Conflict Graphical Threshold of Correlation for Breaking Multicollinearity in Regression Analysis

Younghoon Kim [a], Sangho Shim [b] and Seoung Bum Kim [a,1]

[a] *Industrial Management Engineering, Korea University, Seoul, Korea*

[b] *Managerial Economics and Decision Sciences Kellogg School of Management, Northwestern University, Evanston, IL, USA*

Abstract

Multicollinearity is the most challenging problem caused by tendency that independent variables in regression analysis are highly correlated. The multicollinearity reduces the reliability of estimated regression coefficients. In this study, we introduce a way of deciding the threshold of correlation which indicates the severity of multicollinearity. The way is to draw a conflict graph, which is the minimum vertex cover of multicollinear variables. The simulation results demonstrate that our proposed algorithm can provide an appropriate threshold for reducing large amounts of uncertainty of estimated regression coefficients.

Keywords: Conflict Graph, Multicollinearity, Regression Analysis, Subset Selection, Mixed Integer Optimization

[1] Email: sbkim1@korea.ac.kr

1 Introduction

Multicollinearity describes the situation in which two or more predictor variables in a regression model are linearly related with a non-trivial degree of correlation [3]. In this situation, the coefficient estimates of the linear regression tend to be unstable from one sample to another [4]. That is, a linear regression model with correlated variables may not produce reliable results about any individual variables, or about which variables are redundant with respect to others [8]. In some sense, the multicollinear variables include the same information about the response variable. If the same information is quantified by different variables in terms of correlation then the variables are redundant. This data redundancy may cause overfitting in regression models.

Several linear methods such as the partial least squares (PLS) regression [5], least absolute shrinkage and selection operator (lasso) [9] and stepwise regression were proposed to address a multicollinearity problem [2]. However, aforementioned methods are not designed to directly resolve multicollinearity because they mainly focused on improving accuracy by selecting important predictor variables. To overcome this limitation, we propose a novel regression modeling that can appropriately accommodate the tradeoff between multicollinearity and the accuracy of a regression model.

We first propose a conflict graphical regression model to address multicollinearity problem. The proposed model uses a mixed-integer program (MIP) and conflict graphs [1]. The objective of the proposed model is to minimize least $L1$ norm error with the following two constraints: (1) the variable selection constraint and (2) conflict graph constraints. The first variable selection constraint limits the number of selected variables less than the number of variables specified by the user. The second constraint forces the model to select one of the variables among highly correlated predictor variables.

To optimize the effect of the conflict graph constraint, a threshold that controls a tradeoff between multocollinearity and regression accuracy should be determined. In this paper, we propose an algorithm, called conflict graphical correlation threshold (CGCT) algorithm to efficiently and appropriately determine this threshold. The CGCT algorithm determines an appropriate threshold that reduces relatively large amounts of uncertainty of estimated regression coefficients by sacrificing small amounts of regression error.

2 Formulation

Given a dataset with m examples and N features, the conflict graphical regression model produces the predictive result $\hat{Y} = A\hat{x}$ where $A \in R^{m\times N}$ is a data matrix, $\hat{Y} \in R^{m\times 1}$, are predicted results and $\hat{x} \in R^{N\times 1}$ is a vector of coefficients. The optimization for the conflict graphical regression model is:

$$\begin{aligned} \underset{x}{\text{Minimize}} \quad & \|Y - Ax\| \\ \text{subject to} \quad & \sum_{i=1}^{N} z_i = K \\ & z_p + z_q \leq 2 + \gamma - |\rho_{pq}|,\ (p,q) \in E \\ & -Mz_i \leq x_i \leq Mz_i,\ i = 1,\ldots,N \\ & Y \in R^m, A \in R^{m\times N}, x \in R^N, z \in B^N. \end{aligned} \tag{1}$$

γ is a parameter that controls the construction of conflict relation between two predictor variables p and q in E, a set of all pairs of variables. If the correlation value between two variables is larger than γ, then the right hand side of the second constraint should be less than 2 (i.e., $z_p + z_q < 2$). Consequently, only one of the multicollinear variables is selected.

Aggregated conflict relations of variables can be represented by a conflict graph $G = (V, E^{'})$, where V and $E^{'}$ are, respectively, vertex and edge sets. The vertex i in V represents regression variable i and the edge (p, q) in $E^{'}$ indicates significant conflict relation between variables p and q.

In the aforementioned conflict graphical regression model, selecting an appropriate parameter γ is crucial to construct the conflict graph G because γ can adjust a tradeoff between the degree of multicollinearity and regression error. In the next section, we present a way to determine the appropriate γ in detail.

