Assessing Multicollinearity via Identification of High Leverage Points in Financial Accounting Data

Norazan Mohamed Ramli[1], Zamalia Mahmud[2], Husein Zakaria[3], Mohammad Radzi Idris[3] and Alizan Abdul Aziz[3]

[1,2]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA
[3]Faculty of Accountancy, Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
[1]Email: norazan@tms.uitm.edu.my

ABSTRACT

Inaccurate and invalid statistical inferences in regression analysis may be caused by multicollinearity due to the presence of high leverage points (HLP) in a data set. Therefore, it is important that high leverage point which is a form of outlier be detected because its existence can lead to misfitting of a regression model, thus resulting in inaccuracy of regression results. In this paper, several methods have been proposed to identify HLP in a financial accounting data set prior to conducting further analysis of regression and other multivariate analysis. The Pearson's correlation coefficient and variance inflation factors (VIF) were used to measure the success of a detection method. Numerical analysis showed that common diagnostics like the twice-mean and thrice-mean rules failed to detect HLP in the given data set whilst robust approaches such as the potentials and diagnostic-robust generalized potentials (DRGP) methods were found to be successful in identifying high leverage point as indicated by lower values of the Pearson's correlation coefficient and variance inflation factors.

Keywords: *Multicollinearity, high leverage points, potentials, diagnostic-robust generalized potentials*

ISSN 1675-7017

Introduction

A key issue in interpreting the regression variate is the correlation among the independent variables. This problem is one of data, not of model specification. The ideal situation for a researcher would be to have a number of independent variables highly correlated with the dependent variable, but with little correlation among them. The assumption of normality for regression is no longer valid when the independent variables are highly correlated and this situation leads to multicollinearity. Multicollinearity may result in large standard errors for the regression coefficients and may result in the regression coefficients being insignificant. Khan (2003) proposed a solution to the problem of multicollinearity caused by the presence of multiple high leverage points.

Another cause of multicollinearity is the existence of HLP (Kamruzzaman and Imon, 2002). High leverage points are generally known to cause masking of outliers in linear regression. In this paper we shall focus on the identification of cases or points that can cause multicollinearity in financial data. In Financial Accounting Data section, a set of financial accounting data which had been identified as having multicollinearity problem will be presented. This is followed by discussions on several common methods for detecting HLP in Methods for Identification of High Leverage Points section. Identification of HLP is presented.in the Numerical Examples section and finally some remarks and conclusion are given in final section.

Financial Accounting Data

Financial accounting is aimed at providing information to parties outside the organization. Financial accounting data involves the recording and summarization of business transactions and events. It relates to the preparation of financial statements for external users such as creditors, investors, and suppliers. The main purpose of financial accounting is to prepare financial reports that provide information about a firm's performance to external parties such as investors, creditors, and tax authorities. The financial statements include the balance sheet, income statement, and statement of changes in financial position. These statements, including related footnotes, President's letter, management's discussion of operations, etc., appear in the annual report. Reporting of the financial position and performance of a firm through financial statements issued to external users on a periodic basis. Financial accounting should be performed according to Generally Accepted Accounting Principles (GAAP) guidelines (Britton and Waterston, 2008).

Since financial accounting data are used to provide information about a firm's performance to external parties, therefore it is important to ensure that the essential data are accurate, useful and functional so it can be used to gauge a firm's performance. External parties may also use the data to suggest improvement for the firm, if necessary.

However on several occasions it had been discovered that business and financial data can be messy and has the tendency to deviate from its randomness and normality assumption due to the nature of data. To some extent data can be a little extreme and considered to be of high leverage. Since the data's messiness can affect the accuracy of relationship measures, it is important to separate the contaminated from uncommon but valid observations in the data set. This issue of HLP should be taken into serious consideration as it relates to the masking of the true data. This would lead to invalid and inaccurate diagnosis of a performance measure (Seaver *et al.*, 1995).

