AMAZING JOURNEY TO ROBUST STATISTICS, DISCOVERING OUTLIERS FOR EFFICIENT PREDICTION

THE **INAUGURAL LECTURES are given by honored faculty members within the University who have obtained the rank of full professor. This event gives the honoree the opportunity to deliver a lecture to fellow faculty and other university guests concerning their work and research interests.**

The context of the lecture itself typically includes a summary of the evolution and nature of the honoree's specialized field, highlights of some of the general issues of that particular field, and a description of how the honoree situates his/her work within their field.

UPM conducts this event to highlight and bring attention to the scholarly work that is being done by its distinguished faculty and to illustrate how the work contributes to mankind as a whole.

INAUGURAL LECTURE series

Professor Dr. Habshah Midi

PENERBIT UPM PRESS
Universiti Putra Malaysia Press
43400 UPM Serdang
Selangor Darul Ehsan

Tel: 03-89468851/89468854
Fax: 03-89416172
Email: penerbit@putra.upm.edu.my
Website: www.penerbit.upm.edu.my

ISBN: 9789673445424

INAUGURAL LECTURE series

Professor Dr. Habshah Midi

UPM UNIVERSITI PUTRA MALAYSIA BERILMU BERBAKTI

AMAZING JOURNEY TO ROBUST STATISTICS, DISCOVERING OUTLIERS FOR EFFICIENT PREDICTION

AMAZING JOURNEY TO ROBUST STATISTICS, DISCOVERING OUTLIERS FOR EFFICIENT PREDICTION

PROFESSOR DR HABSHAH MIDI

AMAZING JOURNEY TO ROBUST STATISTICS, DISCOVERING OUTLIERS FOR EFFICIENT PREDICTION

Professor Dr. Habshah Midi
B.A.Math (Drew), M.App. Stat. (Ohio), PhD Stat. (UKM)

6 MAY 2016

Dewan Taklimat
Universiti Putra Malaysia

Universiti Putra Malaysia Press
Serdang • 2016
http://www.penerbit.upm.edu.my

First Print 2016

UPM Press is a member of the Malaysian Book Publishers Association (MABOPA)
Membership No.: 9802

ISBN 978-967-344-542-4

Typesetting : Sahariah Abdol Rahim @ Ibrahim
Cover Design : Md Fairus Ahmad

Design, layout and printed by
Penerbit Universiti Putra Malaysia
43400 UPM Serdang
Selangor Darul Ehsan
Tel: 03-8946 8855 / 8854
Fax: 03-8941 6172
http://www.penerbit.upm.edu.my

Contents

ABSTRACT

In today's society, statistical techniques are being used widely in education, medicine, social sciences, and applied sciences. They are crucial in interpreting data and making decisions. When one makes a statistical inference, it is very crucial to be aware of the assumptions under which the statistical testing procedures can be applied. The assumptions that are common to almost all statistical tests are that the observations are random, independent and identically distributed, come from a normal distribution and they are equally reliable and should have equal role in determining the results. The last assumption implicitly states that there is no outlier in a data set. Outliers are observations which are markedly different or far from the majority of observations.

In most statistical models, the assumptions of normality of errors, no multicollinearity, homoscedasticity and non-autocorrelated errors are often violated. Another assumption that has received much attention from statisticians in recent years is that the regression analysis must be free from the effect of outliers. Even though the Ordinary Least Squares (OLS) estimates retain unbiasedness in the presence of heteroscedasticity, multicollinearity and autocorrelation, their estimates become inefficient. As such, proper diagnostic checking should first be considered before further data analysis is carried out. The problem gets more complicated, when the violation of homoscedasticity, no multicollinearity, and no autocorrelation, each comes together with the existence of outliers. Methods that are designed to rectify these problems, cannot handle both problems simultaneously. In this regard, proper remedial measures should be taken into consideration to remedy these problems. Hence, some robust methods which are developed to simultaneously remedy these two problems will be illustrated in this inaugural lecture. Robust method is a relatively new method

whereby it is not easily affected by outliers because their effects are reduced. This presentation also focuses on our research, in developing robust diagnostic methods for detecting whether or not outliers, multicollinearity, heteroscedasticity and autocorrelated errors are present in a data.

This presentation also will illustrate some of our developed diagnostic methods to identify high leverage points and also to indicate whether multicollinearity is caused by correlated predictors or high leverage points. This presentation will also illustrates the effects of outliers and high leverage points on panel data model, response surface model and variable selection methods. Outliers are known to have an adverse effect on computed values of various estimates. The immediate consequences of outlier are that they may cause apparent non-normality and the entire classical methods breakdown. Classical methods heavily depend on assumptions. However, in practice those assumptions are difficult to be met. Violations of at least one of the assumptions may produce sub-optimal or even invalid inferential statements and inaccurate predictions.

Since outliers give bad consequences, the need for robust methods become essential to avoid misleading conclusion. Hence, we developed several robust methods pertaining to these issues. Due to space limitations, only some selected developed robust methods will be presented in this inaugural lecture and their mathematical derivations are not shown.

INTRODUCTION

In all aspects of our lives, an amazing diversity of data is available for inspection and analysis. Business managers, government officials, policy makers and professionals require solid facts based on data to justify a decision. They need statistical techniques to support their decisions since statistical analysis of data can provide investigators with powerful tools to interpret data relevant to their decision- making. However, the conclusion drawn from a study is to be trusted only when correct statistical techniques are employed. Furthermore, it is usually unwise to rely on the results of test procedures unless the validity of all underlying assumptions has been checked, and met for a valid inferential statement. We may use diagnostic checking to confirm the validity of these assumptions. When the basic assumptions are not satisfied, proper remedial measures should be taken into consideration.

In today's society, it is very unfortunate that with the easy availability of statistical packages such as SAS and SPSS has driven more statistics practitioners to use the packages blindly in analysing their data. Unfortunately, they often are not aware of the fact that statistical packages just follow the instruction given to them and produce results accordingly. They do not know whether researchers have chosen the correct statistical techniques for their studies. Box (1953) stated that *"now it's really too easy, you can go to the computer and with practically no knowledge of what you are doing, you can produce sense or nonsense at a truly astonishing rate"*. With little knowledge in statistics they rely too much on statistical packages to analyse their data. Unfortunately, they are also not aware of the effect of outliers on various estimates and not aware of the immediate consequences of the presence of outliers. Even one single outlier can have arbitrarily large effect on the estimates. In statistical data analysis, there is only one type

of outlier, but in a regression problem, extra care should be taken because in this situation, there are several versions of outliers exist such as residual outliers, vertical outliers and high leverage points. Any observation that has large residual is referred to as a residual outlier. Vertical outliers (VO) or y-outliers are those observations which are extreme or outlying in y-coordinate.

On the other hand, high leverage points (HLPs) are those observations which are extreme or outlying in *X*-coordinate.
The assumption of normality, that is, the data are a random sample from a normal distribution is the most important assumption for many statistical procedures. Non-normality may occur because of their inherent random structure or because of the presence of outliers. Most of the standard results of a study are based on normality and other assumptions and the whole inferential procedure may be subjected to error if there is a departure from these assumptions. The violation of these assumptions may lead to the use of suboptimal estimators, invalid inferential statements and inaccurate predictions and for these reasons we developed test for normality and heteroscedasticity.

Belsley *et al.* (1980) stated that influential observations were those observations either alone or together with several other observations have the largest impact on the computed values of various estimates. It is often very essential in regression analysis to find out whether HLPs have much impact on the fitting of a model. HLPs not only fall far from the majority of predictor variables, but also are deviated from a regression line (Hocking and Pendelton, 1983; Rousseeuw and Leroy, 1987). Habshah *et al.* (2015) pointed out that HLPs can cause multicollinearity. These leverage points may increase (enhancing observation) or decrease (reducing observation) multicollinearity problem (Bagheri *et al*., 2012b).

This inaugural lecture presents part of my research works being performed with my students and colleagues in the field of robust statistics. Robust statistics is a technique that is less affected by the presence of outliers because their effects have been reduced.

Our research was mainly focused on the robust diagnostic methods and robust parameter estimation in linear, nonlinear, logistic, generalised linear and response surface models. Research on robust variable selection procedure, robust statistical process control and robust methods on panel data has also being performed over a decade. This presentation will cover the concept of outlier and influential observations, diagnostic methods, robust graphical display, robust parameter estimations, robust response surface methodology and robust variable selection technique. Diagnostics are designed to find problems with the assumptions of any statistical procedures. Most of the classical statistical procedures heavily depend on normality assumption of observations. A robust rescaled moment (RRM) test which is fairly robust and possesses higher power of rejection of normality in the presence of outliers is developed in this regard. The Diagnostic Robust Generalised Potential (DRGP) has been developed for the identification of high leverage points because it is responsible for misleading conclusion about the fitting of linear regression, causing multicollinearity and swamping and masking outliers in linear regression. The Robust Modification of the Goldfeld Quant (MGQ) and Robust Modified Breusch Godfrey (MBG) tests are developed to detect heteroscedasticity and autocorrelation problems, respectively. Robust Variance Inflation factor (RVIF) and High leverage Collinearity Influential Observations Methods are established to indicate whether multicollinearity problems exist in a data. The new robust diagnostics methods need to be developed because the non-robust methods fail to detect the existence of those problems

in the presence of outliers. When these problems have been correctly identified, appropriate remedial measures are taken into consideration that provides efficient estimates.

THE NEED FOR ROBUST STATISTICS

The word "Robust" literally means something "very strong". Therefore robust statistics are those statistics which do not breakdown easily. The analogous term used in the literature is Resistant Statistics. It is less affected by outliers by keeping its effect small. In classical setup, the assumptions that are common to almost all statistical test are that the observations are random, independent and identically distributed, come from a normal distribution and equally reliable (there is no outlier in a data). The classical methods heavily depend on assumptions and the most important assumption is that data are normally distributed. Hampel *et al.* (1986) claimed that a routine data set typically contains about 1-10% outliers, and even the highest quality data set cannot be guaranteed free of outliers. The immediate consequence of outlier is that it may cause apparent non-normality and the entire classical methods might breakdown. This is the reason why we need to turn to robust statistics where it does not rely heavily on the underlying assumptions. It is usually unwise to rely on the results of test procedures unless the validity of all underlying assumptions have been checked and met. Violations of these basic assumptions may produce sub-optimal or even invalid inferential statements and inaccurate predictions.

ROBUST DIAGNOSTIC METHODS

A Robust Rescaled Moment Test for Normality in Regression

Most of the statistical procedures heavily depend on the normality assumption of observations. But in practice we often deal with data sets which are not normal in nature. Nevertheless, evidence is available that such departure can have unfortunate effects in a variety of situations. When the errors are not normally distributed, the estimated regression coefficients and estimated error variance are no longer normal and consequently the t and F tests are generally not valid in finite samples. Most of the standard results of statistical tests are based on the normality assumption and the whole inferential procedures may be subjected to error if there is a departure from this and for this reason, test for normality has become an essential part of data analysis.

There are a considerable amount of written papers relating to the performance of various tests for normality in regression (Gel and Gastwirth, 2008). Among them, the Jarque–Bera (JB) test for normality (also known in statistics the Bowman–Shenton test) has become very popular with the statisticians. The JB test statistic is a sum of the sample coefficients of skewness and kurtosis and asymptotically follows a χ^2 distribution with two degrees of freedom. But the main shortcoming of the JB test is that it possesses very poor power when the sample size is small or moderate (Montgomery *et al*., 2001). To overcome this problem, rescaled moment (RM) and robust Jarque–Bera (RJB) tests are developed. Since the RJB is designed as a general statistical test for normality, we suspect that it may not perform well in regression analysis. Hence we propose a robust rescaled moment test (RRM) for normality designed for regression models extending the idea

of Imon (2003) and Gel and Gastwirth (2008). Rana *et al.* (2009) developed the robust rescaled moment (RRM) test statistic as:

$$\text{RRM} = \frac{nc^3}{B_1}\left(\frac{\hat{m}_3}{J_n^3}\right)^2 + \frac{nc^4}{B_2}\left(\frac{\hat{m}_4}{J_n^4} - 3\right)^2, \text{ where } J_n = \frac{A}{n}\sum_{i=1}^{n} |X_i - M|$$

is the average absolute deviation from the sample median, and $A = \sqrt{\pi/2}$. The robust sample estimates of skewness and kurtosis are $\hat{m}_3 / J_n^3$ and $\hat{m}_4 / J_n^4$, where $\hat{m}_3$ and $\hat{m}_4$ are the 3rd and 4th order of the estimated sample moments respectively. Under the null hypothesis of normality, the RRM test statistic asymptotically follows a chi-square distribution with 2 degrees of freedom. B_1 and B_2 are computed similar to the RJB test statistic as suggested by Gel and Gastwirth (2008).

To assess the performance of our proposed test, we consider the shelf-stocking data given by Montgomery *et al.* (2001). These data present the time required for a merchandiser to stock a grocery store shelf with a soft drink product as well as the number of cases of product stocked. We deliberately change one data point to create an outlier. For the original data, all the methods showed that the residuals for these data are normally distributed (Table 1). The standard theory tells us that the normality should break down in the presence of outliers. But it is interesting to observe that both the Jarque-Bera and the RM test fail to detect non-normality here. The RJB test also fails to detect non-normality at the 5% level of significance. But the performance of our RRM test is quite satisfactory in this occasion. It can detect the problem of non-normality even at 1.6% level of significance.

Table 1 Power of normality tests for original and modified shelf-stocking data

Tests	Original data		Modified data	
	Value of Statistic	p-value	Value of Statistic	p-value
JB	1.2643	0.5314	2.1820	0.3359
RM	1.9700	0.3735	3.4524	0.1779
RJB	1.4632	0.4811	5.0890	0.0785
RRM	2.2477	0.3250	8.2475	0.0161

A Robust Modification of the Goldfeld–Quandt Test for the Detection of Heteroscdasticity in the Presence of Outliers

It is a common practice over the years to use the ordinary least squares (OLS) as the inferential technique in regression. Under the usual assumptions, the OLS possesses some nice and attractive properties. Among them, homogeneity of error variances (homoscedasticity) is an important assumption for which the OLS estimators enjoy the minimum variance property. It is now evident that the heteroscedastic problems (when assumption of homoscedastic error variance is not met) affects both the estimation and test procedure of regression analysis, so it is really important to be able to detect this problem for possible remedy. If this problem is not eliminated, the OLS estimators will still be unbiased, but the parameter estimates will have larger standard errors than necessary. The Goldfeld –Quandt (GQ) and Breusch-Pagan test (Goldfeld and Quandt 1965) are quite popular and commonly used in econometrics. But, there is evidence that all these tests suffer huge set back when outliers are present in the data. We have modified the GQ test (Rana *et al.*, 2008) by

integrating robust methods in the formulation of the Modified Goldfeld-Quandt (MGQ) test and is summarised as follows.

$$\text{MGQ} = \frac{MSDR_2}{MSDR_1}$$

where $MSDR_1$ and $MSDR_2$ are the median of the squared deletion residuals for the smaller and the larger group variances, respectively. Under normality, the MGQ statistic follows an F distribution with numerator and denominator degrees of freedom each of $(n - c - 2k)/2$.

To show the merit of our developed test, we consider restaurant food sales data given by Montgomery *et al*.(2001). In this data set there is a relation of income with advertising expense. Again we deliberately put three outliers into the data set by replacing the income of the cases indexed by 1, 27 and 30 with large values. It is very obvious from the plot in Figure 1, that the original data has heterocedastic errors.

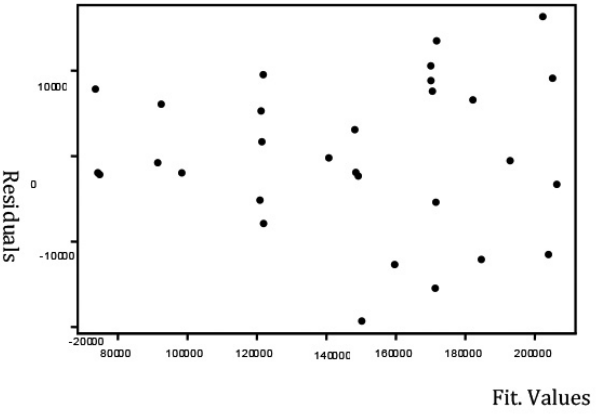

Figure1 Residuals vs. fitted plot for original restaurants food sales data

Table 2 Heteroscedasticity diagnostics for restaurants food sales data

Test	Without Outliers		With Outliers	
	Value of Statistic	***p*-value**	**Value of Statistic**	***p* -value**
Goldfeld-Quandt	4.03671	0.0190	1.074	0.4563
Breusch-Pagan	3.1787	0.0746	0.3799	0.5376
White	4.3575	0.0368	0.0963	0.7562
MGQ	4.9917	0.0090	10.4566	0.0005

Test results as presented in Table 2 show that the three conventional tests perform well in detection of heteroscedasticity but their performances are poor when outliers are present in the data. The MGQ test performs best. Irrespective of the presence of outliers it can successfully detect the heteroscedastic error variance in the data.

Diagnostic-Robust Generalised Potentials for the Identification of Multiple High Leverage Points in Linear Regression

Detection of high leverage values is crucial due to their responsibility for misleading conclusion about the fitting of a regression model, causing multicollinearity problems, masking and/or swamping of outliers etc. It is now evident that most of the commonly used variable selection techniques for model building are affected in the presence of high leverage points and often could produce very misleading conclusions. That is why the identification of high leverage points is essential in linear regression before making any kind of inference.

Much work has been done on the identification of high leverage points (Hoaglin and Welsch, 1978; Huber, 1981; Vellman and Welsch, 1981). However, most of the existing methods fail to identify them because they suffer from masking (false negative) and swamping (false positive) effects. As such, Habshah *et al.* (2009) has formulated a new measure for the identification of HLPs that are called DRGP where the suspected high leverage points are identified by Robust Mahalanobis Distance based on Minimum Volume Ellipsoid (MVE) and then the low leverage points (if any) are put back into the estimation data set after diagnostic checking using generalised potentials to confirm our suspicions.

The generalised potentials for all members in a data set are defined as

$$p_{ii}^{*} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \quad \text{for} \quad i \in R$$

$$= w_{ii}^{(-D)} \quad \text{for} \quad i \in D$$

where $w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \quad i = 1,2,\cdots\cdots,n$, D and R are any arbitrary deleted set and remaining sets of points, respectively. p_i^{*} is considered to be large if

$$p_i^{*} > \text{Median}\,(p_i^{*}) + c\,\text{MAD}\,(p_i^{*})$$

where c equals 2 or 3.

We report a Monte Carlo simulation experiment which is designed to investigate how our newly proposed diagnostic robust generalised potentials perform in the identification of multiple

high leverage points and to compare its performance with other commonly used methods.

