The GSp criterion on subset selection method for regression in the presence of outliers and multicollinearity

A Nabila, Y Widyaningsih and S Nurrohmah
Department of Mathematics, Faculty of Mathematics and Natural Science (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Corresponding author’s email: snurrohmah@sci.ui.ac.id

Abstract. Linear regression models a relation between a response variable with regressor variables and produce regression model. Ordinary least squares (OLS) method is often utilized to estimate parameters of linear regression model. OLS method will produce the best linear unbiased estimates when all the assumptions are met. The “best” means the estimate is unbiased and has the minimum variance. However, in real data, those assumptions are often violated, such as the multicollinearity and outliers. Multicollinearity will produce a large variance of the regression parameters, while outliers will cause a biased estimate. Jackknife ridge M-estimator is recommended to be implemented in the model when these problems are present. Jackknife ridge M-estimator is a parameter estimation method of the regression with robust property, which makes robust to outliers, and ridge method is utilized to overcome multicollinearity. Whereas the jackknife method is utilized to reduce the bias from the result of the ridge method. In multiple linear regression, regression model could possibly contain many candidates for regressor variables. Subset selection method is utilized to select some of those regressors that best predicts the response variable. The selection criterion that can be used is the subset selection method. The selection criterion that can be used is Mallow’s Cp criterion. The reason for this is because the GSp criterion is based on jackknife ridge M-estimator, which is able to solve the problem of outlier and multicollinearity. Furthermore, GSp criterion is implemented to obtain the best model on predicting IQ based on several predictors. It was found that IQ could be predicted by only doing 3 personality test which is test 1, test 3 and test 5 instead of all 5 different personality test.

Keywords: Best model, jackknife ridge M-estimator, predicting IQ, reduce bias, utilized.

1. Introduction
In multiple linear regression, the regression model may consist of many candidates for regressor variables. However, few of the candidates of regressor variables have important information on response variables, so it is needed to select only a subset of the candidates best predicts the response variable. The method that can be used is the subset selection method. Subset selection methods often use least square estimators to estimate parameters, for example Mallow’s Cp criterion. However, least square estimates are very sensitive when there is an outlier. Therefore, a more appropriate criterion was used in subset selection method if there is outlier in the data, that is Sp criterion which is based on robust regression [1].

Moreover, a problem that often occurs in regression model is the presence of multicollinearity. Multicollinearity is a condition where the inter-variable regressors are almost linearly dependent.
Multicollinearity results in increase of variance of the least square estimates, giving the consequence of unstable estimates and thus imprecise estimates of beta. One method that can be used to diagnose the presence of multicollinearity in linear regression models is the variance inflation factor (VIF). Imprecise estimates of beta can be solved by substituting least square estimates with ordinary ridge regression (ORR) estimates. Thus, Draper et al. [2] introduced the Rp criterion on the subset selection method based on the ORR estimate of $\beta$.

However, if outliers and multicollinearity appear together, it will be difficult to estimate the parameters in the subset selection method due to the result parameters are not accurate, where multicollinearity causes variance of parameter estimation larger and outlier causes bias on parameter estimation. To solve this problem, Jadhav et al. [3] introduced the GSp criterion, where the selection criteria for the subset selection method is based on the jackknife ridge Mestimator. This paper aims to find the best subset model by using generalized Sp (GSp) criterion when outliers and multicollinearity appear in the data.

2. Statistical method

2.1. Multicollinearity

Let the data consists of $n$ observations with $p$ measurements or regressors, $X_1, X_2, \ldots, X_p$ to explain the target $Y$. Matrix $X$ is a combine of columns $X_1, X_2, \ldots, X_p$. In multiple linear regression model, multicollinearity is a condition in which the regressor variables in multiple linear regression models almost linearly dependent. Producing correlation matrix amongst the regressors, it can be seen which variables are strongly correlated with the others.