Finance and business research analyst tend to use the financial accounting data to investigate the response or predictor variables that may influence the criterion variable. One issue of concern to the analyst is when these data were used without taking into consideration the effect caused by the presence of HLP. Kamruzzaman and Imon (2002) pointed out that HLP could cause multicollinearity in the predictor variables. HLP is an observation which lies near an extreme of the space of predictor variables while multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.

A Survey of Directors' Renumeration

Based on our diagnosis of identification of HLP in several data sets, we have discovered that the variables in the survey of Directors' Renumeration of MESDAQ companies data was suitable to highlight the problems regarding HLP. The data comprises 108 cases with five continuous predictors and eight categorical variables as shown in Table 1. However, only continuous variables are used as shown Table 1. The reason is that the computational method of HLP does not require the analysis of the categorical variables. Therefore, the search for the most appropriate method and correct identification of HLP is justified.

Table 1: Description of Variables in Director Renumeration Data

Name of Variables	Variable Description	Data Labels
Total Directors' Remuneration	Total payout based on what all executive and non-executive directors of the company received	Director Salary (Y)
Turnover	Based on the net sales or net income as reported in Annual Report.	turnover
Net Profit	The amount of income after deducting all expenses. Known as Profit-after-tax (PAT)	profit
Company Size	Measured by total assets as reported in Annual Report	company size
Equity Capital	Represent share capital of a company.	equity
Earnings Per Share	Profit in pence attributable to each ordinary share in a company	eps

Methods for Identification of High Leverage Points

Leverage Values

Leverage values are commonly used to measure influences in the X-space. Consider a k variable regression model of the form

$$Y = Xb + \in \qquad (1)$$

A vector $\hat{\in}$ denoting the OLS residual vector where $\hat{\in} = Y - \overline{Y} = (I - W)\in$. The weight matrix or leverage matrix is defined as

$$W = X\,(X^T X)^{-1}\,X^T \qquad (2)$$

The leverage values are the diagonal elements w_{ii} of W. On average, w_{ii} should be around $\frac{k}{n}$ (with k is the number of predictor variables that include the intercept term and n represents the number of observations). Any points are considered as high leverage cases if their w_{ii} exceeds a cut-off point of $\frac{2k}{n}$ (Hoglin and Welsch, 1978) and this rule is termed as twice-the-mean-rule. Meanwhile, Vellman and Welsch (1981) proposed the thrice-the-mean-rule and proposed that any points are considered as HLP whenever their w_{ii} exceeds a cut-off point of $\frac{3k}{n}$.

Potentials

A better measure for detecting HLP was proposed by Hadi (1992). According to Hadi (1992), the leverage of the i^{th} point should be based on a fit to the data with the i^{th} case deleted. Suppose the data matrix of k explanatory variables is written as $X = [x_1, x_2, \ldots, x_n]^T$. We define the i^{th} leverage value as

$$w_{ii} = x_i^T (X^T X)^{-1} x_i \tag{3}$$

and the i^{th} potential as

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \tag{4}$$

Note that the $X_{(i)}$ is the data matrix X with the i^{th} row deleted. Kamruzzaman and Imon (2002) gave simplified relationship between w_{ii} and p_{ii} with p_{ii} is defined as follows:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}} \tag{5}$$

Observations with p_i exceeding the cut-off point given in the equation (6) should be declared as HLP (Hadi, 1992).

$$\text{Median } (p_{ii}) + \text{c MAD } (p_{ii}) \tag{6}$$

In the above equation,

$$\text{MAD } (p_{ii}) = \text{Median } \{| p_{ii} - \text{Median } (p_{ii})|\} / 0.6745 \tag{7}$$

where c is a chosen constant such as 2 or 3.