Table 3 Identification of multiple high leverage cases (average) based on 10000 simulations

Per-centage	Sample	No. of	Identification Methods				
	Size	HLPs	Twice mean	Thrice mean	Huber	Potentials	DRGP
5%	n = 20	1	0	0	6	2	2
	n = 40	2	2	2	2	3	3
	n = 60	3	3	3	3	4	4
	n = 100	5	6	6	2	6	6
	n = 200	10	9	9	0	9	12
10%	n = 20	2	2	2	7	2	3
	n = 40	4	3	3	3	3	5
	n = 60	6	5	5	2	4	7
	n = 100	10	7	7	0	6	11
	n = 200	20	13	13	0	11	21
15%	n = 20	3	1	1	9	2	3
	n = 40	6	3	3	3	2	6
	n = 60	9	4	4	0	3	9
	n = 100	15	8	8	0	6	15
	n = 200	30	14	14	0	11	30
20%	n = 20	4	0	0	9	1	4
	n = 40	8	2	2	2	2	8
	n = 60	12	5	5	0	3	12
	n = 100	20	7	7	0	4	20
	n = 200	40	15	15	0	9	40
	n = 20	5	0	0	8	1	5
	n = 40	10	2	2	2	1	10

Table 3 clearly shows the merit of our proposed DRGP method. The number of HLPs detected by DRGP exactly or close to the number of HLPs generated in this experiment. All the commonly used methods failed to identify the high leverage points while the method based on DRGP was successful in identifying high leverage points.

Robust Modification of Breush-Godfrey Test in the Presence of High Leverage Points

The OLS estimates will have optimum properties when all the underlying model assumptions are met. However, practitioners will hardly check the fulfillment of the underlying model assumptions especially the assumptions of random and uncorrelated errors. Most of the time, the assumption of random and uncorrelated errors is taken for granted despite the errors may be correlated with the previous errors. When the error terms are correlated with the previous errors such that $E(\mu_i , \mu_j) \neq 0$, for $i \neq j$, the errors are said to be autocorrelated. This problem mostly happens in time series data.

Autocorrelated errors cause serious problems in linear model. It violates the important properties of the OLS (White and Brisbon, 1980). The parameters estimates obtained are no longer the Best Linear Unbiased Estimators (BLUE) in the sense that their standard errors, $\hat{\sigma}$, are most likely to be underestimated. As the results, the less efficient estimates are obtained because of ignoring the erroneous assumption. The usual t and F tests of significance are no longer convincing. These tests tend to be statistically significant when in fact it is not. The coefficient of determination, R^2 becomes inflated. As such, the estimators would look more accurate as compared to its actual value. All these problems contribute to the failure of the hypothesis testing. Hence,

the autocorrelated errors will most probably provide misleading conclusions about the statistical significance of the estimated regression coefficients (Gujarati and Porter,2009). Therefore, it is very important to detect the presence of autocorrelated errors.

There are quite a number of written articles related to autocorrelation testing procedures (Breusch, 1978; Godfrey, 1978; Durbin and Watson, 1951). Among them, the Breusch-Godfrey (BG) test is the most widely used test to detect the presence of autocorrelated errors. This test is suspected to be affected by high leverage point since it is based on the OLS which is known to be easily affected by outlying observations. Hence robust BG test which is not much affected by high leverage points is proposed for the detection of autocorrelated errors in multiple linear regression (Lim and Midi, 2012; Lim and Midi, 2014). The proposed test incorporates the bounded influence, high efficient and high breakdown MM-estimator (Yohai, 1987) in the Breusch-Godfrey procedure and is called Modified Breusch-Godfrey (MBG) test. Lim and Midi (2012) proposed MBG test as follows:

$$R_M^2 = \frac{\mathbf{SSR}}{\mathbf{SSE} + \mathbf{SSR}}$$

where SSR is the sum of squares regression and SSE is the sum of squares errors of the auxiliary regression using MM estimator. They showed that the distribution of the Lagrange Multiplier statistic of MBG, that is $(n-p)R_M^2$ is approximately Chi-Squares with p degrees of freedom.

The performance of our developed test is shown by real data and simulation study.

For each sample size n = 20, 40, 60, 80, 100 and 200, the n 'good' data are generated according to the following relation:

$$y = 1 + 2X_1 + 3X_2 + u$$

where all the values of X_1 and X_2 are generated from Uniform Distribution, U(0,10). The Uniform Distribution is chosen to ensure that the generated data are free from outliers. This will minimise the chance of generating the outlier observations in the simulation run. The error terms u_t are generated by the first-order autoregressive scheme as follows:

$$U_t = 0.9u_{t-1} + \varepsilon_i$$

with an initial value of u_1 generated from Normal Distribution, N(0,4) in order to ensure there is a strong autocorrelation problem in the dataset when the White noise, ε is generated from Normal Distribution, N(0,1). The performance of BG and MBG tests with 5% and 10% high leverage points in X_1, X_2 and both X_1 and X_2 directions are examined.

The average *p*-values of both BG and MBG tests based on 10,000 simulation runs are presented in Table 4. From the table it is clearly seen that, the BG test has more significance *p*-value than MBG test for detecting autocorrelated errors in the clean datasets. However, the BG test performs miserably in the presence of high leverage points. It is very disappointed to see that BG test can only detect autocorrelated errors in the clean data but fails to diagnose the autocorrelated errors in all levels and all kinds of contamination. Unlike BG test, the MBG test did a credible job. This finding has shown that the MBG test is a robust diagnostic method for autocorrelation. MBG test is not only working well in detecting autocorrelated errors in clean datasets, but it also performs superbly good in identifying autocorrelation in contaminated datasets as compared to classical BG test.

Table 4 The *p*-Values of BG and MBG Tests in the Simulation Study
(Both Positive Directions for β_1 and β_2)

Sample	Test	p-values	5% of HLPs			10% of HLPs		
			X1	X2	X1 and X2	X1	X2	X1 and X2
20	BG	6.72E-03	3.87E-01	4.10E-01	4.11E-01	3.82E-01	4.14E-01	4.10E-01
	MBG	6.90E-03	2.08E-02	2.18E-02	2.26E-02	1.47E-02	2.11E-02	2.14E-02
60	BG	5.03E-06	3.29E-01	4.21E-01	4.56E-01	3.84E-01	4.43E-01	4.62E-01
	MBG	6.27E-06	1.02E-04	1.29E-04	7.81E-04	1.10E-03	9.72E-04	7.23E-04
100	BG	2.08E-11	2.27E-01	3.92E-01	4.43E-01	2.95E-01	4.12E-01	4.55E-01
	MBG	2.20E-11	1.05E-07	1.39E-07	1.40E-07	9.95E-06	1.94E-05	8.78E-06
200	BG	<e-16	8.53E-02	2.83E-01	4.07E-01	1.58E-01	3.42E-01	4.24E-01
	MBG	<e-16	1.23E-14	2.11E-14	2.93E-14	1.41E-13	4.66E-14	4.70E-14

General Road Accident Data in Malaysia

These time series data considered the general road accident data in Malaysia. These data show how the number of road deaths in Malaysia (y) associates with the road length (X_1) and the number of road accidents cases in Malaysia (X_2) from year 1974 to 1999. The data can be obtained from the research paper by Mustafa (2005). Similar to the previous time series data, a normal observation in X_1, X_2 and both X_1 and X_2 directions is arbitrary replaced by a high leverage point in order to get a modified high leverage data in X_1, X_2 and both X_1 and X_2 directions. Figure 2 shows the index plot of residuals for the original data based on OLS estimation. It can be seen very clearly that the residuals are not randomly distributed but followed a cyclical pattern. This provides us a strong evidence to claim that the residuals are not randomly distributed but they are correlated with the previous errors.

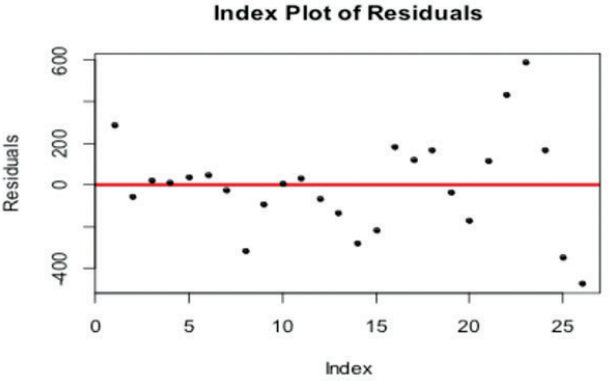

Figure 2 Index Plot of Residuals for General Road Accident Data in Malaysia

The performance of BG and MBG tests in identifying the autocorrelated errors in the original and modified general road accident data in Malaysia are exhibited in Table 5. The BG test is found to have slightly better autocorrelation detection power than MBG test in the original data. However, it is very disappointed to see that the BG test gives misleading findings of no autocorrelated errors in every respect of the contamination made in the original data. The autocorrelation detection power of BG test dropped drastically in the contaminated datasets. It is exciting to note that the MBG test never fails to diagnose the presence of autocorrelated errors in the original as well as in high leverage datasets.

Table 5 Autocorrelation Diagnostics for General Road Accident Data in Malaysia

Tests	BG (*p*-values)	MBG (*p*-values)
No High Leverage Point	6.420e-03	8.191e-03
One High Leverage Point in X_1	7.131e-02	2.968e-03
One High Leverage Point in X_2	6.841e-01	2.443e-02
One High Leverage Point in X_1 and X_2	7.177e-01	1.991e-02

Robust Variance Inflation Factor to Diagnose Multicollinearity

Multicollinearity occurs in a data set when explanatory variables are correlated to each other. Although the OLS estimates are still unbiased in the presence of multicollinearity, its estimates become inefficient (Montgomery *et al*., 2001; Kutner *et al*., 2005; Chatterjee and Hadi, 2006). One of the most important destructive

effects of multicollinearity on regression analysis is non-significant results of individual *t*-tests for some of the important regression coefficients when overall *F*-test confirms the existence of linear relationship between explanatory variables and response variable. Hence, it is imperative to diagnose whether multicollinearity exists in a data.

Variance Inflation Factor (VIF) is one of the most popular multicollinearity diagnostic tools which measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related (Marquardt, 1970). If R^2 is the coefficient determination of each of the explanatory variables when regressed on the other explanatory variable model by OLS method, VIF is given by

$$VIF_j = \frac{1}{1 - R_j^2} \qquad j=1,\ldots,k$$

Moderate or severe collinearity exists in the data set when VIF is between 5 and 10 or exceeds 10, respectively. To prevent misleading conclusions that may be obtained from the classical VIF in the presence of high leverage points, a robust multicollinearity diagnostic method based on robust coefficient determination,should be employed. In this regard, Bagheri and Habshah (2011) proposed two RVIFs, namely the RVIF(MM) and the RVIF(GM(DRGP)).

The proposed RVIF(GM(DRGP)) is defined as follows:

$$RVIF_i(GM(DRGP)) = \frac{1}{1 - RR_i^2(GM(DRGP))} \qquad i = 1,2,\ldots,k.$$

where robust coefficient determination is defined as follows

$$RR^2(GM(DRGP)) = 1 - \frac{\sum_{i=1}^{n} w_{i(GM(DRGP))} e^2_{i(GM(DRGP))}}{\sum_{i=1}^{n} w_{i(GM(DRGP))} (y_i - \bar{y})^2},$$

where

$$\bar{y} = \sum_{i=1}^{n} w_{i(GM(DRGP))} y_i \Big/ \sum_{i=1}^{n} w_{i(GM(DRGP))},$$

e_t and $w_{(GM(DRCP))}$ are the residual and weight, respectively after the algorithm converged.

In this section, the effect of high leverage points on a collinear data set which is taken from Kutner *et al.* (2005) is investigated. Body Fat data set contains 20 observations with three explanatory variables of triceps skinfold thickness (X_1), thigh circumference (X_2) and midarm circumference (X_3). This data set has multicolinearity problem (Kutner *et al*., 2005). In order to modify this data set to have high leverage collinearity- reducing observation, the first observation of the first explanatory variable is replaced with a large value of high leverage point (equal to 300). The results also indicate that only a large value of high leverage point in X_1 ruin the collinearity pattern of the data.

Table 6 exhibits the Classical and Robust VIFs for the original and modified Body Fat data set. It can be observed that for the original data set, the classical VIF and RVIF (GM (DRGP)) indicate the presence of severe multicollinearity in the data set while RVIF (MM) diagnose moderate collinearity in this data set. Thus, RVIF (MM) failed to detect the correct degree of collinearity. However, by modifying the data set through adding a high leverage point, the classical VIF failed to detect collinearity whereas RVIF (GM

(DRGP))) and RVIF (MM) can detect collinearity in this data set. It is interesting to note that the proposed RVIF (GM (DRGP)) can diagnose the degree of multicollinearity correctly (severe multicollinearity) while the RVIF (MM) can only identified moderate collinearity. Hence our new proposed RVIF (GM (DRGP)) is not affected by the added high leverage point and still show the existence of collinearity in this data set.

Table 6 Classical and robust VIF for original and modified Body Fat data set

Variables	Original data set			Modified data set		
	CVIF	RVIF (MM)	RVIF (GM(DRGP))	CVIF	RVIF (MM)	RVIF (GM(DRGP))
X_1	708.8429	5.2997	785.3549	1.1266	5.7297	7.8225
X_2	564.3434	5.4690	656.7576	1.1141	5.4722	628.7662
X_3	104.6060	5.0593	115.0129	1.0363	5.0560	123.6363

Collinearity Influential Observation Diagnostic Measure based on a Group Deletion Approach

High leverage points can induce or disrupt multicollinearity patterns in a data. Observations responsible for this problem are generally known as collinearity-influential observations. Development of collinearity-influential observation diagnostic measures has not been reported extensively in the literature (Hadi, 1988; Sengupta and Behimasankaram, 1997; Bagheri and Habshah, 2012a; Bagheri *et al*., 2012b). There is strong evidence that existing measures that are designed to detect a single observation as collinearity-influential may not be effective in the presence of multiple high leverage collinearity-influential

observations. In this presentation, a novel diagnostic measure for the identification of multiple high leverage collinearity-influential observations is shown (Bagheri *et al*., 2012b).

The proposed high leverage collinearity-influential measures based on DRGP (HLCIM (DRGP)), which is denoted as $\delta_i^{(D)}$ is defined and summarised as follows:

$$\delta_i^{(D)} = \begin{cases} \log \dfrac{k_{(D)}}{k_{(D-i)}} & \text{if } i \in D \text{ and } n(D) \neq 1 \\ \log \dfrac{k_{(D)}}{k} & \text{if } i \in D \text{ and } n(D) = 1 \\ \log \dfrac{k_{(D+i)}}{k_{(D)}} & \text{if } i \in R \end{cases}$$

where D is the group of multiple high leverage points diagnosed by DRGP(MVE) (p_{ii}^{*}) and n(D) is the size of the D group. $k_{(D)}$ and $k_{(D-i)}$ indicate the condition number of the X matrix without the entire group of D high leverage points and without the entire D group minus the i^{th} high leverage points where i belongs to the D group, respectively. Furthermore, $k_{(D+i)}$ refers to the condition number of the X matrix without the entire group of D high leverage points plus the i^{th} additional observation of the remaining group.

The well-known Hawkins, Bradu, and Kass (1984) data is used to show the merit of our proposed method. This artificial three-predictor data set contains 75 observations with 14 high leverage points (cases 1-14). The results in Table 7 show that the existing measures δ^i and l_i can identify the first 13 high leverage points as collinearity-enhancing observations while our proposed $\delta_i^{(D)}$ measure can successfully identify the first 14 observations as high leverage collinearity-enhancing observations.

Table 7 Collinearity-influential measures for Hawkins-Bradu-Kass data

Index	$k_{(i)}$	δ_i	l_i	$\delta_i^{(D)}$
		(-.008)	**(-.004)**	**(-.019)**
		(-0.048)	**(-0.021)**	**(-.022)**
1	13.221	**-0.027**	**-0.012**	**-0.228**
2	13.183	**-0.03**	**-0.013**	**-0.241**
3	13.289	**-0.022**	**-0.01**	**-0.234**
4	13.18	**-0.03**	**-0.013**	**-0.254**
5	13.188	**-0.029**	**-0.013**	**-0.248**
6	13.185	**-0.03**	**-0.013**	**-0.24**
7	13.166	**-0.031**	**-0.014**	**-0.248**
8	13.237	**-0.026**	**-0.011**	**-0.227**
9	13.235	**-0.026**	**-0.011**	**-0.242**
10	13.327	**-0.019**	**-0.008**	**-0.226**
11	13.06	**-0.039**	**-0.017**	**-0.29**
12	13.424	**-0.012**	**-0.005**	**-0.272**
13	13.035	**-0.041**	**-0.018**	**-0.319**
14	17.125	0.26	0.101	**-0.391**
15	13.67	0.006	0.003	-0.005
16	13.752	0.012	0.005	0.01
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
74	13.611	0.002	0.001	-0.002
75	13.651	0.005	0.002	0.009

To investigate the effect of collinearity-influential observations on the collinearity structure of the data, we computed collinearity diagnostics including pair-wise Pearson correlation coefficients, variance inflation factors, and condition indices. These results are presented in Table 8. The results in the table shows that the multicollinearity problem of these data is reflected in the VIF and Condition Number (CN) values. We can see from the table that in the presence of 14 HLPs (original data), the data have multicollinearity but in their absence, there is no multicollinearity. This is referred as High Leverage Collinearity Enhancing Observations.

Table 8 Collinearity diagnostics for Hawkins-Bradu-Kass data

Diagnostics	**Status**	**1**	**2**	**3**
Pearson correlation coefficient	Original data	r_{12}= 0.946	r_{13}= 0.962	r_{23}= 0.979
	Without observations 1 –14	r_{12}= 0.044	r_{13}= 0.107	r_{23}= 0.127
VIF > 5	Original data	13.432	23.853	33.432
	Without observations 1 – 14	1.012	1.017	1.027
Condition index of X matrix > 10	Original data	13.586	7.839	1.00
	Without observations 1 – 14	3.275	2.946	1.00

A New Robust Diagnostic Plot for Classifying Good and Bad High Leverage Points in a Multiple Linear Regression

It is not easy to capture the existence of several versions of outliers in multiple regression analysis by using a graphical method (Rousseeuw and Leroy, 1987). If only one independent variable is being considered, the four types of outliers can easily be observed from a scatter plot of y against the x variables. However, for more than one predictor variable, it is difficult to detect these outliers from a scatter plot. Not much work has been focused on classifying HLP's into good leverage point (GLP) and bad leverage point (BLP).

Rousseeuw and Zomeren (1990) proposed a robust diagnostic plot which is more effective than the non-robust plot for classifying observations into regular observations, vertical outliers, GLPs and BLPs. Rousseeuw and Zomeren plot draws the standardised least median of square residual (LMS) against the robust Mahalanobis distance (RMD) based on minimum volume ellipsoid (MVE), whereby this plot is denoted by LMS-RMD. The non-robust plot draws the Studentised OLS residuals (t_i) against the Mahalanobis distance (MD), we called this plot as OLS-MD plot. We suspect that the robust LMS-RMD diagnostic plot is not very effective in classifying the observations into respective categories since it is based on the robust Mahalanobis distance, which suffers from swamping effects (Bagheri and Habshah, 2015). Moreover, this plot uses Studentised residual which is not very successful in identifying multiple outliers. Habshah *et al.* (2009) showed that the DRGP was very successful in detecting multiple HLPs. In addition, we anticipate that the newly proposed Modified Generalised t (MGt_i) is able to detect multiple outliers. As such

we proposed to improve the classification method of Rousseeuw and Zomeren (1990) by plotting the MGt_i versus DRGP. Our proposed diagnostic plot is called MGt-DRGP plot. The basic rules for classification observation by using the new proposed method are as follows (Mohamed *et al*., 2015a).