To eliminate the effect of the unit of measurement on the data, assumed that response variables and regressor variables in multiple linear regression models have been standardized using the unit length scaling method. Let $W$ is the standardized of matrix $X$. Thus, $WW'$ matrix will be given in the form of correlation matrix as follows [4]:

$$W'W = \begin{pmatrix}
1 & r_{12} & \cdots & r_{1k} \\
r_{12} & 1 & \cdots & r_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
r_{1k} & r_{2k} & \cdots & 1
\end{pmatrix},$$

where $r_{ij}$ is the correlation between regression variables $X_i$ and $X_j$. Therefore, the $WW'$ matrix is often called the correlations matrix between two the regressor variables. If the relation between two independent variables of the regressors is linear, in other words, if two regressors have high correlation, then the $WW'$ matrix will be singular. Let $C$ denote the matrix $(WW')^{-1}$.

The matrix $C$ can be written as follows:

$$C = \begin{pmatrix}
C_{11} & C_{12} & \cdots & C_{1k} \\
C_{21} & C_{22} & \cdots & C_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
C_{k1} & C_{k2} & \cdots & C_{kk}
\end{pmatrix}.$$  

The main diagonal entry of the matrix $C$ is denoted by $C_{jj}, j = 1, 2, \ldots, k$ and can be written as:

$$C_{jj} = (1 - R^2_j)^{-1},$$

where $R^2_j$ is the coefficient of determination for $x_j$ which states the variability of $x_j$ described by other regressor variables through the multiple regression model. $R^2_j$ close to one states that the variability of the regression variable $x_j$ is very well described by other regressor variables, which means that $x_j$ is...
linearly dependent on other regressor variables. VIF (Variance Inflation Factor) for \( j \)-th regression variable is defined as follows:

\[
VIF_j = c_{jj} = \frac{1}{1 - R_j^2}.
\]

According to Montgomery, Peck and Vining, 2012 [4], the criteria used by VIF to express the existence of multicollinearity is when VIF > 10.

2.2. Outlier

Outliers can be defined as an extreme observation that has values far from other observational values. According to Montgomery et al. [4], an outlier is an oddity that signifies a different data point than any other data. When data has a large dimension, it will be difficult to describe the data only by using the plot. Therefore, another way to detect outliers is needed, such as the standardized residual of an observation. Observations that have large values of standardized residual indicate that these observations have considerable distances to the regression model’s estimates.

**Definition 1:**

Statistically, an observation that has a standardized residual of more than 3 (in absolute terms) can already be categorized as an outlier [5].

Standardized residual can be calculated using the following equation:

\[
d_i = \frac{e_i}{\sqrt{MSE}}, \quad i = 1, 2, ..., n; \quad e_i \text{ is } ..., \text{MSE is } ...
\]

Outliers cannot always be discarded because they might have information that is not explained by other data. In general, if there is an error in the measurement of the observation or if the sample point is not a desirable population, then discarding the sample point is normal. However, if the analysis states that the point is a valid sample, then there is no justification for removing the sample point. Outliers have a negative impact on the regression model; such as biased parameter estimates and change the regression line.

2.3. Jackknife Ridge M-Estimator

If the outlier and multicollinearity appear together, one of the parameter estimation methods that can be used is the jackknife ridge M-estimator. Actually, the problem of outlier and multicollinearity can be solved by using M-estimator ridge method only. However, combining the M-estimator and jackknife method is expected to produce a better result, in terms of mean square error, since jackknife method can minimize the bias in ridge regression method. Consider the following linear regression model:

\[
y = \beta_0 + X\beta + \epsilon.
\]

where \( y \) is an observation vector of size \( n \times 1 \) which is dependent variable (response variable), \( \beta_0 \) is an regression parameter vector of size \( n \times 1 \), \( X \) is an observation matrix of size \( n \times k \) where \( k \) is the number of independent variables, \( \beta \) is an regression parameter vector of size \( k \times 1 \), \( \epsilon \) is a random error vector of size \( n \times 1 \).