Diagnostic-Robust Generalized Potentials (DRGP)

Diagnosing collinearity caused by outlying observations was investigated by Hadi (1988). However, Belsley *et al.* (1980) and Belsley (1991) used regression diagnostic to identify the sources of collinearity. On the other hand, a robust and unified approach for the detection of outlying observations which include high leverage points (HLP) was proposed by Habshah *et al.* (2009). In the DRGP method, an approach that based on Mahalanobis distance is initially used to detect the suspected HLP. Any suspected points are confirmed to be HLP if their generalized potential values go beyond a certain cut-off value. In the DRGP method, Robust Mahalanobis Distances (RMD) should be computed first to identify the suspected HLP. If RMD_i represents the robust mahalanobis distance for the ith point, then the RMD_i are defined as

$$RMD_i = \sqrt{[x_i - T(X)]^T [C(X)]^{-1} [x_i - T(X)]} \tag{8}$$

where $T(X)$ is the center of the minimal volume ellipsoid covering at least h points of X and $h = [n/2] + 1$ (see Rousseeuw (1984) for details). We suspect an i^{th} point as a high leverage point when its RMD_i exceeds the following cut off point:

$$\text{Median } (RMD_i) + 3 \text{ MAD } (RMD_i) \tag{9}$$

By checking its corresponding generalized potential value, it will confirm whether a suspected point is really a high leverage point. Following notations used by Kamruzzaman and Imon (2002), we define a deleted set D consists of observations only with its RMD_i exceeding the cutoff point given in equation (9) above and a set R contains $(n - d)$ cases after $d < (n - k)$ cases in D are deleted. Assuming these observations are arranged in the last of d rows of X and Y. We can partition $W = X(X^T X)^{-1} X^T$ as

$$W = \begin{bmatrix} U_R & V \\ & \\ V^T & U_D \end{bmatrix} \tag{10}$$

In this case

$$V = X_R(X^TX)^{-1}X_D^T \text{ is a } (n - d) \times d \text{ matrix,} \tag{11}$$

$$U_R = X_R(X^TX)^{-1}X_R^T \text{ and } U_D = X_D(X^TX)^{-1}X_D^T \tag{12}$$

are symmetric matrices of order (*n*-*d*) and *d* respectively.

We define the i^{th} leverage value that based on the group of deleted cases indexed by *D* as

$$w_{ii}^{(-D)} = x_i^T(X_R^TX_R)^{-1}x_i \tag{13}$$

The generalized potential for an i^{th} point is computed as

$$p_{ii}^* = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \tag{14}$$

for $i \in R$, $p_{ii}^* = w_{ii}^{(-D)}$ and where $i \in D$ with a cut-off point of

$$\text{Median } (p_{ii}^*) + c \text{ MAD } (p_{ii}^*) \tag{15}$$

Any suspected point in set *D* with p_{ii}^* exceeds above cut off point is declared as the true high-leverage point (Habshah *et al.*, 2009).

Numerical Examples

Based on the initial investigation of the Directors' Renumeration data, the correlation coefficient values and variance inflation factors as displayed in Table 2 shows that multicollinearity does exist in the given data set.

Table 2: Collinearity Diagnostics using Pearson's Correlation Coefficient for Directors Renumeration Data before Removing the HLP

Predictor Variables 1 = turnover 2 = profit 3 = company size 4 = equity 5 = eps				
Pearson's correlation coefficient values with the presence of High Leverage Points	$r_{12} = 0.2365$	$r_{13} = 0.6018$	$r_{14} = 0.6044$	$r_{15} = 0.0094$
		$r_{23} = 0.1676$	$r_{24} = 0.4624$	$r_{25} = 0.3507$
			$r_{34} = 0.9069$	$r_{35} = 0.0891$
				$r_{45} = 0.0243$

From the investigation, we believe that the presence of HLP may cause multicollinearity in the predictor variables. HLP were identified using all the methods discussed previously in Section 3. Figure 1 - Figure 4 display the index plots of leverages, potentials and Diagnostic-Robust Generalized Potentials for Directors' Renumeration data. HLP as detected by each of the method were identified and discussed in Section 3.