1. Regular Observation (RO): An Observation is declared as a "RO"if $|\mathrm{MGt}_i| \leq 2.5$ and $p_{ii}^* \leq \mathrm{Median}\,(p_{ii}^*) + c\,\mathrm{MAD}\,(p_{ii}^*)$
2. Vertical Outlier (VO): An Observation is declared as a "VO" if $|\mathrm{MGt}_i| > 2.5$ and $p_{ii}^* \leq \mathrm{Median}\,(p_{ii}^*) + c\,\mathrm{MAD}\,(p_{ii}^*)$
3. GLPs: An Observation is declared as a GLP if $|\mathrm{MGt}_i| \leq 2.5$ and $p_{ii}^* > Median\,(p_{ii}^*) + c\,\mathrm{MAD}\,(p_{ii}^*)$
4. BLPs: An Observation is declared as a BLP if $|\mathrm{MGt}_i| > 2.5$ and $p_{ii}^* > Median\,(p_{ii}^*) + c\,\mathrm{MAD}\,(p_{ii}^*)$

The Aircraft dataset, which is taken from Gray (1985) is used to illustrate the merit of our proposed plot. This dataset contains 23 cases with four predictor variables (Aspect ratio, Lift-to-drag ratio, Weight of the plane, and Maximal thrust) and the response variable is the Cost. The classification of data into regular data, vertical outliers, good and bad leverage points are shown in Figures 3, 4 and 5. It can be observed from Figure 3 that the non-robust plot (OLS-MD) identified one vertical outlier (case 22) and one GLP (case 14). The LMS-RMD plot in Figure 4 detected one vertical outlier (case 16), BLP (case 22) and 2 GLP (cases 14, 20), while the MGt-DRGP plot in Figure 5 identified one vertical outlier (case 16), two BLPs (cases 19 and 22) and one GLP (case 21).

As shown by Mohamed *et al.* (2015a), most of time the classical OLS-MD plot fails to correctly identifies the BLPs. The robust LMS-RMD plot is also not very successful in classifying

observations into four categories. Our new developed MGt-DRGP plot consistently is very successful in classifying observations into regular observations, vertical outliers, good and bad leverage points.

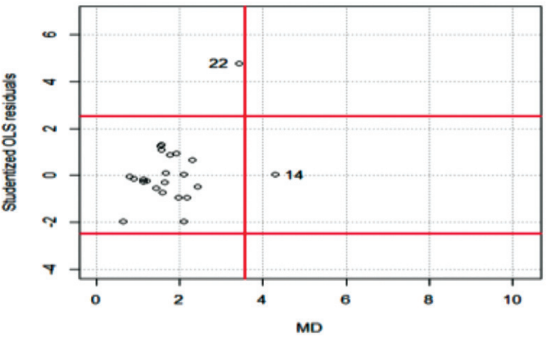

Figure 3 The Studentised OLS res. vs. MD for the Aircraft data

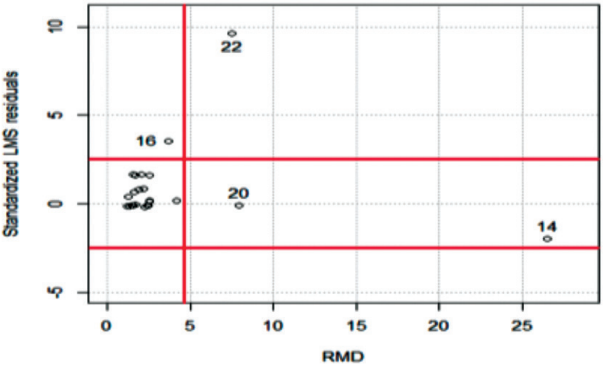

Figure 4 The Standardised LMS res. vs. RMD for the Aircraft data

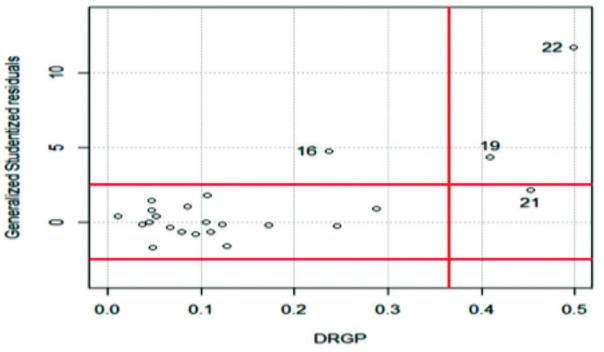

Figure 5 The Mod. Generalised studentised. res. vs. DRGP for the Aircraft data

ROBUST PARAMETER ESTIMATIONS

Robust Jackknife Ridge Regression to Combat Multicollinearity and High Leverage Points

Introduction

Consider the following standard multiple linear regression model:

$$y = X\beta + u$$

it is assumed that y is an ($n \times 1$) vector of the dependent variable, X is an ($n \times p$) and full rank matrix of regressor variables, β is a ($p \times 1$) vector of an unknown regression parameters and u is an ($n \times 1$) vector of the error term with elements are assumed to be independently and normally and identically distributed random variables, such that $E(u) = 0$ and the dispersion matrix $E(uu') = \sigma^2 I$. For the purpose of convenience, it is assumed that all variables are standardised so that the design matrix $X'X$ is in correlation form. The OLS estimator, namely

$$\hat{\beta}_{LS} = (X'X)^{-1} X'y$$

has optimal properties under Gaussian-Markov assumptions. Let $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$ be the matrix of eigenvalues for $X'X$ and γ is a (p×p) matrix of its corresponding eigenvectors whose column are normalized with $\gamma'\gamma = \gamma\gamma' = I$. According to Singh et al. (1986), the linear regression model can be written in canonical form as,

$$y = Z\alpha + u$$

where Z = Xγ and $\alpha = \gamma'\beta$. Since $\gamma'\gamma$, hence Z'Z = γ'X'Xγ = Λ. The OLS estimator for α is given by

$$\hat{\alpha}_{LS} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y$$

since $\alpha = \gamma'\beta$, then $\hat{\beta}_{LS}$ can be written as

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y$$

The MSE for the OLS estimator is given by

$$MSE(\hat{\beta}_{LS}) = MSE(\hat{\alpha}_{LS}) = \sigma^2\Lambda^{-1}$$

Hoerl and Kennard (1970) showed that a solution to the OLS does not always exist and there is no unique solution when the matrix $X'X$ is ill-conditioned (not invertible) due to the multicollinearity problem. Multicollinearity is a major problem in multiple regression, this issues occurs when two or more regressors are highly correlated. In this situation the standard errors of the OLS estimates become large and often the results are confusing and may give misleading conclusions.

There are many methods to address this problem of multicollinearity. The most commonly used methods are Ridge Regression (RR), Latent Root Regression and Jackknife Ridge

Regression (JRR) (Hoerl and Kenard, 1970; Batah *et al*., 2008). However, these estimators are not robust to outliers and leverage point. Unfortunately, neither robust methods nor the RR technique alone is sufficient to address the complicated problem of multicollinearity and outliers (Habshah and Marina, 2007). To circumvent this combined problem, significant works have been done by integrating RR with the robust method to get an estimator that is much less influenced by multicollinearity and outliers. Jadhav and Kashid (2011) suggested using a Jackknife ridge M-estimator to overcome multicollinearity and outliers in the *Y* direction. However most of the suggested methods do not focus on the combined problem of multicollinearity and high leverage points (HLPs). As such, Mohammed *et al.* (2015b) developed two new methods known as Robust Ridge MM (RJMM) and Robust Jackknife Ridge GM2 (RJGM2). The RJGM2 is formulated by incorporating the Generalised M based on Minimum Volume Ellipsoid (GM2) developed by Bagheri and Habshah (2011) and the Jackknife Ridge Regression. Mohammed *et al.* (2015b) have shown that the RJGM2 estimate is given by

$$\hat{\beta}_{RJRR} = \gamma\, \hat{\alpha}_{RJRR}(k)$$

where,

$$\begin{aligned} \hat{\alpha}_{RJRR.}(k) &= [I + k B^{-1}]\tilde{\alpha}_{RRR} \\ &= [I + k B^{-1}][I - k B^{-1}]\tilde{\alpha} \qquad , \quad B = (\Lambda + kI_p) \\ &= (I - k^2 B^{-2})\tilde{\alpha} \end{aligned}$$

$$\hat{k} = \frac{p\tilde{\sigma}^2}{\tilde{\beta}'\tilde{\beta}} \qquad \textit{and} \qquad \tilde{\sigma}^2 = \frac{(y - X\tilde{\beta})'(y - X\tilde{\beta})}{\tilde{\beta}'\tilde{\beta}}$$

The Performance of RJMM and RJGM2

A simulation study is conducted to assess the performance of the proposed methods (RJMM and RJGM2) in the case of the simultaneous presence of the multicollinearity problem and HLPs in a data set. To generate simulated data with a different degree of multicollinearity, we apply a simulation approach given by Lawrence and Arthur (1990). We consider the multivariate linear regression model as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$$

where $\mathcal{E}$ is the error term distributed as $N(0, \sigma^2 I)$. The explanatory variables are generated by,

$$x_{ij} = \rho v_{i4} + (1-\rho^2)^{1/2} v_{ij}, \quad i = 1, 2, \ldots, n; \; j = 1, 2, \text{and } 3.$$

where v_{i1}, v_{i2}, v_{i3}, and v_{i4} are independent standard normal pseudo random numbers, and $p = 3$ is the number of explanatory variables. The explanatory variables are standardised so that the design matrix $X'X$ is in the canonical form. The character ρ^2 denotes the degree of collinearity between x_{ij} and x_{im} for $j \neq m$. In addition, three different values of high collinearity are selected corresponding to ρ = 0.90, 0.95 and 0.99, and four different sets of observations are considered corresponding to n = 20, 30, 50 and 100. The contamination is done by replacing a clean datum in the explanatory variables with HLPs corresponding to various ratios of the HLP, namely τ = 0.05, 0.10 and 0.15. Our proposed RJMM and RJGM2 estimators are compared with existing methods such as Ordinary Least Squares (OLS), Ridge Regression (RR), Jackknife Ridge Regression (JRR), Robust Ridge Regression

based on M-estimator (RRM) and Robust Jackknife Ridge Regression based on M-estimator (RJRM).

Due to space limitations, only one result is shown. However their performances are consistent. It can be observed from Table 8 that when the data have multicollinearity and HLPs,the values of RMSE and Loss for OLS, RR, and JRR are larger than the other robust estimator methods for all possible combinations of n, p and τ. The values of RMSE and Loss for RRM and RJRM are smaller than those for the classical estimator (OLS, RR and RR) but they are less efficient than RJMM and RJGM2 because RRM and RJRM depend on the M-estimator, which is known to be less efficient with HLPs, while the MM-estimator and the MGM2-estimator can do well with HLPs. RJMM and RJGM2 are the best methods in the presence of multicollinearity and HLPs. However, the performance of RJGM2 is better than that of RJMM in all possible cases except in the case of a small sample size, not very strong multicollinearity, and low and moderate HLP ratios ($n = 20$, $\rho = .90$ and $\tau = 0.05$ and 0.10). So, we can say that our proposed methods are the best methods for solving multicollinearity in the presence of HLPs and for producing estimates with lower RMSE and less bias.

Table 9 RMSE and Loss for estimation methods with $\tau = 0.10$ (ratio of HLPs)

	n	**20**		**30**		**50**		**100**	
	ρ	**RMSE**	**Loss**	**RMSE**	**Loss**	**RMSE**	**Loss**	**RMSE**	**Loss**
OLS	0.9	0.399	0.0182	0.3011	0.0176	0.2532	0.0176	0.2107	0.0174
RR		0.2614	0.0179	0.2083	0.0175	0.1898	0.0176	0.1772	0.0174
JRR		0.322	0.018	0.2462	0.0176	0.2147	0.0176	0.1908	0.0174
RRM		0.2142	0.0177	0.1727	0.0175	0.1674	0.0178	0.1635	0.0175
RJRM		0.2662	0.0178	0.1896	0.0174	0.1745	0.0176	0.165	0.0173
RJMM		0.1449	0.0039	0.1132	0.0031	0.0936	0.0028	0.0768	0.0026
RJGM2		0.1572	0.0037	0.097	0.003	0.0881	0.0025	0.0738	0.0026
OLS	0.95	0.5315	0.0188	0.3908	0.0177	0.3157	0.0176	0.2469	0.0174
RR		0.3237	0.0179	0.245	0.0175	0.2117	0.0175	0.1891	0.0173
JRR		0.416	0.0183	0.3065	0.0177	0.2538	0.0176	0.2129	0.0173
RRM		0.2386	0.0176	0.179	0.0174	0.1695	0.0174	0.1635	0.0173
RJRM		0.3161	0.0177	0.2088	0.0173	0.183	0.0173	0.1676	0.0172
RJMM		0.1691	0.0033	0.1257	0.0025	0.0995	0.0022	0.0779	0.002
RJGM2		0.1782	0.0034	0.0917	0.0023	0.0821	0.0025	0.0666	0.002

THE MODIFIED GM-ESTIMATOR BASED ON MGDFF FOR DATA HAVING MULTICOLLINEARITY DUE TO HIGH LEVERAGE POINTS

Introduction

Multicollinearity is a situation of multiple regression model when the independent variables are correlated with each other. However, it is now evident that high leverage points (HLPs) can cause multicollinearity problems (Imon, 2003; Bagheri and Habshah, 2012a). With their presence, VIF value becomes large and VIF value becomes small in their absence. Bagheri and Habshah (2012a) and Bagheri *et al.* (2012b) refer to these situation as High leverage collinearity enhancing observations and High Leverage Collinearity Reducing Observations, respectively. In the previous section, we have illustrated the second situation whereby for multicollinearity which is caused by correlated predictors, in the presence of HLPs, the VIF measure indicate no multicollinearity. We have shown that in such a situation, Robust Jackknife Ridge based on GM2 (RJGM2) is the best solution to remedy the multicollinearity problem. However, this method and any other methods that attempt to remedy multicollinearity problem are not appropriate when multicollinearity is due to HLPs. As such, Habshah *et al.* (2015) proposed a new estimation technique called modified GM-estimator (denoted by MGM) based on modified generalised DFFITS to overcome the multicollinearity problem. The MGM estimates are obtained by solving

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{\mathbf{y_i} - \mathbf{x_i^T}\hat{\beta}}{s\pi_i}\right) x_i = 0$$

where $\psi = \rho'$, is a derivative of redescending function and τ is a weight function aims to downweight high leverage points. Assuming that β_0 is the initial coefficient of the S-estimator, Habshah *et al.* (2015) derived the MGM-estimator from one-step Newton Raphson as

$$\hat{\beta}_{MGM} = \hat{\beta}_0 + (X^T \Psi X)^{-1} X^T W \psi\left(\frac{e_i}{w_i \hat{\tau}}\right) \hat{\tau}$$

where W is an n x n diagonal matrix with w_i , i=1,2,…,n ,

$$\Psi = \text{diag}\left[\psi'\left(\frac{e_i}{\hat{\tau} \times \pi_i}\right)\right]$$

where ψ' is a derivative of Huber's function ψ, the residuals e_i of S-estimator and scale of the residuals, $\hat{\tau} = c(1 + 5/(n-p))\,\text{Median}|e_i|$ and

$$\pi_i = min\left[1, \{\frac{\text{CP}_{\text{MGDFF}}}{\text{MGDFF}}\}\right], \quad i = 1,2,...,n,$$

$$\text{MGDFF}_i = \begin{cases} \sqrt{\dfrac{w^*_{ii(R)}}{1 - w^*_{ii(R)}}}\,\text{MGt}_i & \text{for } i \in R \\ \sqrt{\dfrac{w^*_{ii(R)}}{1 + w^*_{ii(R)}}}\,\text{MGt}_i & \text{for } i \notin R \end{cases}$$

$$\text{MGt}_i = \begin{cases} \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - w^*_{ii(R)}}}, & \text{for } i \in R \\ \dfrac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_R \sqrt{1 + w^*_{ii(R)}}}, & \text{for } i \notin R \end{cases}$$

The Performance of MGM

Two examples and Monte Carlo simulation study were used to investigate the performances of our proposed methods. In this section, we report A Monte Carlo simulation study to assess the performances of our new proposed method (MGM). We consider the following multivariate linear regression model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

where ε is the error term distributed as $N(0,1)$. In the simulation study, we generate an uncorrelated dataset distributed as $N(0,1)$ with three explanatory variables (p = 3), various size of samples (n=30, 50, 100, 200) and various percentage of contaminations (α = 0.05, 0.10). We also considered various explanatory variables (p = 4, 5, 10). The experiment of simulation was repeated 5000 times for consistency. In order to create good and bad leverage points, certain clean observations are replaced by contamination data. To create bad leverage points, the first 100 $\left(\frac{\alpha}{2}\right)$ percent observations for both x and y variables are replaced by contaminated observations distributed as $N(1,10)$. And, to create good leverage points, the last 100 $\left(\frac{\alpha}{2}\right)$ percent observations of x's variable are replaced by contaminated observations distributed as $N(1,10)$. The performance of MGM-estimator is compared with some existing methods such as OLS, ridge regression, MM and GM6. The assessments of the estimators are based on the standard deviation of the estimates and ratio of MSE of the estimator's compared with the OLS estimator for the uncontaminated data (Habshah, 1999; Riazoshams and Habshah, 2010). The MSE and the ratio of MSE are given by;

$$\text{MSE}(\hat{\beta}_j) = (\bar{\beta}_j - \beta_j)^2 + \frac{1}{m}\sum_{j=1}^{m}(\hat{\beta}_j - \bar{\beta}_j)^2$$

$$\text{ratio of MSE}(\hat{\beta}_j) = \frac{\text{MSE}(\hat{\beta}_{j.\text{OLS}})\,[\text{Clean data}]}{\text{MSE}(\hat{\beta}_j)} \times 100\%$$

where, *m* is the replications of simulation experiment. A good estimator is the one that has the smallest value of standard deviation and ratio closest to 100%. Tables 10 and 11 exhibit the VIF, SE and ratio values of the estimates. Due to space constraint the results for (*p*=4, 5 10 and n=20,100) are not shown. However, the results are consistent. It is interesting to observe the results of Tables 10 and 11. For uncontaminated data, the VIF's values are small which suggests that there is no multicollinearity problem in the data. Table 10 also indicates that the performances of all methods are equally good for clean data. The presence of high leverage points changes the situation dramatically. It can be seen from Table 11 that when a certain percentage of HLP's are added to the data, the VIF values become large which indicate that HLP have induced multicollinearity to the data. The high leverage points have changed the data from non-collinearity to collinearity evidenced by high values of VIF's. The performance of the OLS immediately becomes very poor. The ratio of the OLS estimator is much lower than the other estimators and it has the largest values of standard deviations of the estimates. It is interesting to observe from Table 11 that the ridge regression estimator also does not give good results. Although the results of the MM and GM6 estimators are fairly closed, the values of the SE and ratio for the MM estimator is consistently slightly smaller and slightly higher than the GM6, respectively, for all samples sizes. However, it is evident from the results that the MGM estimator consistently has the smallest SE and highest ratio, followed by the MM and GM6

estimators for all possible combinations of n and α. The MGM-estimator consistently provides the most efficient results when multicollinearity is due to HLP.