Let \( \lambda_1, \lambda_2, ..., \lambda_k \), \( k \) is the number of independent variables, be the eigenvalues of the matrices \( X'X \) and \( q_1, q_2, ..., q_k \) are corresponding eigenvectors. Let \( \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k) \), a diagonal matrix with diagonal elements are eigenvalues and \( Q = (q_1, q_2, ..., q_k) \), where matrix \( Q \) represent a matrix with columns are eigenvectors such that \( X'X = Q\Lambda Q' \). To find \( \alpha \), the parameters estimation in \( Q \), thus the above linear regression model can be written in canonical form as follows:

\[
y = \beta_0 + Z\alpha + \epsilon.
\]

where \( Z = XQ \) and \( \alpha = Q' \beta \).
The least squares estimate in canonical form is written as,

\[ \hat{\alpha}_{LS} = \Lambda^{-1} Z' y. \]

Since \( \alpha = Q' \beta \) and \( Q' Q = I \), then the least squares estimates of \( \beta \), that is \( \hat{\beta}_{LS} \), are:

\[ \hat{\beta}_{LS} = Q \hat{\alpha}_{LS}. \]

Note that, since the relation \( \alpha = Q' \beta \), then for any \( \hat{\alpha} \) estimate of \( \alpha \) has the relation \( \hat{\beta} = Q \hat{\alpha} \) and \( MSE(\hat{\alpha}) = MSE(\hat{\beta}) \). Therefore, to find the parameters estimation in \( Q \), we will consider the canonical form only to find out that the least squares estimates of \( \beta \) is \( \hat{\beta}_{LS} = Q \hat{\alpha}_{LS} \).

Let \( (\hat{\beta}_{BM}, \hat{a}_{BM}) \) be an M-Estimator of \( (\beta_0, \alpha) \). Thus, jackknife ridge M-Estimator, \( \hat{a}_{jRM}(r) \), can be derived as follows:

\[ \hat{a}_{jRM}(r) = (1 - r^2 A^{-2}) \hat{a}_{BM}, \quad (3) \]

where \( A = (\Lambda + r I) \) and \( r \) is a ridge constant, its role is to minimize the residual sum of squares, with \( r > 0 \).

2.4. Subset selection method

The subset selection method is used to select a subset of the regressor variable candidates in linear regression model. The following are reasons why a subset selection is needed in regression model:

- Not all regressor variables present in the model provide important information about the response variable \( y \).
- More regressors will produce higher variance of \( y \).
- More regressor variables require greater cost for data gathering.

3. Research method

Note that multiple linear regression models can be written as follows:

\[ y = X_{A} \beta_{A} + X_{B} \beta_{B} + \varepsilon, \quad (4) \]

where \( y \) is an observation vector of size \( n \times 1 \) which is dependent variable (response variable). To create a subset model based on \((p - 1)\) regressor variables, matrix \( X \) and regression parameter vector \( \beta \) are partitioned into \( X = [X_{A} : X_{B}] \) and \( \beta' = [\beta'_{A} : \beta'_{B}] \), respectively. The matrix \( X_{A} \) is a matrix of \( n \times p \) and the matrix \( X_{B} \) is a \( n \times (k - p) \) matrix, whereas \( \beta_{A} \) and \( \beta_{B} \) are vector parameters of \( p \times 1 \) and \( (k - p) \times 1 \), respectively. So the subset model based on \((p - 1)\) regressor variables can be written as follows:

\[ y = X_{A} \beta_{A} + \varepsilon. \quad (5) \]

The main problem of the subset selection is to find the best \( p \)-sized subset of all possible subset models, so that the selected subset of models will provide accurate predictions for the full model.