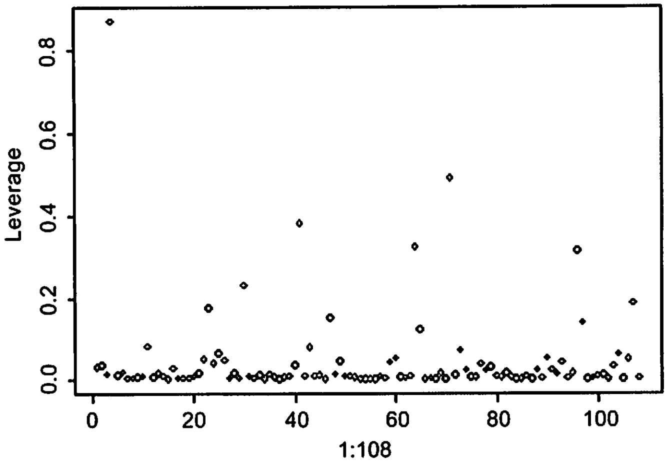

Figure 1: Index Plot of Leverages for Directors Renumeration Data using Twice-the Mean Rule with 10% Identified HLP

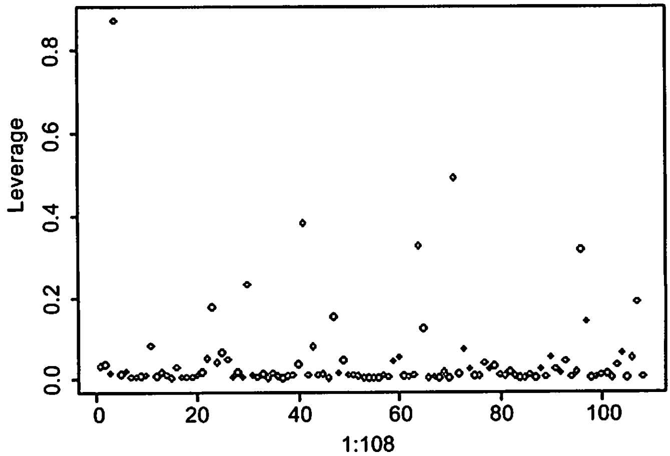

Figure 2: Index Plot of Leverages for Directors Renumeration Data using Thrice-the Mean Rule with only 7% Identified HLP

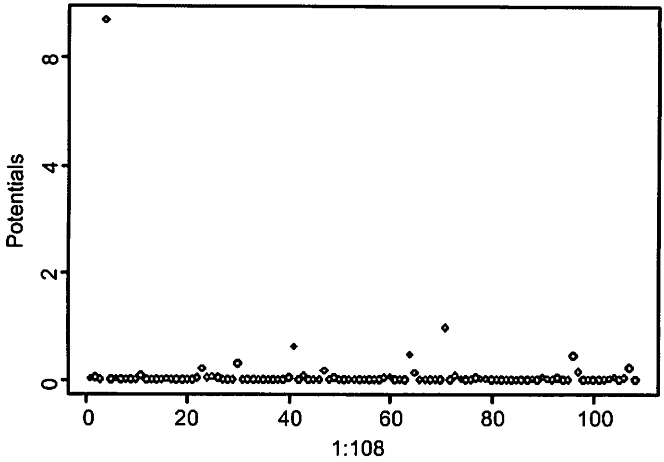

Figure 3: Index Plot of Leverages for Directors Renumeration Data using Hadi's Rule with 23% Identified HLP

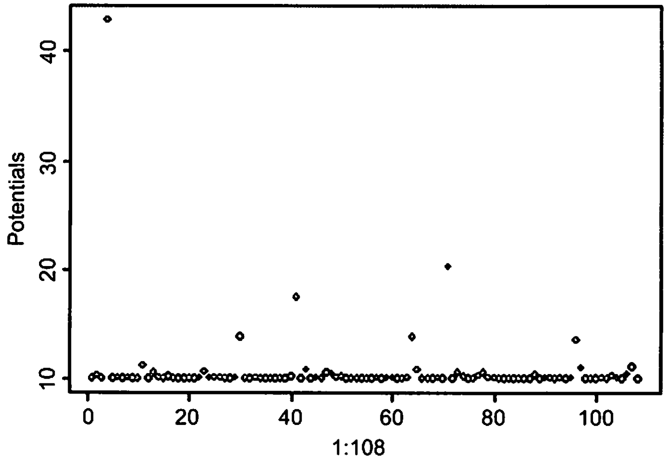

Figure 4: Index Plot of Diagnostic-Robust Generalized Potentials for Directors Renumeration Data with 20% Identified HLP