Table 10 The SE and Ratio of the estimated Ridge, GM6, MM and MGM for clean generated data set

n	Coef.	VIF	Ridge		GM6		MM		MGM	
			S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio
20	β_1	1.14	0.7662	94.96	0.7472	97.38	0.7352	98.97	0.7355	98.93
	β_2	1.11	0.6953	94.61	0.6784	96.96	0.6656	98.83	0.6655	98.84
	β_3	1.12	0.6812	94.47	0.6649	96.78	0.6504	98.94	0.6522	98.67
40	β_1	1.05	0.4432	95.17	0.4363	96.68	0.4275	98.67	0.4279	98.57
	β_2	1.06	0.4012	95.29	0.3916	97.63	0.3883	98.45	0.3877	98.61
	β_3	1.05	0.4911	96.95	0.4905	97.06	0.4835	98.47	0.4851	98.14
100	β_1	1.03	0.3072	94.56	0.2985	97.32	0.2932	99.08	0.2921	99.45
	β_2	1.02	0.2979	95.37	0.2936	96.76	0.2883	98.54	0.285	99.68
	β_3	1.02	0.2494	94.23	0.2407	97.63	0.2367	99.28	0.2351	99.96
200	β_1	1.01	0.2165	95.94	0.2133	97.37	0.2088	99.47	0.208	99.86
	β_2	1.01	0.2127	96.90	0.2087	98.75	0.2066	99.76	0.2069	99.61
	β_3	1.01	0.2145	96.69	0.2138	97.01	0.2083	99.57	0.2078	99.81

Table 11 The SE and Ratio of the estimated OLS, Ridge, GM6, MM and MGM for contamination generated data

α	Coef.	VIF	OLS		Ridge		GM6		MM		MGM	
			S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio	S.E	Ratio
						n = 40						
0.05	β_1	4357	1.1962	22.49	1.2232	22.01	0.3607	74.58	0.3415	78.78	0.2771	97.08
	β_2	4489	1.2128	20.91	1.3571	18.69	0.406	62.46	0.3217	78.83	0.2656	95.5
	β_3	4349	1.2211	22.21	1.2778	21.22	0.4053	66.92	0.3328	81.49	0.2832	95.75
0.1	β_1	8394	1.2393	21.87	1.2782	21.2	0.4119	65.79	0.3616	74.94	0.329	82.38
	β_2	8644	1.2607	20.34	1.2997	19.73	0.4185	61.27	0.3457	74.16	0.3226	79.49
	β_3	8608	1.2572	20.19	1.2663	20.04	0.43	59.02	0.3511	72.28	0.3219	78.85
						n = 200						
0.05	β_1	4023	1.0493	5.51	0.9665	5.98	0.0916	63.12	0.0831	69.56	0.0701	82.45
	β_2	3989	1.0373	6.27	1.1818	5.5	0.1083	60.03	0.0976	66.67	0.0788	82.56
	β_3	4060	1.0717	5.58	1.0381	5.77	0.0961	62.35	0.0859	69.67	0.0752	79.63
0.1	β_1	8147	1.0408	6.11	1.2926	4.92	0.1566	40.62	0.103	61.8	0.0903	70.44
	β_2	8088	1.0647	6.02	1.3696	4.68	0.1447	44.31	0.1031	62.17	0.0945	67.8
	β_3	7940	1.0617	5.53	1.2462	4.71	0.1237	47.47	0.0984	59.66	0.0871	67.42

Habshah Midi

TWO-STEPS ROBUST ESTIMATOR IN HETEROSCEDASTIC REGRESSION MODEL IN THE PRESENCE OF OUTLIERS

Introduction

A commonly used assumption in linear regression is the constancy of error variances or homoscedasticity, mainly because of which the OLS estimators retain the minimum variance property. In a real life situation it is really hard to believe that the error variances will remain constant and that is why the violation of this assumption which causes the heterogeneity of error variances or heteroscedasticity is more prevalent in nature. The main problem with the violation of homoscedasticity assumption is that the usual covariance matrix estimator of the OLS becomes biased and inconsistent.

A large body of literature is now available (Habshah, 2000; Kutner *et al*., 2005; Habshah *et al*., 2009a; Habshah *et al*., 2009b; Rana *et al*., 2012; Siraj-ud-doulah *et al*., 2012) for correcting the problem of heteroscedasticity. The correction for heteroscedasticity is very simple by means of the weighted least squares (WLS) if the form and magnitude of heteroscedasticity are known. The WLS is equivalent to perform the OLS on the transformed variables. Unfortunately, in practice, the form of heteroscedasticity is unknown, which makes the weighting approach impractical. When heteroscedasticity is caused by an incorrect functional form, it can be corrected by making variance-stabilising transformations of the dependent variables or by transforming both sides (Carroll and Ruppert, 1988). However, the transformation procedure might be complicated when dealing with more than one explanatory variable. Montgomery *et al.* (2001), Kutner *et al.* (2004), and others have tried to find the

appropriate weight to solve the heteroscedastic problem when the form of heteroscedasticity is unknown. White (1980) proposed the heteroscedasticity-consistent covariance matrix (HCCM) estimators in this regard. Different forms of HCCM estimators such as the HC0, HC1, HC2, HC3 and HC4 have been proposed (MacKinnon and White, 1985; Davidson and MacKinnon, 1993; and Cribari-Neto, 2004). However, there is no general agreement among statisticians about which of the five estimators of the HCCM (HC0, HC1, HC2, HC3, HC4) should be used (MacKinnon and White, 1985). Chatterjee and Hadi (2006) proposed an estimator which is weight based, but these weights depend on the known structure of the heteroscedastic data. Kutner *et al*. (2005) proposed estimators which do not depend on the known structure of the heteroscedastic data. But the main limitation of the Montgomery *et al.* (2001) estimator is that it cannot be applied to more than one regressor situation. The estimator proposed by Kutner *et al.* (2005) can be applied to more than one variable and it does not depend on the known form of heteroscedasticity, but we suspect this estimator is not outlier resistant.

The weighted least squares also suffer the same problem in the presence of outliers (Maronna *et al*., 2006). We also believe that the HCCM estimators should suffer from the same problem, as they are based on the OLS residuals. Generally speaking, none of the estimation techniques work well unless the effect of outliers in a heteroscedastic regression model is eliminated or reduced by robustifying the WLS or HCCM. Unfortunately, there is not much work in the literature that deals with the estimation of the regression parameters in the presence of both heteroscedasticity and outliers when the structure of heteroscedasticity is unknown. Although Habshah *et al*. (2009a) has proposed this type of robust estimation procedure, but their procedure can be applied to only one regressor.

In this presentation, Habshah *et al.* (2014) proposed a two-step robust weighted least squares (TSRWLS) estimator which can be applied for more than one regressor when the form of the heteroscedasticity is not known. Firstly, for solving the heteroscedastic problem, we estimate the robust initial weights following the idea of Kutner *et al.* (2005) and secondly, we estimate the parameters of the model based on Huber (1981) weighting function in order to reduce the effect of outliers. Habshah *et al.* (2014) summarised the TSRWLS algorithm in the following two steps. In step 1 we form the initial weight and in step 2 we obtain the final weight.

Step1:

i. Find the fitted values $\hat{y}_i$ and the residuals $\hat{\varepsilon}_i$ from the regression model by using the least trimmed of squares (LTS) method.

ii. Regress the absolute residuals, denoted as $\hat{s}_i$ where $s_i = |\hat{\varepsilon}_i|$, on $\hat{y}_i$ also by using the LTS method.

iii. Find the fitted values $\hat{s}_i$ from step 1(ii).

iv. The square of the inverse fitted values would form the initial robust weights, i.e., we obtain $w_{1i} = 1/(\hat{s}_i)^2$.

Step2:

The robust weighting function such as the Huber function (Huber, 1981), the Bisquare function (Tukey, 1977) and the Hampel function (Hampel, 1974) can be used to obtain the final weight. However, in this study, we will use the Huber's weights function which is defined as

$$w_{2i} = \begin{cases} 1 & |e_i| \le 1.345 \\ \dfrac{1.345}{|e_i|} & |e_i| > 1.345 \end{cases}$$

The constant 1.345 is called the tuning constant and e_i is the ith standardised residuals of the LTS obtained from step 1 (i). We multiply the weight w_{1i} with the weight w_{2i} to get the final weight w_i. Finally we perform a WLS regression using the final weights w_i. The regression coefficients obtained from this WLS are the desired estimate of the heteroscedastic multiple regression model in the presence of outliers.

The Performance of the TSRWLS

In this section, we consider a real data to evaluate the performance of the proposed TSRWLS method and compared with the OLS and Kutner *et al.* (2005) method that we call KNN.

Education Expenditure Data

These data are taken from Chatterjee and Hadi (2006) which consider the per capita income on education projected for 1975 as the response variable (Y) while the three explanatory variables are X_1, the per capita income in 1973; X_2, the number of residents per thousand under 18 years of age in 1974, and X_3, the number of residents per thousand living in urban areas in 1970 for all 30 states in USA. According to geographical regions based on the pre-assumption, the states are grouped in a sense that there exists a regional homogeneity. The four geographic regions (i) Northeast, (ii) North centre, (iii) South, and (iv) West. The LTS estimator detected that the observation 49 [Alaska (AK)] is an outlier. The residuals vs. fitted values of OLS (Standardised), KNN and TSRWLS are plotted with and without Alaska. The OLS plot without Alaska clearly indicates a violation of the constant variance assumption. However, the KNN and TSRWLS plot do not show any symmetrical shape like the OLS fit. It shows that for

this 'clean' data (without AK) the non-constancy of error variances is not reflected in KNN and TSRWLS. To see the effect of outliers, we include the observation Alaska and the resulting residuals and fitted values are plotted. We see that OLS residuals are affected in the presence of outliers, but the effect of AK observation is not substantial on KNN and TSRWLS estimators.

Modified Education Expenditure Data

In reality we often have to deal with multiple outliers. For this reason, we deliberately change four data points to generate big outliers. Our changed data points are cases 46, 47, 48 and 50 by taking the value from outside the well known 3- σ sigma normal distance in *Y* direction. In fact, we replace the data points of *Y* for observations 46, 47, 48 and 50 by $| y_{cont.} |$ where $y_{cont.}$ are generated as $\overline{y} \pm 9s_y$, with $\overline{y}$ and s_y as the respective mean and standard deviation of *Y*. In this situation, it is more likely that these points would become big outliers. With this modified data, now we have five outliers (since these data already contained one outlier, i.e., Alaska). When the LTS is employed to the data, all 5 outliers are identified.

The plots of the residuals against the fitted values of the OLS, KNN and TSRWLS for the modified data are illustrated in Figure 6(a) - 6(f). It is observed from Figure 6(a) and 6(b) that in the presence of outliers the patterns of residuals are completely destroyed. That is, the OLS and KNN are greatly affected by outliers and so they are not good estimators for the remedy of the heteroscedastic problem when outliers are present. It is interesting to note that in Figure 10(c), the TSRWLS shows the scatter plot of the residuals except the data points which are outliers. The residual-fitted plots without the 10% outliers for the OLS, KNN and the TSRWLS are shown in Figure 6(d) - 6(f). Figure 6(d) signifies

that the OLS cannot remedy the problem of heteroscedasticity but the KNN and proposed TSRWLS are successful as it is expected. It re-emphasises our concern that the KNN might be good in the absence of outliers whereas our proposed TSRWLS might be good in the presence or absence of outliers since it is keeping the scatter plot in both situations.

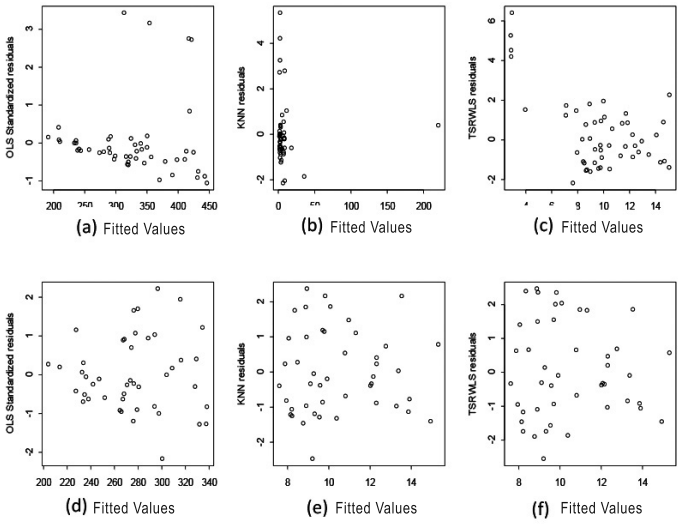

Figure 6 The OLS, KNN and TSRWLS fitted values vs. residuals plots with 10% outliers, (a)-(c) ; without 10% outliers, (d)-(f).

We know that graphical displays are always very subjective and that is why we would like to present some numerical summaries of the examples considered above. Here, we compare the performance of the proposed TSRWLS estimator with the existing estimators, such as the OLS, KNN and five versions of the HCCM estimators. Table 12 displays the summary statistics such as estimates of the parameters and their standard errors. It also

considers three different situations: when there are no outliers, with only one outlier (AK), and with 5 outliers. In the absence of outliers, all estimators perform equally in terms of parameter estimates and their standard errors and the resulting values are relatively close. But things change dramatically when outliers are present in the data. All estimators except the TSRWLS are strongly affected by outlier(s). We observe that the OLS and the KNN estimators not only have more bias in comparison to the TSRWLS, but also the sign of $\hat{\beta}_{3OLS}$ and $\hat{\beta}_{3KNN}$ have been changed in some occasions. By looking at the results of standard errors it is clear that both the OLS and the KNN estimators together with the five versions of HCCM (not shown) break down easily even in the presence of a single outlier. They produce much higher standard errors as compared with the TSRWLS estimator and things deteriorate when multiple outliers are present in the data. It can be concluded from Table 12 that the proposed TSRWLS is the best overall estimator as it possesses less bias and standard errors as compared to other estimators in the presence of heteroscedasticity and outliers. We have examined the performance of the proposed TSRWLS estimator and compare its performance with other existing estimators. Although the KNN, HCCMs and TSRWLS estimators are reasonably close to one another in the presence of heteroscedasticity with clean data, but the TSRWLS is the most reliable estimator as it possesses the least bias and standard errors. However, the performance of KNN and HCCMs are much inferior to the TSRWLS when contamination occurred in the data.

Table 12 Regression estimates of the Education Expenditure Data

		$\hat{\beta}_0$	$\hat{\beta}_1$	β_2	β_3
Without outliers	OLS	-277.577	0.0483	0.8869	0.0668
	KNN	-334.422	0.055	0.9809	0.0599
	TSRWLS	-283.24	0.0508	0.8827	0.0573
With AK outlier	OLS	-556.568	0.0724	1.5521	-0.0043
	KNN	-423.721	0.062	1.1782	0.0519
	TSRWLS	-365.479	0.0543	1.0779	0.0633
With multiple outliers	OLS	-452.07	0.0821	0.82	0.1936
	KNN	-536.69	0.1219	1.0639	-0.0983
	TSRWLS	-391.536	0.0605	1.0815	0.0626
Standard Errors of Estimators					
Without outliers	OLS	132.4229	0.0121	0.3311	0.0493
	KNN	108.2248	0.0111	0.2642	0.0419
	TSRWLS	105.9811	0.0106	0.2732	0.0422
With AK outlier	OLS	123.1953	0.0116	0.3147	0.0514
	KNN	96.883	0.0107	0.2313	0.0405
	TSRWLS	102.6924	0.0105	0.2486	0.0402
With multiple outliers	OLS	464.4632	0.0437	1.1864	0.1938
	KNN	182.047	0.0204	0.4591	0.0397
	TSRWLS	161.8082	0.017	0.3932	0.063

The empirical study reveals that the proposed estimator is outlier(s) resistant. Larger bias in estimates and standard errors, and smaller values of robustness measures clearly prove that the OLS, KNN and the five versions of HCCM are easily get

affected by outliers. To the contrary, both graphical and numerical evidences signify that the TSRWLS is capable of rectifying the problems of heteroscedasticity and outliers at the same time. Thus, the TSRWLS estimates emerge to be conspicuously more efficient and more reliable in comparison with other estimators considered in this inaugural lecture.

ROBUST PARAMETER ESTIMATION FOR LINEAR MODEL WITH AUTOCORRELATED ERRORS

Introduction

The Ordinary Least Squares (OLS) method is the most favorite technique for estimating the parameters of the multiple linear regression model because it is easy to understand and ease to apply. In many occasions, the assumptions of the Classical Linear Regression Model (CLRM) are taken for granted by statistics practitioners without any rigorous check. One of the importance assumptions that always being violated is the random and uncorrelated errors in the dataset. Autocorrelated errors cause the OLS estimators to lose their Best Linear Unbiased Estimators (BLUE) properties (White and Brisbon,1980). When the residuals are correlated with the previous errors which means $E(u_i, u_j) \neq 0$ for $i \neq j$, the variance $\hat{\sigma}^2$ is likely to be underestimated by the true σ^2. Consequently, less efficient estimates are obtained in the sense that the usual *t* and *F* tests of significance are no longer valid. These tests may show statistically significant when in fact it is not. The coefficient of determination, R^2 becomes inflated which wrongly indicates that the data fits the model well but in fact it is not. Hence, autocorrelated errors may provide misleading conclusions about the statistical significance of the regression coefficients (Gujarati and Porter, 2009). Therefore, appropriate

remedial measure must be taken after detecting the presence of autocorrelation problems.

In order to correct for autocorrelation and to obtain the parameters estimate, one often uses Generalized Least Square (GLS) procedures such as Cohrane Orcutt iterative method and Cochrane-Orcutt Prais-Wisten two-step or iterative procedures (Gujarati and Porter, 2009). Among these procedures the Cochrane-Orcutt Prais-Winsten (COPW) iterative method (Prais and Winsten, 1954) is the most popular measure in econometrics to obtain estimators with the optimum BLUE properties. Nonetheless, this procedure is based on the OLS estimates, which is not robust and therefore easily affected by high leverage points. Many statistics practitioners are unaware of the fact that high leverage points have an unduly effects on the OLS estimates. (Habshah *et al*., 2009; Riazosham *et al*., 2010)

Therefore Habshah *et al.* (2013) proposed a robust method for estimating the parameters of linear model with autocorrelated errors in the presence of high leverage points. The proposed robust method is formulated by incorporating the bounded influence, high asymptotic efficiency and high breakdown MM-estimator into the Cochrane-Orcutt Prais-Winsten (RCOPW) iterative method. This new procedure is named as Robust Cochrane-Orcutt Prais-Winsten (RCOPW) iterative method and the algorithm consists of six steps.

The parameters estimate of $\hat{\beta}^*$ in RCOPW iterative method can be expressed in the following matrix form:

$$\hat{\beta}^* = (\mathrm{X}^{*'}\mathrm{W}\mathrm{X}')^{-1}\mathrm{X}^{*''}\mathrm{W}\mathrm{y}^*$$

where W is the weights matrix of Iteratively Reweighted Least Squares (IRLS) in the MM estimator procedure. Lim

(2014) showed how the parameter estimate of $\hat{\beta}_0$ and $\hat{\beta}_j$ for $j = 1,2,3,..,k$ can be obtained.

The Performance of RCOPW

The robustness of RCOPW iterative method is examined by the Monte Carlo simulation study (not shown) and numerical example.

