, \delta \sim N\left(1/\sigma^2 A_{\delta\tau}Z'Y, A_{\delta\tau}\right),$$

$$\sigma^2|Y, \tau^2, \delta, \theta \sim \text{IG}\left(\frac{n + a_1}{2}, \frac{1}{2}(Y - Z\theta)'(Y - Z\theta) + b_1\right),$$

$$\tau_j^2|Y, \sigma^2, \delta, \theta \sim \text{IG}\left(\frac{a_2 + 1}{2}, \frac{\delta_j^2d_j^2 + b_2}{2}\right),$$

$$\delta_j|\delta, \theta, \tau^2 \sim \text{Bernoulli}(\hat{\delta}_j)$$

where $A_{\delta\tau} = (\sigma^{-2}Z'Z + (D_{\delta\tau}RD_{\delta\tau})^{-1})^{-1}$, $\delta = (\delta_1, \delta_2, \ldots, \delta_{p+q+1})'$, $d_j^2 = 1$ if $\delta_j = 0$ and $d_j^2 = c^2$ if $\delta_j = 1$, and

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

Therefore the best subset of variables is selected according to the information contained in the $\hat{\delta}$. The posterior probability $p(\delta_j = 1|Y)$ for the regressor $z_j$ 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 $p(\delta_j^{(m)} = 1|Y)$.

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.
Donato and Forte (2017) introduce the R package, \texttt{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 \texttt{Bvs}, \texttt{PBvs}, and \texttt{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 set to different $\beta$s in (2.1), $\beta_1 = (1, 1, 0, 0)'$ and $\beta_2 = (1, 1, 0, 0, 0, 0)'$. Also, we consider three different sample 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 $X = \{x_{ij}\} \sim \mathcal{N}(0,3^2C)$, where $C = (|i-j|)$, $i = 1; \ldots; n$, $j = 1; \ldots; p$. In addition, we generate errors using $\mathcal{N}(0,1)$, and 100 data sets for each case.

We apply SSVS using the hyperparameter settings: prior correlation matrix $R \equiv I$, and $\delta$ prior (3.5) with $p_j = 0.5$ which yields $\pi(\delta) \equiv 1/2^{p+j+1}$. We use three different values of $c^2_p = c^2$, $c^2 = 10, 25, 100$. For $c^2_j$, we consider two cases: $c^2_p = c^2_j = c^2$ or $c^2_p \neq c^2_j$ where $c^2_p$ denotes common prior variance for regression coefficients and $c^2_j$ 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 \texttt{Bvs} ($p + q + 1 < 20$) and \texttt{GibbsBvs} ($p + q + 1 \geq 20$) function in \texttt{BayesVarSel}. For choosing prior probabilities of models and $\theta$, we use two prior probabilities of models (ScottBerger, Constant) and five prior probabilities of $\theta$ 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 1: Priors used in \texttt{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                       |
Simultaneous outlier detection and variable selection via DBRM and SSVS

Table 2: Criteria of outlier detection

| Detection      | True        | Non-outlier | Sum       |
|----------------|-------------|-------------|-----------|
|                | n_{CD}      | n_{ID}      | n_{D}     |
|                | n_{IU}      | n_{CU}      | n_{U}     |
| 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.

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

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

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

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.

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

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

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 $\pi(\delta)$ of BayesVarSel, is “Constant”, which stands for $\pi(\delta) = 1/2^{p+q}$ and intercept ($\beta_0$) 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 $\tau^2 = \tau^2$. In contrast, the method using BayesVarSel shows different results depending on the prior probability of $\theta$. 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 $\theta$.