Table 3 provides percentages of identified HLP by each method. To further check whether these identified HLP actually caused the multicollinearity problem in the given data set, we recalculated the correlation coefficient values and variance inflation factors with and without the presence of these observations. Table 4 gives collinearity diagnostics using Pearson's correlation coefficient for Directors'

Renumeration data after removing the identified HLP. Meanwhile, Table 4 compares the collinearity diagnostics using variance inflation factors (VIF). VIF_j is calculated as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 of the artificial regression with the j^{th} independent variable as a "dependent" variable. If $\text{VIF}_j > 10$ is clear evidence that the estimation of the regression parameters are affected by multicollinearity. For details of the VIF computation, please refer to Fox and Monette (1992).

Table 3: Percentage of Identified HLP

Method	Percentage (%)
Twice mean rule	10
Thrice mean rule	7
Potentials	23
DRGP	20

Based on collinearity diagnostics using Pearson's correlation coefficient (Table 4), we notice that there exists a strong correlation between company size and equity. This is indicated by the Pearson's correlation coefficient value of 0.9069. Pearson's correlation coefficients were recalculated after removing the identified HLP. We observe that each method had detected different number of HLP. The Twice-Mean Rule had identified that the data consist 10% high leverage whilst the Thrice-Mean Rule had identified only 7. Meanwhile, higher percentages of HLP were detected using Hadi's (23%) and the Diagnostic-Robust Generalized Potentials (20%).

In Table 5, the Variance inflation factors (VIF) with and without the presence of identified HLP were used to check for the success of each method in detecting the exact HLP. Note that the variables are suspicious of significant multicollinearity problem if their VIF values exceeding a cutoff value of 10. In this case, we observe that prior to removal of the identified HLP, the VIF values for the company size and equity were very high, *i.e.,* 10.6117 and 12.8656, respectively. The VIF

Table 4: Collinearity Diagnostics using Pearson's Correlation Coefficient for Directors Renumeration Data after Removing the Identified HLP

Method used to identify HLP	Pearson's correlation coefficient values after removing the HLP (Predictor Variables: 1 = turnover, 2 = profit, 3 = company size, 4 = equity, 5 = eps)			
Twice mean rule	r_{12}=0.5231	r_{13}=0.4088	r_{14}=0.5925	r_{15}=0.2324
		r_{23}=0.3381	r_{24}=0.6496	r_{25}=0.6552
			$\boldsymbol{r_{34}}$**= 0.8256**	r_{35}=-0.0037
				r_{45}=0.2350
Thrice mean rule	r_{12}=0.5810	r_{13}=0.4807	r_{14}=0.6588	r_{15}=0.3115
		r_{23}=0.3976	r_{24}=0.6969	r_{25}=0.6644
			$\boldsymbol{r_{34}}$**= 0.8380**	r_{35}=0.0550
				r_{45}=0.3455
Potentials	r_{12}=0.5099	r_{13}=0.3918	r_{14}=0.6273	r_{15}=0.2365
		r_{23}=0.0814	r_{24}=0.6145	r_{25}=0.6460
			r_{34}=0.6811	r_{35}=-0.0233
				r_{45}=0.3150
DRGP	r_{12}=0.5536	r_{13}=0.3904	r_{14}=0.6078	r_{15}=0.3546
		r_{23}=0.1742	r_{24}=0.5616	$\boldsymbol{r_{25}}$**= 0.8402**
			r_{34}=0.7393	r_{35}=0.0230
				r_{45}=0.3420