Time Series Data

We consider the Poverty data given by Murray (2006). The dataset contains 24 observations that gives U.S. Poverty Rates (*y*), Unemployment Rates (X_1) and GDP Growth Rates (X_2) from year 1980 to year 2003. Here the performance of COPW and RCOPW iterative methods are examined in the original data and in the presence of high leverage points. Three types of high leverage points are studied. The first type of the high leverage point is the high leverage in X_1 direction. A good observation in X_1 is simply replaced by a high leverage point. The second type of high leverage point is the high leverage point in X_2 direction. A good observation in X_2 is randomly replaced by a high leverage point. The third type of high leverage point is the high leverage point in both the X_1 and X_2 directions. For this case, a pair coordinates observation in X_1 and X_2 directions are randomly replaced by a high leverage point. There are many definitions of high leverage point. In this study, the high leverage point is taken as value which is beyond 3 deviation scope from its mean. The DRGP is applied to ensure that the contaminated data points are the high leverage points in the data.

Figure 7 shows the scatter plot of the current residuals (Res1) versus lagged residuals (Res(-1)) for the original data based on OLS estimation for dataset from 1980 to 2002. It can be seen very clearly from the residuals plot that the data has a strong positive

autocorrelation as many of the residuals are clustered in the first and the third quadrants of the plot. The OLS, COPW and RCOPW iterative methods are applied to estimate the regression coefficients for dataset from 1980 to 2002. The data in 2003 is used to compare the one step ahead forecast for the regression model based on these three estimations. The comparison of the parameters estimates obtained by COPW and RCOPW iterative methods are exhibited in Table 13. It is interesting to see that COPW and RCOPW are equally good when there is no contamination in the dataset. The estimated values and the standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained by COPW and RCOPW in the original data are almost the same.

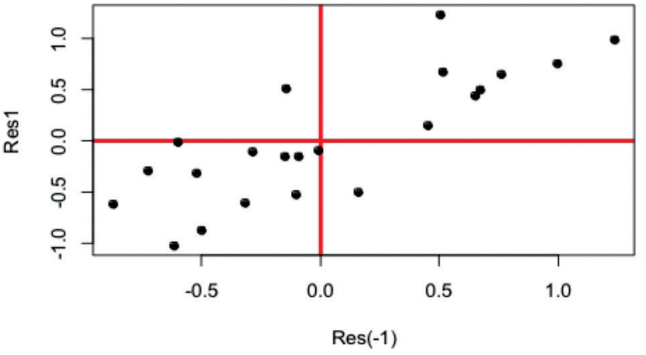

Figure 7 Current Residuals (Res1) Versus Lagged Residuals (Res(-1))

From the *p*-values of the MBG test, it can be seen that the autocorrelation problems are effectively corrected by RCOPW iterative procedure when there is a high leverage point contaminated in the data in all directions. The *p*-values become non significance after the RCOPW iterative procedure. But the COPW iterative method fails to correct the autocorrelation problems when there is a high leverage point in X_2. The *p*-value of MBG after COPW

iterative process becomes even smaller than before the iterative process takes place.

The estimators obtained by COPW estimation in contamination datasets can be completely different from that one obtained in the original dataset. The estimated value of $\hat{\beta}_1$ obtained by COPW estimation in the original data is 0.626. However, the COPW estimated value has changed drastically to 0.050 when there is a high leverage point in both X_1 and X_2 directions. Similarly, the estimated value of $\hat{\beta}_2$ obtained by COPW estimation in the original data is 0.067. Disappointedly, the estimations provided by COPW estimation when there is a high leverage point in X_1 and in both X_1 and X_2 directions have turned to negative values. The estimated values are -0.038 and -0.057 respectively.

Unlike COPW estimation, the parameters estimate obtained by RCOPW in high leverage datasets are very close to the parameters estimate obtained by RCOPW in the original datasets. The estimated value of $\hat{\beta}_1$ in the original data is 0.644 and the estimated values provided by RCOPW in the contaminated datasets are in the range (0.598 to 0.662). The estimated value of $\hat{\beta}_2$ obtained by RCOPW in the original data is 0.071 and the estimated values provided by RCOPW in the contaminated datasets are in the range (0.052 to 0.065). In addition, the standard errors of RCOPW parameters estimate are very much smaller than the one obtained by COPW estimation especially when there is a high leverage point in X_1 and also in X_1 and X_2 directions. This shows that RCOPW estimation provides a more consistent parameters estimate than COPW estimation. The regression model based on RCOPW estimation gives a very close one step ahead forecast to the actual value of y (12.50). The difference between the forecast values based on RCOPW regression model and the actual value of y is only around 0.10 unit. However, the difference is at least

0.14 unit if the regression model is based on COPW estimate. It is worth to mention that the OLS regression model which does not account for the nature of the autocorrelation gives a very far different step ahead forecast to the actual value of y, the difference is more than 0.80 unit. The results from this example show that the RCOPW estimation is the best method for correcting both autocorrelation and high leverage point's problems.

Table 13 Performance of COPW and RCOPW Iterative Methods in the Original and Modified Poverty Data

Estimated Values	Original Data			One High Leverage Point in X_1		
	OLS	**COPW**	**RCOPW**	**OLS**	**COPW**	**RCOPW**
$\hat{\beta}_1$	0.615	0.626	0.644	0.203	0.063	0.606
se	0.094	0.100	0.102	0.076	0.050	0.032
$\hat{\beta}_2$	0.106	0.067	0.071	0.038	-0.038	0.053
se	0.073	0.036	0.038	0.109	0.061	0.038
MBG (*p*-values)	9.25e-05	8.48e-01	8.74e-01	9.58e-03	9.00e-01	9.39e-01
One step ahead forecast y_{24} =12.50	13.35	12.36	12.46	13.38	12.21	12.34

Estimated Values	One High Leverage Point in X_2			One High Leverage Point in X_1 and X_2		
	OLS	COPW	RCOPW	OLS	COPW	RCOPW
$\hat{\beta}_1$	0.599	0.579	0.598	0.193	0.050	0.662
se	0.102	0.103	0.088	0.068	0.043	0.027
$\hat{\beta}_2$	0.023	0.017	0.065	-0.169	-0.057	0.052
se	0.050	0.022	0.020	0.092	0.056	0.035
MBG (*p*-values)	6.38e-05	5.86e-05	5.12e-01	5.57e-03	7.88e-01	7.89e-01
One step ahead forecast y_{24} =12.50	13.33	12.32	12.53	13.35	12.19	12.43

ROBUST TWO STAGE ESTIMATOR IN NONLINEAR REGRESSION WITH AUTOCORRELATED ERROR

Introduction

The Nonlinear model is commonly used by statistics practitioners in many applied sciences such as econometrics, engineering, biology and physical sciences to model a response variable to a set of independent variables (see Bates and Watt, 1988; Ratskowky, 1987;Seber and Wild, 2003).

Consider the general nonlinear model:

$$y = f(\theta) + \varepsilon$$

where $y = [y_1, y_2, ..., y_n]^T$ is $n \times 1$ response vector, $f(\theta) = [f(x_1;\theta), ..., f(x_n;\theta)]^T$ is $n \times 1$ vector of model function $f(x_i;\theta)$'s, $x_i = [x_{i1}, x_{i2}, ..., x_k]^T$ is predictor (design) vector and $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is $n \times 1$ vector of errors which are independent identical distributed (iid) with mean zero and unknown variance σ^2. The parameters of the model are often estimated by using the nonlinear least squares (NLLS) method because of tradition and ease of computation. Under the usual assumption, the NLLS estimates possess desirable properties. A commonly violated assumption is known as autocorrelated errors, occurs when the errors are correlated with the previous errors. This problem usually occurs in the situation when the data are collected over time (see White and Brisbon (1980)). Unfortunately many statistics practitioners are not aware that analysing such data based on the NLLS method posed many drawbacks. Seber and Wild (2003), proposed two stage estimator (CTS) to rectify this problem. Nevertheless, the problem is further complicated when the violation of the independent error terms come together with

the existence of outliers. It is now evident that outliers may have an unduly effect on the the NLLS estimates (see Habshah (1999)). By ignoring the outliers and erroneously assuming that the errors are independent, the NLLS technique is used for estimating the parameters. Consequently less efficient estimates are obtained as a result of employing an incorrect model on the erroneous assumption. The CTS method alone cannot rectify both problems of outliers and autocorrelated errors. This problem motivates us to establish a new and more efficient estimator that can rectify with these two problems simultaneously. However, the development of such method has not been published extensively in the literature. Sinha *et al*.(2003), proposed Generalized M (GM) estimator to estimate the parameters of the model when the errors follow autoregressive (AR) error process. Riazoshams *et al.* (2010) developed a new method that they call Robust Two Stage Estimator (RTS) to remedy the problem of autocorrelated errors which come together with the existence of outliers. The proposed method consists of two steps whereby in the second step, the RTS estimate is obtained by minimising

$$h(\theta) = \sum_{i=1}^{n} \rho\left(\frac{Ry_i - Rf(x_i;\theta)}{\sigma}\right)$$

where $\rho(.)$ is an influence function. For correlated errors, let V be positive definite correlation matrix of ε_i 's and the variance matrix of errors are denoted as $\text{var}(\varepsilon) = \sigma^2 V$. Let $V = U^T U$ be the Cholesky decomposition, where U is the upper triangular matrix and defined $R = (U^{-T})^{-1}$.

The Performance of RTS Estimator

In this section, the robustness of our newly proposed robust two stage estimator is assessed by using real life data set. This data set that we refer to as chloride data is taken from Bates and Watts (1988) which presents the relationship between the chloride concentration (%) and time. (see Sredni, 1970). They considered the following model for the data

$$f(x_i, \theta) = \theta_1(1 - \theta_2 e^{-\theta_3 x_i}) + \varepsilon_i$$

where the ε_i is the error terms which follows the AR(1) process. Nevertheless, Lin and Wei (2004) enumerated that the error terms follow a SAD(1) (Special Ante Dependence) error process, which is close to the AR(1) process. In order to see the effect of outliers, we deliberately changed three data points, that is the 2^{nd}, 3^{rd} and 4^{th} observations corresponding to y values (17.60,17.90,18.30) with higher values (20.60, 20.90, 21.30). The Nonlinear Least Squares (NLLS), Classical Two Stage (CTS) and Robust Two Stage (RTS) estimators were then applied to the original and the modified data. Tables 14 and 15 present the parameter estimates, the residual standard errors and the percentage variances accounted for, which are denoted as $100\overline{R}^2$ = 100[1-(residual mean square/total mean square)] for the original and the modified data.

It can be observed from Table 14 that when there is no outlier in the data, the three estimates are reasonably closed to each other. Nonetheless, as expected, the CTS estimator performs slightly better than the RTS and the NLLS as evidenced by its smallest residual errors. The results of Table 15 signify that the presence of outliers changes things dramatically. The NLLS and the CTS estimates immediately are affected by outliers. It can be seen that the residual standard errors of the NLLS and the CTS estimates

have increased markedly and their goodness of fit measures have decreased. The parameter estimates of the NLLS and the CTS have changed drastically. Nevertheless, the RTS seems to be only slightly affected by outliers revealed by the values of the RTS estimates, residual standard errors and the value of $100\overline{R}^2$, which seem to be only slightly changed. It looks like the NLLS estimator is easily affected by outliers and autocorrelated errors followed by the CTS.

Table 14 The parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\ \hat{\varphi}$ of the chloride data (original)

Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\varphi}$	$\hat{\sigma}$	$100\hat{R}$
NLLS	38.8653	0.8290	0.1606	---	0.2017	99.71078
CTS	38.8443	0.8258	0.1600	0.654	0.1991	98.75913
RTS	39.2077	0.8230	0.1559	0.643	0.2016	99.98373

Table 15 The parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\ \hat{\varphi}$ of the chloride data (modified)

Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\varphi}$	$\hat{\sigma}$	$100\hat{R}$
NLLS	65.2632	0.8173	0.0528	---	0.6477	96.49654
CTS	52.5806	0.7895	0.0771	0.499	0.6391	89.89496
RTS	38.4889	0.8151	0.1611	0.794	0.3085	99.97967

Monte Carlo Simulation

Here we report a Monte Carlo simulation study that is designed to assess the performance of the RTS estimates. The simulation study was carried out as follows. We considered a logistic growth curve model with the following function

$$y_i = \frac{2570}{1+41e^{-0.1x_i}} + \varepsilon_i, i = 1,...,n$$

Where x_i is uniformly distributed on interval [3, 51]. In this simulation study, we considered different sample sizes that varied from 20, 50, and 100 and different errors processes, that is AR(1), AR(2) and AR(3). However, we only show the results for AR(1) process. For AR(1) process, we considered $\phi = -0.3$, $\sigma_a = 50$, $\sigma^2 = \sigma_a^2 / (1-\phi^2)$. We then generate $\varepsilon_1 \sim N(0, \sigma^2)$ $\alpha_i \sim N(0, \sigma_a^2), i = 2, \ldots, n,$ and the remaining errors are computed from recursion relation.

In order to study the effect of outliers on the NLLS, CTS and RTS estimates, the data were contaminated with different percentage of outliers, that is 5%, 10%, 15% and 20%. The contaminated data points were generated following Fox (1972) algorithm by using Type I outlier, where the replacement outliers (RO) technique is applied. The Bernouli process is used to isolate the outliers (see Marona *et al.*(2006)).

Due to time constraint, in each simulation run, there were 200 replications. The mean estimated values, the bias, the variance and the Root Mean Squared Error (RMSE) of each estimate were computed based on 200 runs. In order to simplify the presentation of the results, we only report the percentage robustness measure, that is the ratio of the (RMSEs) of the estimators compared with the CTS estimator for clean data which have autocorrelated errors.

For quick interpretation, graphical results for these robustness measures are presented in Figure 8. A good estimator is the one which has robustness measure, which is closest to 100%. It can be observed that when there is no outlier in the data, the robustness measures of the three methods are fairly closed to each other and they are closed to 100%. However, when contamination occurs in the data, the robustness measures for all estimates decreased irrespective of the sample sizes, percentage of outliers and type of autocorrelation process.

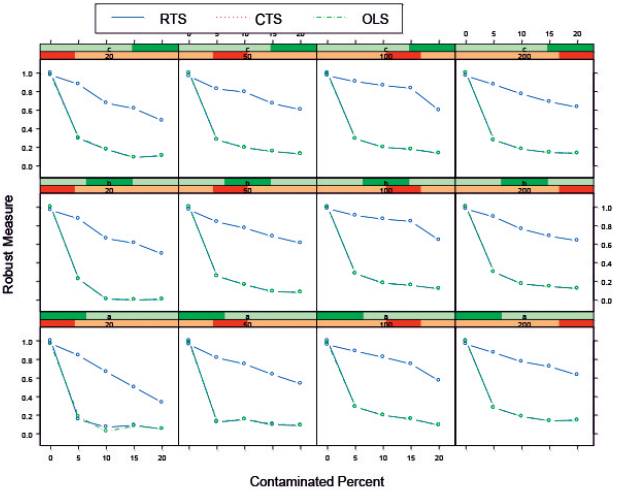

Figure 8 Robustness Measure (%) of the three estimates, AR(1) process

The RTS method outperforms the CTS and the NLLS method evident by its highest values of robustness measures for all the simulation runs. It is worth mentioning that the robustness measure of all estimates is decreased with an increased in the percentage of outliers. The results seem to be uniform for different sample of size n= 20, 50, 100 and 200, and different percentage of outliers. These results agree reasonably well with the results of real data that the RTS emerges to be the most efficient estimator, followed by the CTS and the NLLS when both problems of outliers and autocorrelated errors occur together. It seems that the performances of the CTS and the RTS estimators are equally good in a well behaved data and they are slightly better than the NLLS. The CTS is a good technique for correcting autocorrelated errors but it is easily affected by outliers. Thus, in this situation, it is not reliable. In this paper, we proposed a RTS method where it can remedy both problems of outliers and autocorrelated errors at

the same time. The numerical example and simulation experiment indicate that the RTS is more efficient than the NLLS and CTS for handling the problems of outliers and autocorrelated errors.

ROBUST CENTERING IN THE FIXED EFFECT PANEL DATA

Introduction

Panel data refers to the pooling of observations on a cross-section of households, countries, firms, etc. over multiple time series (Baltagi, 2005). For the past decade, there has been an increasing trend on the use of panel data in the research of economics and finance.

The fixed effect linear panel data model can be formulated as below:

$$y_{it} = \alpha_i + x'_{it}\beta = \varepsilon_{it} \qquad (1)$$

where $i = 1, ..., n$ are individual units observed at time series $t = 1,..., T$. y_{it} is the dependent variable, α_i are the unobservable time-invariant individual effects, β Is $K \times 1$ and x_{it} is the i-th observation on K explanatory variables. The ε_{it} denote the error terms which are assumed to be uncorrelated across time and individual units. The assumption of strict no endogeneity is applied.

The classical Within Groups estimator is obtained by firstly transformed the data within each time series by the mean:

$$\tilde{y}_{it} = y_{it} - \frac{1}{n}\sum_{t=1}^{n} y_{it}$$

and

$$\tilde{x}_{it} = x_{it} - \frac{1}{n}\sum_{t=1}^{n} x_{it}$$

The procedure is known as data centering and became an essential part by which the unobserved individual effects are eliminated. It follows from (1) that:

$$\tilde{y}_{it} = \tilde{x}_{it}\beta + u_{it}$$

where u_{it} is the new error term. Thus, the classical Within Estimator, $\hat{\beta}$w can be determined by the OLS method which minimises the function:

$$\sum_{t=1}^{n}(\tilde{y}_{it} - \tilde{x}'_{it}\beta)^2 = \sum_{t=1}^{n}(r_{it})^2$$

As already mentioned, the OLS produces the best linear unbiased estimator (BLUE) under the usual assumptions of normally distributed, independent and identically distributed errors. However, outliers can immediately alter the normal setting of the data and lead to unreliable estimates of the model. The damaging effect of outliers can be more crucial for the Within Group estimator. The classical data transformations will introduce a lot more outliers into the transformed data due to the non-robust property of the mean. Data in the contaminated time series will be affected in which the values will be greatly inflated or deflated. Thus, a robust data transformation is required to rectify this problem. Bramati and Croux (2007) and Verardi and Wagner (2011) replaced the centering by the mean with the median centering:

$$\hat{y}_{it} = y_{it} - median\{y_{it}\}$$

and

$$\hat{x}_{it} = x_{it} - median\{x_{it}\}$$

for $1 \leq i \leq n$ and $1 \leq t \leq T$. Median is chosen simply because it is the simplest robust measure to be derived and also due to its min max property. Median also has the highest breakdown point, in which data can be contaminated up to 50% before the estimate becomes useless. Once data has been robustly transformed by the median, Bramati and Croux (2007) employed the Robust Within Group GM-estimator (RWGM) to estimate the parameters of the panel data model. Generally, the GM-estimators are solutions to normal equations:

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{\mathbf{y_i} - \mathbf{x_i}^{\mathrm{T}} \hat{\beta}}{s\pi_i}\right) x_{\mathbf{i}} = 0$$

In this presentation, Robust Within Group MM-estimator (RWMM) and RWGM based on MM centering are proposed. It is important to note that prior to utilising the proposed estimators, the data centering procedures need to be employed. As already mentioned, the commonly used mean centering procedure is very sensitive to outliers. As an alternative, the median centering is put forward. However, centering by the median produces nonlinearity to the resulting data and affects the equivariance properties of the robust estimators (Bramati and Croux, 2007). Moreover, in an uncontaminated data, median is known to be less efficient than the mean (Maronna *et al.* 2006). This will certainly affect the efficiency of robust estimators in the absence of outliers. Thus, different type of robust centering is proposed in order to bring back linearity into the transformed data and at the same time provide more efficiency. Hence Midi and Bakar (2015) proposed centering to be done by MM-estimate of location called MM-centering. The proposed centering procedure is incorporated in the establishment of the proposed robust Within Group Estimator.