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 | CS | AN | Max | Mode |
| 30  | SSVS   | $c^2 = 100$      | $\tau_0^2 = \tau_1^2$ | 0.90 | 2.13 | 4 | 2 | 0.88 | 0.99 | 3.12 | 5 | 3 |  
|     |        | $\tau_0^2 < \tau_1^2$ | 0.88 | 2.15 | 4 | 2 | 0.76 | 0.98 | 3.26 | 5 | 3 |  
|     |        | $c^2 = 25$      | $\tau_0^2 = \tau_1^2$ | 0.89 | 2.07 | 4 | 2 | 0.93 | 0.99 | 3.08 | 6 | 3 |  
|     |        | $\tau_0^2 < \tau_1^2$ | 0.88 | 2.20 | 4 | 2 | 0.88 | 0.99 | 3.16 | 6 | 3 |  
| 50  | SSVS   | $c^2 = 100$      | $\tau_0^2 = \tau_1^2$ | 0.93 | 2.07 | 4 | 2 | 0.98 | 0.99 | 3.00 | 4 | 3 |  
|     |        | $\tau_0^2 < \tau_1^2$ | 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 |  
| 100 | SSVS   | $c^2 = 100$      | $\tau_0^2 = \tau_1^2$ | 0.97 | 2.02 | 4 | 2 | 0.95 | 0.95 | 4.95 | 6 | 5 |  
|     |        | $\tau_0^2 < \tau_1^2$ | 0.95 | 2.05 | 3 | 2 | 0.92 | 0.95 | 4.98 | 6 | 5 |  
|     | BayesVarSel | FLS | 0.95 | 2.03 | 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 |  
|     |        | $c^2 = 25$      | $\tau_0^2 = \tau_1^2$ | 0.98 | 2.03 | 3 | 2 | 0.99 | 0.99 | 10.31 | 12 | 10 |  
|     |        | $\tau_0^2 < \tau_1^2$ | 0.96 | 2.04 | 3 | 2 | 0.65 | 0.99 | 10.41 | 12 | 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 $\tau_0^2 = \tau_1^2$, our method performs best (PD is over 0.95 and PD + OS ≈ 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 $\beta_1 = (1, 1, 0, 0)'$. To identify outlier candidates, we calculate the absolute values of the intercept estimators ($\hat{\gamma}_i^*$) 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 ($Y$ and $Z$) is sorted by the order of $\hat{\gamma}_i^*$ 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 $\tau_0^2 = \tau_1^2$. 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 $\beta_1$, $\beta_2$ 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 | CS | AN | Max | Mode |
| 30  | SSVS         |    |    |     |      |    |        |    |    |     |      |
|     | $c^2 = 100$ | 0.73 | 3.44 | 7 | 3 | 0.73 | 0.91 | 3.14 | 6 | 3 |
|     | $c^2 = 25$  | 0.71 | 3.47 | 6 | 3 | 0.58 | 0.91 | 3.40 | 6 | 3 |
|     | $c^2 = 10$  | 0.78 | 3.20 | 6 | 3 | 0.89 | 0.91 | 3.14 | 4 | 3 |
| 50  | BayesVarSel  |    |    |     |      |    |        |    |    |     |      |
|     | $c^2 = 100$ | 0.82 | 3.12 | 7 | 3 | 0.66 | 0.91 | 3.20 | 6 | 3 |
|     | $c^2 = 25$  | 0.88 | 3.01 | 4 | 3 | 0.82 | 0.91 | 3.02 | 5 | 3 |
|     | $c^2 = 10$  | 0.84 | 3.06 | 5 | 3 | 0.68 | 0.91 | 3.18 | 6 | 3 |
| 100 | BayesVarSel  |    |    |     |      |    |        |    |    |     |      |
|     | $c^2 = 100$ | 0.82 | 3.12 | 7 | 3 | 0.66 | 0.91 | 3.20 | 6 | 3 |
|     | $c^2 = 25$  | 0.88 | 3.01 | 4 | 3 | 0.82 | 0.91 | 3.02 | 5 | 3 |
|     | $c^2 = 10$  | 0.84 | 3.06 | 5 | 3 | 0.68 | 0.91 | 3.18 | 6 | 3 |

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 ($\gamma_i^*$) 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 $\gamma_i^*$ and 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 $\tau_\beta^2 = \tau_\gamma^2$. 
Intercept estimates in DBRM on simulated data with $n = 50$, $p = 4$, $\beta_1 = (1, 1, 0, 0)'$, true outlier location = (2, 13, 15, 25, 33), and size of true outlier = (10, 7, 7, 7).