3 Conflict Graphical Correlation Threshold Algorithm

The degree of multicollinearity can be calculated by the mean value of variance inflation factor (VIF) of variables [6]. The VIF measures how much the variance of an estimated regression coefficient is increased by multicollinearity and is defined as follows:

$$VIF_i = \frac{1}{1 - R_i^2},\ i = 1,\ldots,N, \tag{2}$$

where R_i^2 is the coefficient of determination obtained by regressing A_i on the other predictor variables. Generally, VIF values larger than 10 indicate serious multicollinearity problems [7]. Our proposed algorithm utilizes the VIF threshold logic to detect variables with large multicollinearity.

Having found multicollinear variables, we define conflict relations to generate a conflict graph constraint in formulation (1). The conflict relations can be formed if the correlation value between two variables is larger than the parameter γ.

Algorithm 1 Conflict Graphical Correlation Threshold (CGCT)

procedure CGCT(A)
  Initialize:
  $M \leftarrow \emptyset, G \leftarrow \emptyset$
  ▷ Identification of multicollinear variables ($VIF[i] > 10$)
  for $i = 1 \rightarrow N$ **do**
    $VIF[i] \leftarrow VIF$ value of i^{th} variable.
    if $VIF[i] > 10$ **then**
      $M \leftarrow M \cup \{i\}$
    end if
  end for
  ▷ Determination of γ
  Calculate the correlation matrix $C \leftarrow corr(A)$
  for $i \in M$ **do**
    $G \cup \max\{C_{ij} |\ j = 1, 2, \ldots N$ *except* $i\}$
  end for
  $\gamma \leftarrow \min G$
return γ
end procedure

Fig. 1. is a toy example that describes three different graph structures by changing the parameter γ. The black nodes denote multicollinear variables whose VIF values are larger than 10, and the white nodes denote the variables that do not have multicollinearity. The edges in a graph are created if the correlation between two connected variables are larger than γ. Fig. 1(a) displays the case where γ=1. In this case, because there should be no edges in a graph, the model fails to detect true multicollinear variables. On the other hand, Fig. 1(c) shows the case where all edges are connected (γ=0). In this case, the model mistakenly detects the variables that do not have

multicollinearity. Obviously, both cases generate a serious problem. In the proposed CGCT algorithm, we change γ values from 0 to 1 until we find the γ^* (appropriate γ) where all the multicollinear variables are covered by conflict graph edges. This is shown in Fig. 1(b).

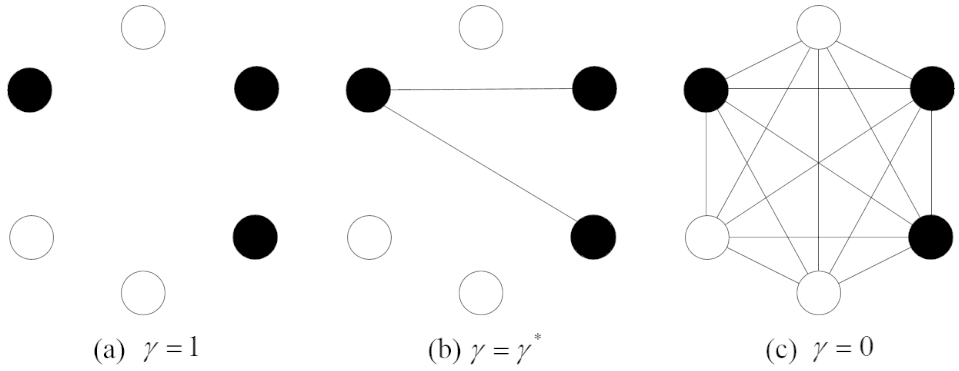

Fig. 1. Three conflict graph structures in terms of the different parameters γ.

4 Experimental Results

To examine the usefulness and applicability of the proposed CGCT algorithm, we generated the dataset containing 500 observations with 40 variables where each variable follows the normal distribution. Among 40 variables, 10 of them were highly correlated with each other. The regression errors and average VIFs were calculated by decreasing γ in formulation (1). Fig. 2. shows a clear tradeoff between the average VIF and regression errors. The appropriate value of the parameter γ obtained from the proposed CGCT algorithm was 0.83. To check how this parameter value affect VIF and regression accuracy, we plugged this value into formulation (1). We found that the degree of multicollinearity decreased 61.5%, while the regression error increased just 5.7%. We believe this simulation result demonstrated the usefulness of the proposed CGCT algorithm in that by using the γ value determined from the proposed CGCT algorithm, relatively large amounts of the multicollinearity decreased by sacrificing small amounts of regression errors.