3.1. The GS\(_p\) Criterion

GS\(_p\) criterion are criteria that can be used in the subset selection method if outliers and multicollinearity appear together in the data. This criterion is based on the jackknife ridge M-Estimator of \( \beta \). Note the multiple linear regression model described previously, the prediction vector of \( y \) based on the estimated jackknife ridge M-Estimator of \( \beta \) i.e:
\[ \hat{y}_k = X\hat{\beta}_{JRM} = H_{JRM}y, \]

where \( \hat{y}_k \) is a prediction vector of full model, \( \hat{\beta}_{JRM} = XR(X'WX)^{-1}X'Wy \) and \( H_{JRM} = XR(X'WX)^{-1}X'W \) is the prediction matrix for full model based on jackknife ridge M-Estimator. JRM stands for jackknife ridge M-Estimator based on full model.

Let \( \hat{\beta}_{AJRM} \) be the estimated jackknife ridge M-Estimator of \( \beta_A \), then the prediction vector of \( y \) based on jackknife ridge M-Estimator of \( \beta_A \) is:

\[ \hat{y}_p = X_A\hat{\beta}_{AJRM} = H_{AJRM}y, \]

where \( \hat{\beta}_{AJRM} = X_A R(X_A'WX_A)^{-1}X_A'Wy \) and \( H_{AJRM} = X_A R_A(X_A'W_AX_A)^{-1}X_A'W_A \) is the prediction matrix for the subset model based on the jackknife ridge M-Estimator. Therefore, based on the two predictive vectors \( \hat{y}_k \) and \( \hat{y}_p \), we have \( G_{S_p} \) criterion which will be explained in Definition 2. AJRM stands for jackknife ridge M-Estimator based on subset model.

**Definition 2**: The \( G_{S_p} \) criterion is defined as follows:

\[ G_{S_p} = \frac{\sum_{i=1}^{n}(\hat{y}_{ik} - \hat{y}_{ip})^2}{\sigma^2} - \text{tr}[(H_{JRM} - H_{AJRM})(H_{JRM} - H_{AJRM})] + p, \]

where:

- \( \hat{y}_{ik} \) is the predicted value of \( y_i \) based on full model,
- \( \hat{y}_{ip} \) is the predicted value of \( y_i \) based on the subset model,
- \( p \) is the number of parameters in the subset model,
- \( H_{JRM} = XR(X'WX)^{-1}X'W \) is a prediction matrix based on full model,
- \( H_{AJRM} = X_A R_A(X_A'W_AX_A)^{-1}X_A'W_A \) is the prediction matrix for the subset model based on the jackknife ridge M-Estimator,
- \( \sigma^2 \) is variance error based on full model.

In the above equation the value of \( \sigma^2 \) is unknown, therefore its value will be estimated using the appropriate estimates and is based on the jackknife ridge M-Estimator, that is:

\[ \hat{\sigma}^2 = \frac{(y - X\hat{\beta}_{JRM})(y - X\hat{\beta}_{JRM})}{n - k}. \]

Algorithm for subset selection based on \( G_{S_p} \) criterion is as the following

1. Algorithm for subset selection based on criteria \( G_{S_p} \)

   **Step 1**: Calculate the statistic \( G_{S_p} \) for \( 2^{k-1} - 1 \) possible subset model, where \( k - 1 \) is the number of regressor variables in the full model.
Step 2: Select the subset model that has the minimum regressor variable among other subset models and satisfies:
\[ G S_p \leq p. \]

Step 3: Equivalent to step 2, the subset selection can be done by looking at the plot between \( G S_p \) versus \( p \), then selecting the subset model closest to the line.

The following are theorems that can be used to support the use of the criteria \( G S_p \) described above in selecting a suitable subset model, if the outliers and multicollinearity appear together in the data.