### Table 5: High frequency models simulated data using SSVS

| Model | Model selection | Outlier detection | Probability |
|-------|-----------------|-------------------|-------------|
| 1     | {\beta_1, \beta_2} | \{y_{13}, y_{25}, y_{33}, y_{25}, y_{15}\} | 0.99965     |
| 2     | {\beta_1} | \{y_{13}, y_{25}, y_{33}, y_{25}, y_{15}\} | 0.00005     |
| 3     | {\beta_1, \beta_2} | \{y_{13}, y_{25}, y_{33}, y_{25}, y_{15}\} | 0.00010     |
| 4     | {\beta_2} | \{y_{13}, y_{25}, y_{33}, y_{25}, y_{15}\} | 0.00010     |
| 5     | {\beta_1, \beta_2} | \{y_{25}, y_{33}, y_{25}, y_{15}\} | 0.00005     |
| 6     | {\beta_1, \beta_2} | \{y_{13}, y_{25}, y_{33}, y_{25}\} | 0.00005     |

### Table 6: Estimation results simulated data using SSVS

| Parameter | Quantile (95%) | Median | Mean | sd | Inclusion probability |
|-----------|----------------|--------|------|----|-----------------------|
| $\beta_1$ | 0.816 | 1.589 | 1.201 | 1.202 | 0.196 | 0.9999 |
| $\beta_2$ | 1.175 | 2.067 | 1.623 | 1.623 | 0.227 | 0.99995 |
| $\beta_3$ | -0.071 | 0.820 | 0.354 | 0.359 | 0.227 | 0 |
| $\beta_4$ | -0.487 | 0.227 | -0.121 | -0.122 | 0.182 | 0 |
| $\gamma_{13}$ | -1.149 | -0.639 | -0.896 | -0.895 | 0.129 | 0.99995 |
| $\gamma_{25}$ | -1.180 | -0.681 | -0.932 | -0.932 | 0.125 | 1 |
| $\gamma_{33}$ | -1.173 | -0.682 | -0.927 | -0.928 | 0.125 | 0.999 |
| $\gamma_2$ | -1.612 | -1.123 | -1.368 | -1.368 | 0.124 | 0 |
| $\gamma_{15}$ | 1.140 | 1.691 | 1.418 | 1.418 | 0.140 | 0.99995 |
| $\gamma_{24}$ | -0.151 | 0.316 | 0.085 | 0.084 | 0.119 | 0 |
| $\gamma_{44}$ | -0.097 | 0.397 | 0.154 | 0.153 | 0.126 | 0 |
| $\gamma_{39}$ | -0.433 | 0.068 | -0.185 | -0.183 | 0.127 | 0 |
| $\gamma_{32}$ | 0.000 | 0.480 | 0.239 | 0.240 | 0.122 | 0 |
| $\gamma_{50}$ | 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, $\beta_1$ and $\beta_2$ are included in the final model, and observation 18 is determined as an outlier.
Simultaneous outlier detection and variable selection via DBRM and SSVS

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

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

| Model | Index set | Model selection | Outlier detection | Probability |
|-------|-----------|-----------------|-------------------|-------------|
| 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 |
|-----------|----------------|--------|------|----|-----------------------|
| Variable selection | $\beta_1$ | 21.045 | 35.307 | 28.155 | 28.177 | 3.621 | 0.9999 |
| | $\beta_2$ | 18.938 | 29.414 | 24.186 | 24.178 | 2.643 |
| Outlier detection | $\gamma_5$ | -5.043 | 0.273 | -2.429 | -2.414 | 1.354 | 0 |
| | $\gamma_{17}$ | -2.277 | 3.543 | 0.620 | 0.609 | 1.474 |
| | $\gamma_{33}$ | -1.299 | 4.622 | 1.697 | 1.684 | 1.502 |
| | $\gamma_{26}$ | -5.038 | 0.324 | -2.361 | -2.362 | 1.365 |
| | $\gamma_{31}$ | -6.267 | -0.638 | -3.527 | -3.512 | 1.419 |
| | $\gamma_{18}$ | 8.725 | 14.084 | 11.377 | 11.386 | 1.358 | 0.9997 |
| | $\gamma_{11}$ | -0.476 | 9.577 | 4.590 | 4.574 | 2.558 |

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 $\tau^2$. 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