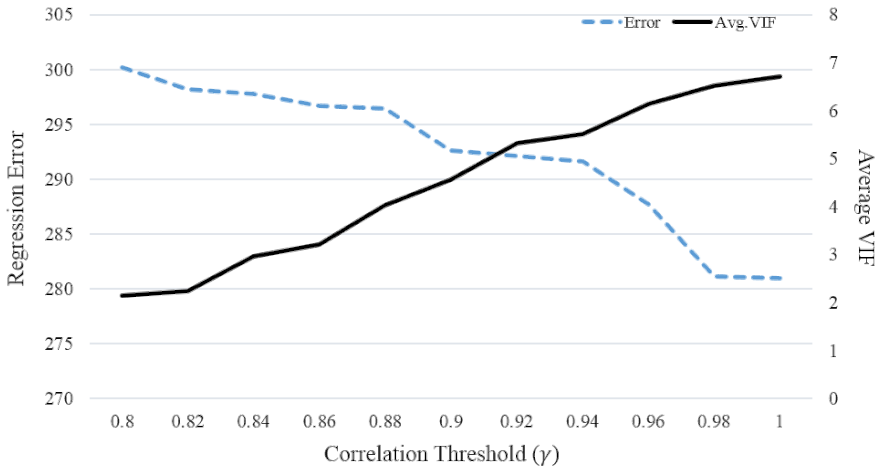

Fig. 2. Tradeoff relationship between average VIF and regression errors.

To visually confirm the above result, we generated a stem plot in which a stem indicates the VIF value of each predictor variable. Note that in Fig. 3(a) the VIFs of the nine variables are larger than 10 indicating their serious multicollinearity. Fig. 3(b) displays the result of the conflict graphical regression model based on the γ obtained from the proposed CGCT algorithm, showing that the multicollinear variables were successfully removed.

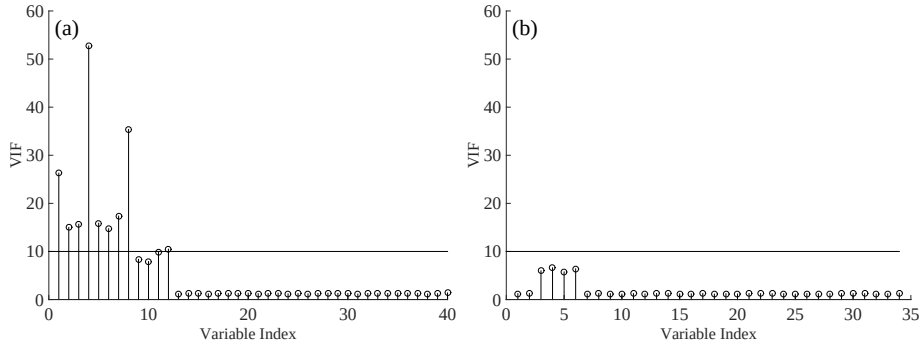

Fig. 3. VIF value for each variables (a) in the original dataset (b) after removing multicollinearity by the conflict graphical regression model.

5 Conclusion

In this study, we have proposed the CGCT algorithm to determine appropriate parameter that controls the tradeoff between multicollinearity and regression errors in a conflict graphical regression model. The experiments with simulated data demonstrated the usefulness of the proposed CGCT algorithm. In our further research, we plan to investigate the theoretical bound of regression errors in terms of changing the parameter γ.

References

[1] A. Atamturk, G.L. Nemhauser, M.W. Savelsbergh, Conflict graphs in solving integer programming problems, *European Journal of Operational Research* **121** (2000) 40-55.

[2] Chong, Il-Gyo, and Chi-Hyuck Jun, Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems* **78** (2005) 103-112.

[3] Dormann, Carsten F., et al., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography* **36** (2013) 027-046.

[4] Farrar, Donald E., and Robert R. Glauber, Multicollinearity in regression analysis: the problem revisited, *The Review of Economic and Statistics* (1967) 92-107.

[5] Geladi, Paul, and Bruce R. Kowalski, Partial least-squares regression: a tutorial, *Analytica chimica acta* **185** (1986) 1-17.

[6] Kutner, and M. H., *Applied linear statistical models*, (Chicago, Irwin, 1996).

[7] Obrien, Robert M., A caution regarding rules of thumb for variance inflation factors, *Quality & Quantity* **41** (2007) 673-690.

[8] Silvey, S. D., Multicollinearity and imprecise estimation, *Journal of the Royal Statistical Society. Series B (Methodological)* (1969) 539-552.

[9] Tibshirani, Robert, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* (1996) 267-288.