**Theorem 1:** If the model subset with the number of parameters \( p \) is true, then:
\[
E \left( \sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2 \right) \equiv \sigma^2 \text{tr} \left[ (H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM}) \right].
\]

**Proof:**
Note that \( \hat{y}_{ik} \) and \( \hat{y}_{ip} \) are the \( i \)-th prediction value of \( y \) based on the estimated jackknife ridge M-Estimater for full model and subset model. Thus, the sum of the squares of the difference between \( \hat{y}_{ik} \) and \( \hat{y}_{ip} \) can be written as:
\[
\sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2 = (\hat{y}_k - \hat{y}_p)' (\hat{y}_k - \hat{y}_p) = (H_{JRM}y - H_{AJRM}y)'(H_{JRM}y - H_{AJRM}y) = y'(H_{JRM} - H_{AJRM})(H_{JRM} - H_{AJRM})y.
\]

Then we will look for the expected value of both sides of the equation above as follows:
\[
E \left[ \sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2 \right] = E[y'(H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM})y].
\]

Since \( [(H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM})] \) is a symmetric matrix, then by [6], the expected value of the above equation can be written as:
\[
E \left[ \sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2 \right] = \sigma^2 \text{tr} \left[ (H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM}) \right] + \beta'X'[(H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM})]X\beta.
\]

However, if the subset model with the number of parameters \( p \) is correct, then
\[
\beta'X'[(H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM})]X\beta \equiv 0 \text{ so,}
\]
\[
E \left( \sum_{i=1}^{n} (\hat{y}_{ik} - \hat{y}_{ip})^2 \right) \equiv \sigma^2 \text{tr} \left[ (H_{JRM} - H_{AJRM})'(H_{JRM} - H_{AJRM}) \right].
\]
Theorem 2: If the subset model with the number of parameters \( p \) is true, then:

\[
[G_S] \cong p.
\] (10)

Proof:

If the model subset matches, Theorem 1 can be written as follows:

\[
E \left[ \sum_{i=1}^{n} \left( \frac{\hat{y}_{ik} - \hat{y}_{ip}}{\sigma} \right)^2 \right] \cong \text{tr}\left[(H_{jrm} - H_{AJRM})(H_{jrm} - H_{AJRM})\right].
\]

so,

\[
E[G_S] \cong \text{tr}\left[(H_{jrm} - H_{AJRM})(H_{jrm} - H_{AJRM})\right] - \text{tr}\left[(H_{jrm} - H_{AJRM})(H_{jrm} - H_{jrm})\right] + p.
\]

Next look at figure 1, the subset model with the minimum regressor variable that lies close to the line \( GS = p \) is a subset model consisting of the regressor variables \( x_1 \), \( x_2 \), and \( x_3 \), where the result is equivalent to the subset of the model generated by just looking at the value \( GS \).

4. Data application

Data on IQ measurement based on personality test is used in this research. Data is downloaded from https://www.ncss.com/ [7]. Data consists of measurement of IQ, denoted as \( y_i \), on 15 people and scores on 5 different personality tests they underwent, namely test 1 \( (x_{i1}) \), test 2 \( (x_{i2}) \), test 3 \( (x_{i3}) \), test 4 \( (x_{i4}) \) and test 5 \( (x_{i5}) \). The scores obtained from these five personality tests are measured on a scale of 1–100.

Multiple linear regression model for IQ person’s data is given by:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i.
\]

with \( i = 1, 2, ..., 15 \), where \( y_i \) is the IQ score of the \( i \)-th person, \( x_{ij} \) is the value of \( j \)-th the personality test score of \( i \)-th person, \( j = 1, 2, ..., 5 \).

It can be shown that there is an outlier, the second observation with a standardized residual value of -3.331, and multicollinearity problem in the data. regressor variables \( x_1 \) (VIF = 40.89), \( x_2 \) (VIF = 35.62), \( x_4 \) (VIF = 82.86), are greater than 10, then the \( GS_p \) criterion in the subset selection method can be used to get the best model. Since there are 5 predictors, we calculate the value of \( GS_p \) for \( 2^5 - 1 \) possible subset models Using R. The following is given a table of \( GS_p \) values for \( 2^5 - 1 \) possible subset models (table 1) and plots between values \( GS_p \) versus \( p \) for IQ data (figure 2).