Table 5: Collinearity Diagnostics using Variance Inflation Factors (VIF) for Directors Renumeration Data

Method	Status	1 = turnover	2 = profit	3 = company size	4 = equity	5 = eps
	With HLP	1.6242	2.5510	**10.6117**	**12.8656**	1.1732
Twice mean rule	Without HLP	1.6400	3.6772	4.0443	6.7346	1.9854
Thrice mean rule	Without HLP	1.8778	3.5761	4.6234	7.9020	1.9919
Potentials	Without HLP	1.7423	3.5632	2.8443	4.8336	1.7634
DRGP	Without HLP	1.8077	5.5929	2.7159	4.1996	3.7849

Significant if VIF > 10

were found to have decreased drastically after the identified HLP were isolated in the VIF calculation. It was discovered that the robust methods perform better than the other two methods in identifying the correct HLP. This was evident from the VIFs (without the identified HLP) which had shown improvement in terms of tremendous value decrease. The VIFs based on robust methods was also found to be the lowest among the other methods.

Conclusion and Recommendation

In this study, we have demonstrated four methods namely, Twice-Mean Rule, Thrice-Mean Rule and Potentials, Hadi's Rule and Diagnostic-Robust Generalized Potentials that were used to detect and identify HLP. Based on the analysis of the methods using Director's Renumeration data, it had shown that HLP had caused multicollinearity among the predictor variables. While each method successfully demonstrates the ability to detect the points, the main concern is to ensure that each method is able to identify a large number of correct HLP.

From the analysis, we observed that methods such as Twice-Mean Rule, Thrice-Mean Rule, Hadi's Rule and Potentials and Diagnostic-Robust Generalized Potentials each manage to detect HLP in the data but what differentiates between them is in the number of HLP being detected. Based on the comparison of all methods, the analysis had shown that robust methods, namely the Potentials and Diagnostic-Robust Generalized Potentials outperform the Twice-Mean Rule and the Thrice-Mean Rule methods in the detection of HLP in the Directors' Remuneration data. The outcome was indicated by the large number of HLP detected as well as a tremendous decrease in the VIFs values of the predictor variables when Diagnostic-Robust Generalized Potentials was applied on the data. These outcomes are important to research analysts when they need to investigate the relationship between the criterion and predictor variables in multiple linear regression. It is quite often that researchers tend to overlook the test for multicollinearity assumption once they were satisfied with the assumption testing, namely the normality and residual analysis. It is when the individual variable shows some insignificant results, only then it is realized that it was due to the existence of HLP which had caused multicollinearity.

Based on the analysis, it is recommended that research analyst gives great attention to the existence of HLP as it has shown to cause multicollinearity among the predictor variables. When data set contain outliers, it is also suggested that a more robust method be used to obtain the regression equation for the purpose of predicting and investigating the relationship between the variables. Future work shall include comparisons made on several regression methods that are suitable for fitting the data that contains HLP.

References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York, Wiley.

Belsley, D.A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, New York, Wiley.

Britton, A. and Waterston, C. (2008). *Financial Accounting*, Pearson.

Fox, J. and Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistician Association*, 87, 178-183.

Habshah, M., Norazan, M.R. and Imon, A.H.M.R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics,* 36, 507-520.

Hadi, A.S. (1988). Diagnosing collinearity influential observations. *Computational Statistics and Data Analysis*, 7, 143-159.

Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*, 14, 1-27.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. *The American Statistician Association,* 32, 17-22.

Kamruzzaman, M. and Imon, A.H.M.R. (2002). High leverage point: Another source of multicollinearity. *Pakistan Journal of Statistics*, 435-448.

Khan, M.A.I. (2003). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. *International Journal of Statistical Science*, 2, 37-50.

Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871-880.

Seaver, B.L. and Konstantinos P. (1995). The impact of outliers and leverage points for technical efficiency measurement using high breakdown procedures. *Journal of Management Science*, 41, 6, 937-956.

Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics. *Am. Statist.*, 27, 234-242.