The proposed Robust Within Group estimator is summarised in two steps as follows;

Step 1 : Employ the proposed MM-centering procedure to the data.

Step 2 : Estimate the parameter of the panel data by using the RWGM proposed by Bramati and Croux (2007) or using our proposed RWMM.

The Performance of RWGM and RWMM

The performances of the MM-centering will be compared to the median centering for the two robust estimators by the Monte Carlo simulation. Following Bramati and Croux (2007), the dependent variable is set to accord the fixed effect linear panel model by generating $\varepsilon_{it} \sim N(0,10)$, $\alpha \sim U(0,20)$ and the vector of the slope coefficients β set equal to a vector of ones. The explanatory variables are generated from a multivariate standard normal distribution where 1 is a $K \times 1$ vector of ones.

Data are contaminated either randomly over all observations (random contamination) or concentrating the contamination in a few times series (block concentrated contamination). Both types of contaminations are done at two different locations; in y-direction and x-direction or leverage. All together, four different types of contamination cases are studied; vertical outliers, leverage, block concentrated vertical outliers and block concentrated leverage, at 5% and 10% level of contamination. The non-contaminated case is also studied for comparisons. For the block concentrated contamination, a few time series are randomly selected from the panel data set and be contaminated only up to 50% as suggested by Bramati and Croux (2007).

Vertical outliers or outliers in the *y*-direction are generated by inflating the randomly chosen *y*'s from a few time series with ~*N*(20.1). Further, to generate block concentrated *x*-outliers or the leverage points, we inflate *x*'s of the contaminated 's with data points from *K*-variate normal distribution $N(10 \times 1,1)$. This is done in order to create influential leverage points. In the experiments, we considered panel datasets of T = 5, 10, 15 and 20; representing small, medium, and large time series, each with n = 25,50,100 and 200 units for small, medium and large samples. Univariate regression is considered where K = 1 with M = 1000 Monte Carlo replications.

Once panel datasets are generated, data are immediately transformed by applying the classical mean centering and two other types of robust centering procedures - the median centering and MM-centering. The classical β coefficients are estimated by the OLS and the robust coefficients are estimated by the RWGM and RWMM estimators. The average mean square error (MSE) for each case is calculated by comparing the robust estimator's parameter estimates to the true parameter values using the formula:

$$MSE(\hat{\beta}) = \frac{1}{M} \sum_{j=1}^{M} (\hat{\beta}^{(j)} - \beta)^2$$

where $\hat{\beta}^{(j)}$ is the estimated slope in *j*th-replication. The root mean square error (RMSE) is given by $[MSE(\hat{\beta})]^{1/2}$. Following Riazoshams *et al.* (2010) the performance of each technique is evaluated based on the percentage of robustness measures using the ratio of the RMSEs of the estimators compared with the WG-mean centering based estimator for the good data. The robustness measures for different types of contaminations are presented. High percentage indicates the improved performance of the robust estimators. The robustness measures of the simulated panel data

set in the uncontaminated data is not shown. The overall results show that the RWMM provide better estimates than the RWGM in the uncontaminated panel data. Hence, RWMM under MM-centering provides the most efficient and consistent results for the uncontaminated data.

Results for block concentrated contaminations are produced in Table 16 for vertical outliers. It is observed that the classical WG estimations are largely affected in both types of outliers; only less severe when contaminated vertically. On the other hand, both RWGM and RWMM estimators are able to provide improved estimations under the two robust centering. Their performances are seen to increase with the increase of number of time series, T but rather low under the median centering. More stable and greater performances are observed for the robust estimations under MM-centering compared to robust estimations under median centering.

Similar results are obtained when blocks or time series are contaminated in the x-direction. Leverage points are known to cause severe effects to the classical estimates, resulting in low percentage on the robustness measures in all cases. Under the median centering, RWGM and RWMM are able to provide good results with increasing trend as T increases. Once again, the better results are found under the MM-centering regardless of the size of time series. It is also observed that RWMM performs more superior than RWGM under different types of robust centering and contamination levels.

The poor performances of robust estimators in the median-centred data may due to the non-linearity in the median transformed data. Under the robust MM-centering, linearity is brought back into the data and provided improved performances for both RWGM and RWMM. In both newly proposed robust centering, data are required to be centered close to the value of the mean in the non-

contaminated data. This also explains the increased performances for the uncontaminated data of both robust estimators, and hence the ability to provide efficient estimates under normality. MM-centering is found to provide more efficient, stable and consistent results to both RWGM and RWMM.

Simulation study indicates that data transformation under MM-centering provides more stable and superior results than transformation by the median. The performances of robust estimations under the newly proposed procedures have also improved vastly in small data sets, with small number of time series. Overall results showed that the performances of RWMM are more superior than RWGM under different types of contamination levels, sample size and number of time series.

Table 16 Robustness Measures (%) of simulated panel data sets for block concentrated vertical contamination

Cont. Level	N	T	WG	RWGM	RWMM	RWGM	RWMM
			Mean Centering	Median Centering		MM Centering	
Block Vertical 5%	25	5	9.2	50.3	60.6	83.7	91.6
		10	9.6	64.5	74.7	87.6	94.1
		15	10.9	67.4	75.4	89.7	94.9
		20	9.9	72.2	81.1	91.2	96.6
	100	5	8.9	28.7	36.4	83.9	90.4
		10	9.2	43.0	53.9	87.6	94.9
		15	10.1	43.9	53.5	87.6	94.4
		20	9.9	50.9	60.7	89.6	96.4
	200	5	8.9	20.6	26.5	82.5	90.2
		10	9.4	32.9	42.1	90.0	95.2
		15	10.2	32.7	39.9	88.6	94.3
		20	10.5	41.4	50.8	90.9	96.1

Block	25	5	6.6	49.5	57.8	81.0	87.5
Vertical		10	6.5	63.9	73.0	88.9	94.1
10%		15	7.4	65.8	72.8	89.8	94.0
		20	7.4	72.7	79.2	91.4	95.1
	100	5	6.4	27.9	33.1	81.0	85.9
		10	6.6	41.3	49.6	87.4	91.6
		15	7.0	41.6	48.9	89.5	93.4
		20	7.2	48.5	56.2	88.8	92.5
	200	5	6.2	20.5	25.0	83.4	87.7
		10	7.0	33.1	40.6	91.4	95.1
		15	7.2	31.3	37.2	88.2	92.1
		20	7.5	38.9	46.2	90.3	93.9

ROBUST ESTIMATOR IN RESPONSE SURFACE DESIGN WITH HETEROSCEDASTIC CONDITIONS

Introduction

Response Surface Methodology (RSM) which was first introduced by Box and Wilson in 1951 (Hill and Hunter, 1966) involves the use of statistical and mathematical tools for modelling and analysing a problem in which a response variable of interest is influenced by several variables. The main objective of RSM is to optimise the response and to find the combination of conditions that provides the highest response. RSM helps industrial world to realise how several input variables potentially influence some performance measures of a process and product. The relationship between a set of independent variables (also known as *control*, or *input* variables) and a response is determined by a mathematical model called regression model. Multiple regression analysis is one of the regression models useful for modelling and analysing the relationship between a response and control variables required in RSM. In general, regression analysis is routinely applied in most applied sciences to observe the change in the response variable by changing any one of the control variables in the situation that the control variables are considered to be fixed. One of the predominant regression analysis techniques in RSM is Ordinary Least Squares Method (OLS). The popularity of OLS in industrial applications is due to its easy computation, universal acceptance, and elegant statistical properties.

In applications, the normality of error distribution assumption will be inefficient in the presence of outlying observations in a data set resulting in less reliable estimates of the model parameters (Montgomery *et al*., 2001; Kutner *et al*., 2004; Montgomery, 2009). The first step in RSM is to construct an approximation

model for the response y. This approximation model is usually the second-order polynomial model to be fitted between the response variable (quality characteristics) and a number of input variables. The main aim is to find the best optimal settings of interest for the input variables or the best values of design parameters that optimise the response variable. Typically the main emphasis is on optimising (minimises or maximises) the mean (location) value of y where the variance (scale) is assumed to be small and constant. These assumptions may not be valid in real-life practice. Nonetheless, only constructing a response surface model for the mean may not be adequate and optimisation result can be misleading. Robust design is one of the most important process and quality improvement methods that focus on determining the optimum operating conditions with the ultimate aim of minimising variations in the quality characteristics while keeping a process mean at the customer-identified target value. Originally, RSM was designed to address only single response, but many real lives industrial applications involve optimisation of more than one response variables. Therefore, the dual response approach (developed by Myers and Carter, 1973) is used to tackle such problem (see Vining and Myers, 1990; Park and Cho, 2003; Shaibu and Cho, 2009). Basically in dual response surface optimisation, two models are established for the mean and for the standard deviation of the response y. Then the two fitted response models are optimised simultaneously in a region of interest. The experiments are repeated m times to measure the variability of y.

The OLS method is often used to estimate the parameters of the models. It is important to mention that the OLS regression estimates which are often used in RSM are also not appropriate for real-world industrial problems containing outliers. The problems get more complicated when outliers and heteroscedastic

errors come together. Goethals and Cho (2011) employed the Reweighted Least Squares (RLS) method to estimate the model parameters when the assumptions of constant error variances are violated. Although the RLS based method can rectify the heteroscedastic error, but it is not robust when outliers occur in the data. In this situation, the RLS based method cannot handle both problems at the same time. We need to improve this method that can remedy the problem of heteroscedastic errors and dampen the effects of outliers. In this respect, Shafie (2015) proposed to incorporate robust MM estimator in the formulation of the Two-Stage Robust (TSR-MM based) procedure. The TSR-MM based method consists of two steps whereby the $\hat{\beta}_{TSR-MM}$ estimate is obtained by minimising;

$$\min_{\hat{\beta}_{TSR-MM}} \sum_{i=1}^{n} \rho\left(\frac{y_i^* - f^*(x_i;\beta)}{\sigma}\right)$$ using MM estimation technique

where

$y_i^* = y_i \times w_i$, $f_i^*(x_i,\beta) = f(x_i,\beta) \times w_i$, and $\varepsilon_i^* = \varepsilon_i \times w_i$,

The weight is defined as the square of the inverse fitted values of $\hat{s}_i$, $w_i = \frac{1}{\hat{s}_i^2}$ (obtained from the first step). Subsequently, they employed the TSR-MM based method to estimate the parameters of the second-order polynomial models for the process mean $(\bar{y})$ and process standard deviation (s) of the response y. The fitted response functions for the process mean and process standard deviation are as follows:

$$\hat{\mu}(x) = b_{0(TSR-MM)} + x'b_{TSR-mm} + x'B_{TSR-MM}x$$

$$\hat{\sigma}^2(x) = c_{0(TSR-MM)} + x'c_{TSR-MM} + x'C_{TSR-MM}x$$

where $b_{0(TSR-MM)}, b_{TSR-MM}, B_{TSR-MM}, c_{0(TSR-MM)}, c_{TSR-MM}, C_{TSR-MM}$ were estimates of the coefficients based on TSR-MM estimator.

The usual method in replicated responses problem is to firstly compute the sample mean and sample standard deviation of *y* and construct the process mean and process standard deviation functions. Once the fitted response function for the process mean and process variance have been established, the optimum operating conditions of control factors are obtained by minimising the following

$$\text{minimise } MSE = \hat{\sigma}^2(x) + (\hat{\mu}(x) - t_0)^2$$

where t_0 is the customer-identified target value for the quality characteristics of interest.

The performance of TSR-MM based estimator

In this section, we report a Monte Carlo simulation study that is designed to assess the performance of the TSR-MM based estimator. In this simulation study, firstly, the responses *Y* were generated randomly from a normal distribution. Following Park and Cho (2003), five responses are generated from each distribution with $\mu(x_i)$ and $\sigma(x_i)$ at each control factor settings $x_i = (x_{i1}, x_{i2}, x_{i3}), i = 1, 2, \ldots, 27$. The total number of iterations is 500, each having 27 design points, and 135 responses. $\mu(x)$ and

$\sigma^2(x)$ are given as follows:

$$\mu(x) = 50 + 5(x_1^2 + x_2^2 + x_3^2), \quad \sigma^2(x) = 100 + 5\{(x_1 - 0.5)^2 + x_2^2 + x_3^2\}$$

Secondly, to see how the lack of a normal distribution affects the estimators, the response Y are also generated from other distribution such as double exponential distribution, which has heavier tails distribution that is prone to produce a few outliers. $\sigma^2(x)$ is generated accordingly to induce heteroscedasticity of the error variances. To further investigate the effect of outliers, the data were contaminated by generating outliers. Since the OLS model is known to be not reliable in the presence of outliers, it is not included in the comparison. For each distribution specified above, two statistical measures such as bias and mean squared error (MSE) using RLS and TSR-MM based methods were considered as decision criteria to judge the performance of the estimators. The result of Breusch-Pagan test indicates that the error variance of this experiment is not constant. Table 17 illustrates the estimated bias and MSE of the optimal mean response $\hat{\mu}(x)$ for response surface model with heteroscedastic errors based on RLS and TSR-MM based methods. Assuming that the target value for this experiment is $t_0=50$. It can be observed that in the presence of heteroscedascity and without contaminated data, as expected, the RLS based estimate is slightly better than the TSR-MM based. However, for non-normal data having heteroscedastic errors, the TSR-MM based method is more efficient than the RLS based method evidence by having smaller bias and MSE.

Table 17 Estimated Bias and MSE of the Estimated Optimal Mean Response for Heteroscedascity Data Using RLS based and TSR-MM based Methods

	RLS based		TSR-MM based	
Distribution	Bias	MSE	Bias	MSE
Normal	3.83	24.90	3.90	26.12
Normal (contaminated)	9.17	133.48	3.37	19.55
Double Exponential	4.65	40.95	4.15	29.54

Numerical Results

The merit of the newly proposed robust TSR-MM based estimator is assessed using numerical example.

Printing Process Data

This experiment introduced by Box and Draper (1987), was conducted to determine the effect of the three control variables: x_1 (speed), x_2 (pressure), and x_3 (distance) on the characteristic of a printing process y, that is on the machine's index to apply colored inks to package labels (y_1, y_2, y_3). The experiment is a 3^3 factorial design with three replicates at each of the 27 design points. In order to see the effect of outliers in the heteroscedasticity data, we deliberately changed three response points, that is the 8th, 15th, and 27th observation corresponding to y_1 (259 to 9259), y_2 (568 to 8656), and y_3 (1161 to 11161). The plot of residual against fitted values suggests that there is a moderate heteroscedasticity problem. The result of Breusch Pagan test indicates that the error variances of this experiment are not constant. The optimum response based on least-squares (OLS), Reweighted Least Squares (RLS based), and Two-Stage Robust (TSR-MM based) estimations were then applied to the data. Table 18 exhibits the estimated

optimum settings, mean, variance, and MSE of the estimated mean response. The mean squared error is obtained by the MSE relation where, $\text{MSE} = \hat{\sigma}^2(x) + (\hat{\mu}(x) - t_0)^2$ with $t_0 = 500$. It can be seen from Table 18 that the estimated mean response based on RLS achieves the target i.e. 500 and has the smallest value of MSE.

Table 18 The Estimated Optimum Settings, Mean, Variance, and MSE of the Estimated Mean Response

Model	x^*	Mean	Variance	MSE
OLS	(1.000, 0.060, -0.243)	494.657	1988.550	2017.099
RLS	(0.9966, 0.9967, -0.7190)	500	$8.043e^{-11}$	$5.161e^{-10}$
TSR-MM based	(1.000, 1.000, -1.000)	497.86	492.29	496.85

The results of Table 19 signify that in the presence of outliers, changes things dramatically. The OLS and RLS based immediately are affected by outliers. It can be seen that the standard errors of the OLS and RLS estimates increased markedly, and their objective target have deviated. Nevertheless, as expected, the TSR-MM based estimate only slightly affected by outliers revealed by smaller values of the standard errors, and MSE and achieve the objective target.

Table 19 The Estimated Optimum Settings, Mean, Variance, and MSE of the Estimated Mean Response for Modified dataset

Model	x^*	Mean	Variance	MSE
OLS	(-0.637, 0.353, 1.000)	342.01	14448.21	39407.22
RLS	(0.777, -1.000, 1.000)	444.97	12482.95	15511.21
TSR-MM based	(1.000, 0.1278, -0.3421)	497.62	793.13	798.81

It can be concluded that the performances of the optimum mean response of the RLS and the TSR-MM based estimators are equally good in a heteroscedascity data without outliers. The RLS based estimator is a good technique for solving heteroscedascity problem but it is easily affected by outliers. Hence, they are not reliable. The numerical example and simulation experiment indicate that the TSR-MM based method offers a substantial improvement over the other existing methods for handling the problems of outliers and heteroscedastic errors in response surface model.

ROBUST STABILITY BEST SUBSET SELECTION FOR AUTOCORRELATED ERRORS

Introduction

In the last part of this inaugural lecture, the issue on the variable selection technique for high dimensional data is discussed. It is now evident that the classical variable selection methods such as fitting all the possible subsets and using stepwise selection procedures failed to correctly select the important variables in the final model. Moreover, those procedures are not practical because they are very time consuming. The problem becomes

more complicated when autocorrelated errors come together with the existence of outliers in a data set. There are many variable selection techniques such as Forward Selection and Multi-Split procedures, but they do not discuss the issue of the combined problems of outliers and correlated errors. As such Hassan *et al.* (2015) developed a new robust variable selection technique that we call Robust Multi-Split –AIC (R.Multi-Split-AIC) and Robust Multi-Split-BIC(R.Multi-Split-BIC). Since the formulation of the proposed methods are very long and mathematically complex and also because of space limitations, we only describe the algorithm. The developed methods consist of three steps whereby in the first step, robust Cochrane-Orcutt method of Midi *et al.* (2013) is employed, followed by using $\sqrt{n}$ Reweighted Fast Consistent and High (RFCH) breakdown estimator which is developed by Olive and Hawkins, (2010). Finally, the BIC and AIC procedures are applied to the concentrated data (Hassan *et al*.,2015). It is very important to highlight that a good variable selection technique is the one that has the ability to correctly choose the important variables to be included in the final model so that it will have high predictive power. The merit of our proposed method is illustrated by using numerical example and simulation study.

The Performance of the Proposed Method

A simulation study that was designed to assess the performance of our proposed robust variable selection techniques is conducted under two different outlier scenarios. However, we only report one scenario. In this experiment, we consider multiple linear regression model with the following relation:

$$Y = 7X_1 + 6X_3 + 5X_4 + 7X_6 + 7X_9 + 0\,[X_D\,] + e$$

where D = 2,5,7,8,10. A design matrix was generated from a multivariate normal distribution with covariance structure $cov(X_j ; X_K) = \rho^{|j-k|}$ where $\rho = 0.5$, j,k =1,2,…..10 and n = 500. The random errors ε were drawn from a standard normal distribution.