![Figure 1. Flowchart of steps to finding the best model.](image-url)
Figure 2. Plot $GS_p$ vs $p$

It can be seen in figure 2, a subset of models with minimal regressor variables that are close to the line $GS_p = p$ is a subset of models consisting of regressor variables $x_1$, $x_3$, dan $x_5$, where the results are equivalent to the resulting subset of models by only looking at value $GS_p$. So it can be concluded that, the best model for one's IQ data is a model that consists of regressor variables $x_1$, $x_3$, dan $x_5$.

Note table 1 shows $GS$ values for $2^5 - 1$ subset models, the results from algorithm for subset selection based on criteria $GS$. Five is the number of regressor variables in the full model. A subset model will be selected which has the minimum regressor variables among other subset models and satisfies $GS \leq p$. So according to table 1, the subset model to be chosen as the best model for one's IQ data is a subset model consisting of the regressor variables $x_1$, $x_3$, dan $x_5$ because it has a value of $GS = 3.67 \leq p = 4$. Actually, the subset model that has the value of $GS$ closest to $p$ is the subset model consisting of the regressor variables $x_1$, $x_3$, $x_4$ and $x_5$ with the value $GS = 4.96$, but because the subset model to be selected is a subset model that has the minimum regressor variable among other subset
models, then the subset model selected as the best model for one's IQ data is a subset model consisting of the regressor variables \( x_1, x_3, \) and \( x_5 \).

Next look at figure 2, the subset model with the minimum regressor variables that lies close to the line \( GS = p \) is a subset model consisting of the regressor variables \( x_1, x_3, \) and \( x_5 \), where the result is equivalent to the subset of the model generated by just looking at the value \( GS \) at table 1. So it can be concluded that, based on the result of Jackknifed ridge M-estimator method, the best model for one's IQ data is a model consisting of regressor variables \( x_1, x_3, \) and \( x_5 \). Based on the data, it was found that IQ could be predicted by only doing 3 personality test which is test 1, test 3 and test 5 instead of all 5 different personality test, which can reduce the time, cost and energy required.

5. Conclusion

Based on the result of this study, if the outlier and multicollinearity appear together in the data, a parameter estimation method that can be used is the Jackknifed ridge M-estimator. Jackknifed ridge M-estimator is a regression coefficient estimate that it is unaffected by the outlier and uses ridge regression to overcome multicollinearity problems and also uses the jackknife method to reduce the bias of the ridge method. Thus, multicollinearity and outlier problems can be solved together. After the estimated parameters that is not affected by the outlier and multicollinearity obtained, then the estimated parameters will be used in the subset selection method to get the best model. Selection criteria in the subset selection method used when outliers and multicollinearity appear together is \( GS_p \) criterion. This is because the \( GS_p \) criterion is based on jackknifed ridge M-estimator that has been proven to overcome the problem of outlier and multicollinearity together.

Acknowledgments

This work was financially supported by Universitas Indonesia under research grant PITTA 2018.

References

[1] Dorugade A V and Kashid D N 2010 J. Stat. Comput. Simul. 80 1211-24
[2] Draper N R and Smith H 1998 Applied Regression Analysis 3rd edition (New York: John Wiley & Sons, Inc)
[3] Jadhav N H, Kashid D N and Kulkarni S R 2014 Stat. Methodol. 19 44-59
[4] Montgomery D C, Peck E A and Vining G G 2012 Introduction to Linear Regression Regression Analysis 5th edition (New York: John Wiley & Sons, Inc)
[5] Jadhav N H and Kashid D N 2011 J. Stat. Theory Pract. 5 659-73
[6] Kashid D N and Kulkarni S R 2002 Commun. Stat. Theory Methods 31 795-811
[7] NCSS Statistical Software 2017 available at https://www.ncss.com/