To create the autocorrelation problem we considered the following setting:

$$\left.\begin{aligned} Y^* &= Y_{[2:n]} + \rho Y_{[1:(n-1)]} \\ X^* &= X_{[2:n]} + \rho X_{[1:(n-1)]} \end{aligned}\right\}$$

where ρ = 0.9 . As in (Agostinelli and Salibian-Barrera, 2010) outliers were generated by replacing 10% of the original values with high leverage points and vertical outliers. The vertical outliers were generated as asymmetric outliers, where ε = 0.10 and the errors were generated as e~$(1–\varepsilon)N(0,1) + \varepsilon N(20,1)$. To create the leverage points, each covariate was contaminated with 10% outlying observations generated from $N(50,1)$. For each case, we generated 500 independent simulated datasets. The problem of autocorrelated errors first be rectified and then randomly split each of the dataset into training n^{tr} (70%) and test n^{ts} sets (30%). The proposed robust stability selections (R.Multi Split-AIC and R.Multi Split-BIC), the existing stability selections (Multi Split-AIC and Multi Split-BIC) and the Single-split all-subsets-AIC and the single-split all-subsets-BIC methods were then applied to the training datasets. This process was repeated 500 times. The average Root Mean Squares Errors (RMSE) of the test sets over 500 simulation runs and the percentage chances for each variable of the training sets being selected in the final model over 500 simulation runs are presented. The potential variables being selected are also exhibited in the tables. The best method is the one that has the lowest RMSE and selects the correct variables (variables X1, X3, X4, X6, X9) in the final model with no noise variable. The results of the study show that when there is no

outlier in the data, all the six methods able to choose all the correct variables in the final model. The results indicate that our proposed method is comparable with other existing methods. Nevertheless, the results change dramatically in the presence of outliers in a data set. It can be observed from Table 20 that the classical Multi-Split-AIC and Multi-Split-BIC methods are much affected in the presence of high leverage and vertical outliers. Both methods have the highest RMSEs and tend to be underfitting. In this situation, both the Single-split-AIC and Single-split-BIC variable selection techniques also fail to select the correct variables.

Both methods tend to be over-fitting because they also select noise variables in the final model. It is interesting to observe that our proposed variable selection methods consistently have the least RMSE and successfully chosen the correct variables in the final models without selecting any noise variable.

Table 20 Selected variables, average RMSE, and percentage for each variable being selected for high leverage and vertical outliers (Threshold=71.41)

	Single-split-AIC	**Single-split-BIC**	**Multi-Split-AIC**	**Multi-Split -BIC**	**R. Multi-Split -AIC**	**R. Multi-Split -BIC**
RMSE	0.039	0.039	21.93	22.29	0.036	0.036
1	100	100	43.6	16.5	100	100
2	99.89	97.72	28.7	5.4	2.16	2.16
3	100	100	66.7	45.4	100	100
4	100	100	49.6	25.2	100	100
5	17.22	1.72	97.5	78.1	1.04	1.04
6	100	100	100	99.9	100	100
7	15.8	2.08	16.8	2.8	0.49	0.49
8	19.38	2.79	16.3	3.4	1.31	1.31
9	100	100	97.1	92.3	99.9	99.9
10	16.65	2.66	16.2	2.1 0	1.27	1.27
Selected Variables	1,2,3,4,6,9	1,2,3,4,6,9	5,6,9	5,6,9	1,3,4,6,9	1,3,4,6,9

Air Quality Data

In this study, an hourly air pollution data which are taken from the Department of Environment (DoE), Malaysia is used to further assess the performance of our method.

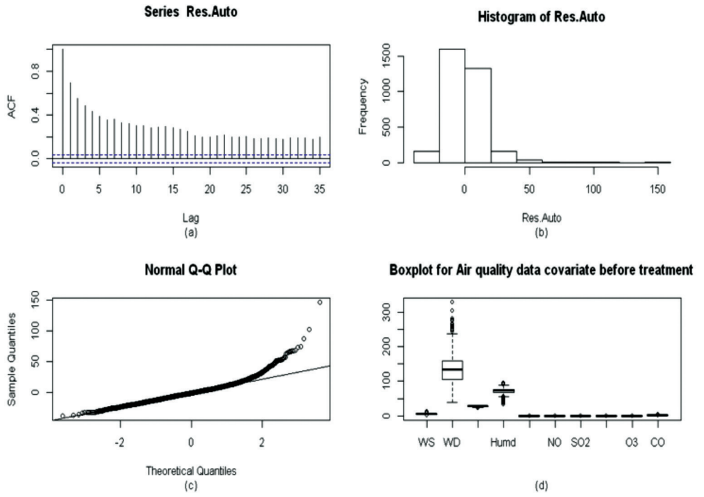

Figure 9 QQ-Plot, histogram of residuals and plot of PM10 vs each component of air quality data, Seberang Prai, Pinang

The data consists of the PM10 concentration and ten independent variables, of which six are pollutant variables (sulphur dioxide (SO2), nitrogen dioxide (NO2), nitrogen monoxide (NO), nitrogen oxide (NOx), carbon monoxide (CO) and ozone (O3)) and four are meteorological variables (wind speed (WS), wind direction (WD), temperature (Temp) and relative humidity (Hum)). PM10 is a particulate matter 10 micrometers or less in diameter of solid or semi-solid material found in the air. The value of each variable was recorded from the monitoring station at Seberang Perai, Penang on an hourly basis every day from January 2005

to December 2013. For the purpose of the statistical analysis, the hourly data were converted to a daily average, giving 3,287 readings. Missing values and calibration hours of certain variables are replaced by the coordinate medians for these variables.

Let us first observed the plots in Figure 9. Both the histogram (b) and the quantile–quantile (q-q) plot (c) of Figure 10 show that the residuals are contaminated with a heavy-tailed mixture distribution. Since some points in the qq-plot do not fall on the straight line and the histogram is skewed to the right, this indicates that this data is not normal. Thus, we suspect that there are outliers in this dataset. Figure 9(d) also shows that there are some leverage points in each covariate. Figure 9(a) indicates the existence of autocorrelation or serial correlation between the residuals, and it seems that there is a high order auto-regression AR(P).

Our proposed robust R.Multi-Split-AIC and R.Multi-Split-BIC and the existing methods were then applied to the data (3287 observations) to investigate which important variables influenced PM10. The dataset consists of 3287 observations, which include the PM10 as the response variable and the ten independent variables already mentioned. Since the air quality data are taken in time sequence, the Durbin Watson (DW) test is applied to the data to check the existence of autocorrelation problem. The results of Durbin Watson statistics for the original air quality data ($p<0.01$) confirmed the existence of autocorrelation and no autocorrelation ($p>0.05$) after treating the autocorrelation problem. After correcting the autocorrelation problem, the data is then randomly divided into training (70%) and test sets (30%).

This process is repeated 3,000 times. The RFCH is used to concentrate the training and test set data. Following Meinshausen and Bühlmann (2010), each training and each test set are split randomly into two sets of equal size and this process is repeated

50 times. The six variable selection methods were then applied to the first part of the training data set. The variables that are selected in the final model are determined. The best method is the one that has the lowest average of RMSE.

The results in Table 21 show that the RMSE of our proposed method, based on both AIC and BIC, is the smallest compared to the existing methods. This suggests that our proposed method correctly identified the potential variables, namely WD, Temp, Hum, SO2, NO2, O3 and CO, to be included in the final model. The Single-split-AIC method selects eight covariates, while the single-split-BIC method selects only six covariates. The classical Multi-Split-AIC selects seven covariates and Multi-Split-BIC selects five covariates.

It is interesting to observe that our proposed methods select all the pollutant variables except NOx and NO and all the meteorological variables except WS. From the results in Table 21, we can clearly infer that the R. Multi-Split-AIC and R. Multi-Split-BIC methods are more efficient than the classical methods, because the final model that is selected by these methods is sufficient to include all the non-zero covariates and has the smallest RMSE. The results of the model validation suggest that WD, Temp, Hum, SO2, NO2, O3 and CO should be included in the final model.

Table 21 Selected variables, average RMSE, and percentage chance for each variable being selected, for air quality data (Threshold=67.08)

	Single-split-AIC	Single-split-BIC	Multi-split-AIC	Multi-split-BIC	R. Multi-split-AIC	R. Multi-split-BIC
RMSE	0.51	0.51	0.51	0.51	0.4	0.4
WS	8.53	0.4	24.24	0.64	23.07	6.6
WD	100	76.86	73.34	14.24	100	99.77
Temp	100	100	100	100	100	100
Hum	100	100	100	100	100	100
NOx	91	45.93	79.30	61.22	54.43	26.3
NO	96.5	47.8	87.66	67.54	49.6	24.63
SO2	89.23	13.63	7.12	0.06	99.93	91.33
NO2	10.46	54.36	32.42	47.06	71	84.77
O3	100	100	100	100	100	100
CO	100	100	100	100	100	100
Selected Variable	2,3,4,5,6,7,9,10	2,3,4,9,10	2,3,4,5,6,9,10	3,4,5,8,9	2,3,4,7,8,9,10	2,3,4,7,8,9,10

The real air quality data and simulation experiments show that our proposed methods successfully and consistently select the correct variables in the final model with the smallest RMSE. The commonly used methods failed to correctly select the correct variables in the final model. Hence, we can consider our proposed methods as a better variable selection method and strongly recommend using them especially when outliers and autocorrelated errors occur in the data.

CONCLUSION

No statistical technique can be used to eliminate or explain all of the uncertainty in the world. Nonetheless, statistics can be used to quantify that uncertainty. That is the reason why statistical techniques have been used widely to help policy makers make decisions. One cannot just use statistical techniques blindly without prior knowledge or sound knowledge in statistics. We have illustrated some topics in statistical analysis where researchers often are not aware of the bad consequences of using classical methods when outliers are present in a data set. To get a valid inference, appropriate statistical techniques should be used and a proper adequacy checking of the underlying assumptions are to be performed. When the basic assumptions are not satisfied, proper remedial measures should be taken into considerations. The classical methods heavily depend on assumptions. The most important assumption in classical method is that data are normally distributed. All classical procedures are based on this assumption. It is very unfortunate that the presence of outliers in a data set may caused apparent non-normality and all the classical procedures breakdown in their presence. Thus, in the presence of outliers, we recommend robust methods to assist statistics practitioners making correct decision. By ignoring the correct statistical techniques and adequacy checking will lead to invalid inferences and inaccurate predictions. Consequently, policy makers become ignorant of the

fact and they are bound to rely on meaningless and misleading results to make decisions and that may bring disaster to a community or to a country.

BIBLIOGRAPHY

Bagheri, A., & Midi, H. (2011). On the performance of robust variance inflation factors. *International Journal of Agricultural and Statistical Sciences*, *7*(1), 31-45.

Bagheri, A., & Habshah, M. (2012a). On the performance of the measure for diagnosing multiple high leverage collinearity-reducingobservations. *Mathematical Problems in Engineering*. vol. 2012, Article ID 531607, 16 pages, 2012. doi:10.1155/2012/531607

Bagheri, A., Habshah, M., & Imon, R. H. M. R. (2012b). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Communications in Statistics-Simulation and Computation*, 41(8), 1379-1396.

Bagheri, A., & Habshah, M. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. *Statistics and Operations Research Transactions, 39*(1), 51-70.

Baltagi, B.H. (2005). *The econometrics of panel data*. New York: John Wiley & Sons.

Batah, F. S. M., Ramanathan, T. V., & Gore, S. D. (2008). The efficiency of modified jackknife and ridge type regression estimators. *A Comparison*. *Surveys in Mathematics and its Applications*, *24*(2), 157-174.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley & Sons.

Belsley, D.A., Kuh, E., & Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York :Wiley.

Bramati, M.C., & Croux, C. (2007). Robust estimators for the fixed effects panel data model. *Econometrics Journal*, *10*(3), 521–540.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic model. *Australian Economic Papers*, *17*, 334-355.

Carroll, R.J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman and Hall.

Chatterjee, S., & Hadi, A.S. (2006). *Regression analysis by example* (4th ed.). New York: Wiley.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis, 45*, 215-233.

Davidson, R., & MacKinnon, J.G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Fox, J. (1972). Outliers in time series. *Journal of the Royal Statistical Society* (B) *34*, 350-363.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

Hocking, R. R., & Pendleton, O. J. (1983). *The regression dilemma*. *Communications in Statistics-Theory and Methods*, *12*(5), 497-527.

Gray, J.B. (1985). Graphics for Regression Diagnostics. *In Proceedings of the Statistical Computing Section, pp. 102-107, American Statistical Association (ASA), Washington, DC, USA*, 1985.

Gel, Y. R., & Gastwirth, J. L. (2008). A robust modification of The Jarque-Bera Test of normality. *Economics Letters*. *99*(1), 30-32.

Godfrey. L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica,, 46*, 1293-1301.

Goethals, P. L., & Cho, B. R. (2011). Solving the optimal process target problem using response surface designs in heteroscedastic conditions. *International Journal of Production Research*, *49*(12), 3455-3478.

Gujarati, D.N., & Porter, D. C. (2009). *Econometrics*. New York: McGraw-Hill, Ch 12 pp. 412-466.

Habshah, M., (1999*)*. Preliminary estimators for robust non-linear regression estimation. *Journal of Applied Statistics*, *26*(5), 591-600.

Habshah, M. (2000). Heteroscedastic nonlinear regression by using tanh phi function. *Sains Malaysiana*, *29*,103-118.

Habshah, M., & Zahari, M. (2007). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Jurnal Teknologi*, *47*(1), 59-74.

Habshah, M., Lim, H. A., & Rana, M. S. (2013). On the robust parameter estimation for linear model with autocorrelated errors. *Advanced Science Letter, 19*(8), 2494-2496

Habshah, M., Norazan, M.R., & 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*(5), 507-520.

Habshah, M., Rana, S., & Imon, A.H.M.R. (2009a). The performance of robust weighted least squares in the presence of outliers and heteroscedastic. *WSEAS Transition of Mathematics*, *8*, 351 – 361.

Habshah, M., Rana, S. & Imon, A.H.M.R. (2009b). Estimation of parameters in heteroscedastic multiple regression model using leverage based near-neighbors. *Journal of Applied Sciences*, *9*, 4013-4019.

Habshah, M, Rana, L. S., & Imon, A.H.M.R. (2014). Two steps robust estimator in heteroscedastic regression model in the presence of outliers. *Economic Computation & Economic Cybernetics Studies & Research, 48*(3).255-272.

Habshah M., Mohammed, A. M., Imon A. H. M. R., & Sohel R. (2015). The modified GM-estimator based on MGDFF for data having multicollinearity due to high leverage points, *Under review, submitted for Journal Applied Statistics.*

Hadi, A. S. (1988). Diagnosing collinearity-influential observations. *Computational Statistics & Data Analysis*, *7*(2), 143-159.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*, 383-393.

Hassan, U., Habshah, M., & Rana, S. (2015). Robust stability best subset selection for autocorrelated data based on robust location and dispersion estimator*, Journal of Probability and Statistics*, Vol 2015, 8 pages.

Hawkins, D. M., Bradu, D., & Kass, G. V. (1984*)*. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, *26*(3), 197-208

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, *12*(1), 55-67.

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

Huber, P.J. (1981). *Robust statistics*. New York: Wiley.

Imon, A. H. M. R. (2003). Regression residuals, moments and their use in tests for normality. *Communications in Statistics - Theory and Methods*, *32*(5), 1021-1034.

Jadhav, N. H., & Kashid, D. N. (2011). A jackknifed ridge m-estimator for regression model with multicollinearity and outliers. *Journal of statistical theory and practice*, *5*(4), 659-673.

Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2005). *Applied linear regression models* (5th ed.).. New York: MacGRAW-Hill.

Lawrence, K. D., & Arthur, J. L. (Eds.). (1990). *robust regression: analysis and applications*. New York: Marcel Dekker.

Lim, H. A., & Midi, H. (2012). Robust Autocorrelation testing in multiple linear regression model with autocorrelated errors. *International Journal of Mathematics and Computers in Simulation*, *6*(1), 119-126.

Lim, H. A., & Midi, H. (2014). The performance of robust modification of Breusch-Godfrey Test in the presence of outliers. *Pertanika Journal of Science & Technology*, *22*(1), 81-94.

Lim, H.A. (2014). Robust Estimation Technique and Robust Autocorrelation Diagnostic for Multiple Linear Regression Model with Autocorrelated Errors. PhD Thesis, UPM

Lin, J.G., & Wei, B.C. (2004). *Testing for heteroscedasticity and correlation in nonlinear regression with correlated errors*. *Communications in Statistics-Theory and Methods*, *33*, 251–275.

MacKinnon, J.G., & White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 53-57.

Maronna, R.A., Martin, R.D., & Yohai, V.J. (2006). *Robust statistics: Theory and methods*. New York: John Wiley.

Midi, H., & Bakar, N. M. A. (2015). The performance of robust-diagnostic F in the identification of multiple high leverage points. *Pak. J. Statist*, *31*(5), 461-472.

Montgomery, D.C., Peck, E.A., & Vining, G.G. (2001). *Introduction to linear regression analysis*. New York: Wiley.

Mohammed A., Habshah, M., and Imon, A. H. M. R. (2015a). A New Robust Diagnostics Plot for Classification Good and Bad High Leverage Points in A Multiple Linear Regression Model. *Mathematical Problems in Engineering* vol. 2015, Article ID 279472, 12 pages, 2015. doi:10.1155/2015/279472.

Mohammed, A. Habshah, M, & Rana, L. S. (2015b). Robust jackknife ridge regression to combat multicollinearity and high leverage points in multiple linear regressions. *Economic Computation & Economic Cybernetics Studies & Research, 49*(4).

Mohd Shafie Mustafa (2015). Robust Outlier Detection and Estimation in Response Surface Methodology. PhD Thesis UPM.

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th Ed.).. John Wiley and Sons, Inc.

Montgomery, D. C., Peck, E. A., & Vining, G. G (2001). *Introduction to linear regression analysis* (3rd Ed.). John Wiley and Sons, Inc.

Mustafa, M.N.(2005). Overview of current road safety situation in Malaysia. *Highway planning Unit, Road safety Section, Ministry of Works*, 5-9.

Park, C., & Cho, B. R. (2003). Development of robust design under contaminated and non-normal data. *Quality Engineering*, *15*, *3*, 463-469.

Rana, S., Habshah, M., & Imon, A.H.M.R. (2012). *Robust Wild Bootstrap for Stabilizing The Variance of Parameter Estimates in Heteroscedastic Regression Models in The Presence of Outliers*. Mathematical Problems in Engineering, Article ID 730328, 2012, 14 pages.

Rana, S., Habshah, M., & Imon, A.H.M.R. (2008). A robust modification of the Goldfeld-Quandt Test for the detection of heteroscedasticity in the presence of outliers. *Journal of Mathematics and Statistic*s, *4*, 277-283.

Rana, S., Habshah, M., & Imon, A.H.M.R. (2009). A Robust Rescaled Moment Test for Normality in Regression. *Journal of Mathematics and Statistics, 5*(1), 54-62.

Ratkowsky, D. A., (1983). *Nonlinear regression modeling*. New York: Marcel Dekker.

Riazoshams, H., Midi, H. B., & Sharipov, O. S. (2010). The performance of robust two-stage estimator in nonlinear regression with autocorrelated error. *Communications in Statistics-Simulation and Computation*, *39*(6), 1251-1268.

Rousseeuw, P. J., & Leroy, A. (1987), Robust Regression and Outlier Detection, New York: Wiley.

Rousseeuw, P., & Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Associations*. *85*, 633-639.

Sengupta, D., & Bhimasankaram, P. (1997).On the roles of observations in collinearity in the Linear Model. *Journal of American Statistical Association*. *92*, 1024-1032.

Seber, G., A. F. & Wild, C. J. (2003). *Nonlinear regression*. New York: John Wiley & Sons, Inc.

Siraj-ud-doulah, M., Rana, S., Midi, H., & Imon, A.H.M.R (2012). New Robust Tests for Detection of ARCH Effect. *Economic Computation and Economic Cybernetics Studies and Research*, *46*, 251- 259.

Tukey, J.W. (1977). *Exploratory data analysis*. Cyprus: Addison-Wesley Publishers.

Vellman, P.F., & Welsch, R.E. (1981). Efficient computing of regression diagnostics. *American* S*tatistician. 27*, 234-242.

Vining, G. G., & Myers, R. H. (1990). *Combining Taguchi and response surface philosophies: A dual response approach*. *Journal of Quality Technology*, *22*, 34-45.

White, G. C., & Brisbon, I. L. (1980). Estimation and comparison of parameters in stochastic growth model for barn owls. *Growth*. *44*, 97-111.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica*, *48*, 817-838.

BIOGRAPHY

Habshah Midi completed her secondary education at Sekolah Menengah Sains Selangor in 1974 and continued with her undergraduate studies at the *Drew University, USA*, graduating with *Bachelors Degree in Mathematics* in 1979. In pursuance for academic excellence, Habshah Midi continued her education at *The Ohio State University, USA,* where she received the *Master of Applied Statistics* in 1981. In 1981, she joined Universiti Putra Malaysia as a lecturer and pursued her Ph.D (Statisctis) at Universiti Kebangsaan Malaysia where she was awarded her doctorate in 1995. Habshah supervises and co-supervises Ph.D and Master students with more than 20 graduated from Universiti Putra Malaysia. A Ph.D student was co-supervised with Dato' Ir. Dr Radin Umar Radin Sohadi, the ex-Vice Chancellor of UPM, who graduated in 2006. To date, Habshah is supervising eleven Ph.D students and four Master students. In 2013, she received certificate of appreciation from UPM to be recognised as *5 Star and Role Model Supervisor*. Additionally, Habshah Midi sits in as an internal and external examiner for post-graduate students.

Having published more than 100 papers in international and local citation-indexed journals, her research interests focus on *robust statistics, regressions, experimental designs, quality control, sampling techniques, bootstrapping techniques* and *application of statistical methods to real life problems*. Habshah Midi is concerned with the usage of the right statistical techniques in research. Due to that fact, she has organised many workshops and attended many as the invited speaker that addresses the issue, like *The Right Way to Analyse Statistical Data* workshops and spoke on "The Use and Abuse of Statistics" at the *National Statistics Conference, Putrajaya International Convention Centre,* organised by Statistics Department, Malaysia in 2006. She is also invited to Majlis

Peperiksaan Malaysia to conduct a workshop on the "*The Right Way to Conduct Sampling Techniques in Research.*"

Her progressive nature and interest in research have garnered more than a million Ringgit in research grants over the years. She is actively involved in the Malaysia Institute of Statistics Society and has been appointed as the Vice President since 2014. Prior to this, she was the council members of the society. She is often invited as invited speaker in statistical conferences and also has served as research grant evaluator, journal papers reviewer and external reviewer for academic promotions. Habshah's active involvement in research and exhibitions was further recognised when she was awarded several silver medals and several gold medals in research exhibitions. Her research entitled, *The Misconception of Some Statistical Techniques in Research* won gold medal in the 2009 *PRPI* exhibition. Habshah also serves as a statistics consultant to several organisations including Majlis Peperiksaan Malaysia from 2007-2010, whereby the ex-Vice Chancellor of UPM, Tan Sri Dr. Nik Mustapha bin R. Abdullah cum the chairman of the *Performance of the MUET study committee* appointed her as the head of the researcher. She has been invited four times by Majlis Peperiksaan Malaysia to present the progress of this study at the *MPM* meetings. She has successfully completed a written report in 2009 on "Laporan Kajian Pencapaian Malaysia University English Test (MUET) 2002-2006". The then Deputy Prime Minister cum the Minister of Education, Malaysia, approved the report for public viewing. Habshah Midi is also involved as questions setter, one of the panel questions evaluator, and group examiners' leader for the matriculation program's mathematics paper. Additionally, she was appointed by the Ministry of Education as an *Ahli Panel Temuduga dan Penggubal Soalan Tugasan Program Penilaian Guru Cemerlang Gred Khas C* (KUP: Matematik). She also actively

involved in the community and industrial linkages and has been the programme head (2012, 2013, 2015) for the Nobel Laureates Outreach Camp - 'Scientist of Tomorrow' for selected MRSM Students from 45 MRSM in Malaysia. She was also the programme head and on of the speaker for Nobel Laureates Outreach Camp 2015 - Training of Trainers Camp for future Nobel Laureates.

Habshah Midi was also the head of the Curriculum Committee for the Department of Mathematics, became a member of Faculty Sciences Curriculum Committee and was the co-ordinator of the *Master of Applied Statistics Program*, from 1998-June 2004. She was the head of the *Laboratory of Applied and Computational Statistics, Institute for Mathematical Research* (INSPEM), Universiti Putra Malaysia, from 2004-May 2010. Currently, she is the Deputy Dean of the Faculty of Science, UPM since 2010.

As an academician, Habshah Midi is dedicated to the quality of teaching, hence is constantly involved in module writings namely the module for *MTH3003 (Statistics for Applied Sciences), MTH3401 (Statistics and Probability I)* and *MTH 3406 (Quality Control Techniques)*. Her persistence in ensuring the success of her teaching delivery is reflected in an above 4.5 (90%) of student's teaching evaluation and have been awarded *"Excellence in Teaching*" from the Faculty of Science, every year since it was introduced in 2002. In 2007, Habshah was awarded the *Fellowship Naib Canselor (Excellent in Teaching)*. In 2010, she was nominated by UPM for the National Academic Award (Excellence in Teaching).

ACKNOWLEDGEMENT

In the name of ALLAH the Most Gracious and the Most Merciful.

I am most grateful to ALLAH SWT for His Blessings and Guidance and making me the recipient of a life generous with love and support. First and foremost I would like to thank Universiti Putra Malaysia, Ministry of Science, Technology and Innovation, and Ministry of Higher Learning for funding my research grant. I also would like to thank UPM for giving me this career opportunity and chance to contribute my knowledge to students and societies through my academic activities over the last 35 years. My sincere appreciation to my PhD supervisor, co-researchers and students whose names are impossible to list at the Department of Mathematics, Faculty of Science and Institute for Mathematical Research, UPM. Not forgetting the relentless support from superiors, colleagues and supporting staff, for which I am most grateful. My utmost appreciation goes to my late mother and father, my brother and sisters and families for their prayers and love. However, my heartfelt thanks and gratitude must go to my husband, Azmi, my children, Nur Liyana, Ahmad Azfar, Nur Izzati, Ahmad Syahmi and Nur Sabrina for their endless love, support, sacrifice and constant prayers. My special thanks to my husband that always give me inspiration, patience and constantly guide me to be a good Muslim. Abang Azmi, "I love you so much." For all who have touched my life so graciously, they have indeed touched my heart, for which I am forever indebted.

LIST OF INAUGURAL LECTURES

1. Prof. Dr. Sulaiman M. Yassin
   The Challenge to Communication Research in Extension
   22 July 1989

2. Prof. Ir. Abang Abdullah Abang Ali
   Indigenous Materials and Technology for Low Cost Housing
   30 August 1990

3. Prof. Dr. Abdul Rahman Abdul Razak
   Plant Parasitic Nematodes, Lesser Known Pests of Agricultural Crops
   30 January 1993

4. Prof. Dr. Mohamed Suleiman
   Numerical Solution of Ordinary Differential Equations: A Historical Perspective
   11 December 1993

5. Prof. Dr. Mohd. Ariff Hussein
   Changing Roles of Agricultural Economics
   5 March 1994

6. Prof. Dr. Mohd. Ismail Ahmad
   Marketing Management: Prospects and Challenges for Agriculture
   6 April 1994

7. Prof. Dr. Mohamed Mahyuddin Mohd. Dahan
   The Changing Demand for Livestock Products
   20 April 1994

8. Prof. Dr. Ruth Kiew
   Plant Taxonomy, Biodiversity and Conservation
   11 May 1994

9. Prof. Ir. Dr. Mohd. Zohadie Bardaie
   Engineering Technological Developments Propelling Agriculture into the 21st Century
   28 May 1994

10. Prof. Dr. Shamsuddin Jusop
    Rock, Mineral and Soil
    18 June 1994

11. Prof. Dr. Abdul Salam Abdullah
    Natural Toxicants Affecting Animal Health and Production
    29 June 1994

12. Prof. Dr. Mohd. Yusof Hussein
    Pest Control: A Challenge in Applied Ecology
    9 July 1994

13. Prof. Dr. Kapt. Mohd. Ibrahim Haji Mohamed
    Managing Challenges in Fisheries Development through Science and Technology
    23 July 1994

14. Prof. Dr. Hj. Amat Juhari Moain
    Sejarah Keagungan Bahasa Melayu
    6 August 1994

15. Prof. Dr. Law Ah Theem
    Oil Pollution in the Malaysian Seas
    24 September 1994

16. Prof. Dr. Md. Nordin Hj. Lajis
    Fine Chemicals from Biological Resources: The Wealth from Nature
    21 January 1995

17. Prof. Dr. Sheikh Omar Abdul Rahman
    Health, Disease and Death in Creatures Great and Small
    25 February 1995

18. Prof. Dr. Mohamed Shariff Mohamed Din
Fish Health: An Odyssey through the Asia - Pacific Region
25 March 1995

19. Prof. Dr. Tengku Azmi Tengku Ibrahim
Chromosome Distribution and Production Performance of Water Buffaloes
6 May 1995

20. Prof. Dr. Abdul Hamid Mahmood
Bahasa Melayu sebagai Bahasa Ilmu-Cabaran dan Harapan
10 June 1995

21. Prof. Dr. Rahim Md. Sail
Extension Education for Industrialising Malaysia: Trends, Priorities and Emerging Issues
22 July 1995

22. Prof. Dr. Nik Muhammad Nik Abd. Majid
The Diminishing Tropical Rain Forest: Causes, Symptoms and Cure
19 August 1995

23. Prof. Dr. Ang Kok Jee
The Evolution of an Environmentally Friendly Hatchery Technology for Udang Galah, the King of Freshwater Prawns and a Glimpse into the Future of Aquaculture in the 21st Century
14 October 1995

24. Prof. Dr. Sharifuddin Haji Abdul Hamid
Management of Highly Weathered Acid Soils for Sustainable Crop Production
28 October 1995

25. Prof. Dr. Yu Swee Yean
Fish Processing and Preservation: Recent Advances and Future Directions
9 December 1995

26. Prof. Dr. Rosli Mohamad
Pesticide Usage: Concern and Options
10 February 1996

27. Prof. Dr. Mohamed Ismail Abdul Karim
Microbial Fermentation and Utilization of Agricultural Bioresources and Wastes in Malaysia
2 March 1996

28. Prof. Dr. Wan Sulaiman Wan Harun
Soil Physics: From Glass Beads to Precision Agriculture
16 March 1996

29. Prof. Dr. Abdul Aziz Abdul Rahman
Sustained Growth and Sustainable Development: Is there a Trade-Off 1 or Malaysia
13 April 1996

30. Prof. Dr. Chew Tek Ann
Sharecropping in Perfectly Competitive Markets: A Contradiction in Terms
27 April 1996

31. Prof. Dr. Mohd. Yusuf Sulaiman
Back to the Future with the Sun
18 May 1996

32. Prof. Dr. Abu Bakar Salleh
Enzyme Technology: The Basis for Biotechnological Development
8 June 1996

33. Prof. Dr. Kamel Ariffin Mohd. Atan
The Fascinating Numbers
29 June 1996

34. Prof. Dr. Ho Yin Wan
Fungi: Friends or Foes
27 July 1996

35. Prof. Dr. Tan Soon Guan
Genetic Diversity of Some Southeast Asian Animals: Of Buffaloes and Goats and Fishes Too
10 August 1996

36. Prof. Dr. Nazaruddin Mohd. Jali
Will Rural Sociology Remain Relevant in the 21st Century?
21 September 1996

37. Prof. Dr. Abdul Rani Bahaman
Leptospirosis-A Model for Epidemiology, Diagnosis and Control of Infectious Diseases
16 November 1996

38. Prof. Dr. Marziah Mahmood
Plant Biotechnology - Strategies for Commercialization
21 December 1996

39. Prof. Dr. Ishak Hj. Omar
Market Relationships in the Malaysian Fish Trade: Theory and Application
22 March 1997

40. Prof. Dr. Suhaila Mohamad
Food and Its Healing Power
12 April 1997

41. Prof. Dr. Malay Raj Mukerjee
A Distributed Collaborative Environment for Distance Learning Applications
17 June 1998

42. Prof. Dr. Wong Kai Choo
Advancing the Fruit Industry in Malaysia: A Need to Shift Research Emphasis
15 May 1999

43. Prof. Dr. Aini Ideris
Avian Respiratory and Immunosuppressive Diseases- A Fatal Attraction
10 July 1999

44. Prof. Dr. Sariah Meon
Biological Control of Plant Pathogens: Harnessing the Richness of Microbial Diversity
14 August 1999

45. Prof. Dr. Azizah Hashim
The Endomycorrhiza: A Futile Investment?
23 October 1999

46. Prof. Dr. Noraini Abdul Samad
Molecular Plant Virology: The Way Forward
2 February 2000

47. Prof. Dr. Muhamad Awang
Do We Have Enough Clean Air to Breathe?
7 April 2000

48. Prof. Dr. Lee Chnoong Kheng
Green Environment, Clean Power
24 June 2000

49. Prof. Dr. Mohd. Ghazali Mohayidin
Managing Change in the Agriculture Sector: The Need for Innovative Educational Initiatives
12 January 2002

50. Prof. Dr. Fatimah Mohd. Arshad
Analisis Pemasaran Pertanian di Malaysia: Keperluan Agenda Pembaharuan
26 January 2002

51. Prof. Dr. Nik Mustapha R. Abdullah
Fisheries Co-Management: An Institutional Innovation Towards Sustainable Fisheries Industry
28 February 2002

52. Prof. Dr. Gulam Rusul Rahmat Ali
Food Safety: Perspectives and Challenges
23 March 2002

53. Prof. Dr. Zaharah A. Rahman
Nutrient Management Strategies for Sustainable Crop Production in Acid Soils: The Role of Research Using Isotopes
13 April 2002

54. Prof. Dr. Maisom Abdullah
Productivity Driven Growth: Problems & Possibilities
27 April 2002

55. Prof. Dr. Wan Omar Abdullah
Immunodiagnosis and Vaccination for Brugian Filariasis: Direct Rewards from Research Investments
6 June 2002

56. Prof. Dr. Syed Tajuddin Syed Hassan
Agro-ento Bioinformation: Towards the Edge of Reality
22 June 2002

57. Prof. Dr. Dahlan Ismail
Sustainability of Tropical Animal-Agricultural Production Systems: Integration of Dynamic Complex Systems
27 June 2002

58. Prof. Dr. Ahmad Zubaidi Baharumshah
The Economics of Exchange Rates in the East Asian Countries
26 October 2002

59. Prof. Dr. Shaik Md. Noor Alam S.M. Hussain
Contractual Justice in Asean: A Comparative View of Coercion
31 October 2002

60. Prof. Dr. Wan Md. Zin Wan Yunus
Chemical Modification of Polymers: Current and Future Routes for Synthesizing New Polymeric Compounds
9 November 2002

61. Prof. Dr. Annuar Md. Nassir
Is the KLSE Efficient? Efficient Market Hypothesis vs Behavioural Finance
23 November 2002

62. Prof. Ir. Dr. Radin Umar Radin Sohadi
Road Safety Interventions in Malaysia: How Effective Are They?
21 February 2003

63. Prof. Dr. Shamsher Mohamad
The New Shares Market: Regulatory Intervention, Forecast Errors and Challenges
26 April 2003

64. Prof. Dr. Han Chun Kwong
Blueprint for Transformation or Business as Usual? A Structurational Perspective of the Knowledge-Based Economy in Malaysia
31 May 2003

65. Prof. Dr. Mawardi Rahmani
Chemical Diversity of Malaysian Flora: Potential Source of Rich Therapeutic Chemicals
26 July 2003

66. Prof. Dr. Fatimah Md. Yusoff
An Ecological Approach: A Viable Option for Aquaculture Industry in Malaysia
9 August 2003

67. Prof. Dr. Mohamed Ali Rajion
The Essential Fatty Acids-Revisited
23 August 2003

68. Prof. Dr. Azhar Md. Zain
Psychotheraphy for Rural Malays - Does it Work?
13 September 2003

69. Prof. Dr. Mohd. Zamri Saad
Respiratory Tract Infection: Establishment and Control
27 September 2003

70. Prof. Dr. Jinap Selamat
Cocoa-Wonders for Chocolate Lovers
14 February 2004

71. Prof. Dr. Abdul Halim Shaari
High Temperature Superconductivity: Puzzle & Promises
13 March 2004

72. Prof. Dr. Yaakob Che Man
Oils and Fats Analysis - Recent Advances and Future Prospects
27 March 2004

73. Prof. Dr. Kaida Khalid
Microwave Aquametry: A Growing Technology
24 April 2004

74. Prof. Dr. Hasanah Mohd. Ghazali
Tapping the Power of Enzymes-Greening the Food Industry
11 May 2004

75. Prof. Dr. Yusof Ibrahim
The Spider Mite Saga: Quest for Biorational Management Strategies
22 May 2004

76. Prof. Datin Dr. Sharifah Md. Nor
The Education of At-Risk Children: The Challenges Ahead
26 June 2004

77. Prof. Dr. Ir. Wan Ishak Wan Ismail
Agricultural Robot: A New Technology Development for Agro-Based Industry
14 August 2004

78. Prof. Dr. Ahmad Said Sajap
Insect Diseases: Resources for Biopesticide Development
28 August 2004

79. Prof. Dr. Aminah Ahmad
The Interface of Work and Family Roles: A Quest for Balanced Lives
11 March 2005

80. Prof. Dr. Abdul Razak Alimon
Challenges in Feeding Livestock: From Wastes to Feed
23 April 2005

81. Prof. Dr. Haji Azimi Hj. Hamzah
Helping Malaysian Youth Move Forward: Unleashing the Prime Enablers
29 April 2005

82. Prof. Dr. Rasedee Abdullah
In Search of An Early Indicator of Kidney Disease
27 May 2005

83. Prof. Dr. Zulkifli Hj. Shamsuddin
Smart Partnership: Plant-Rhizobacteria Associations
17 June 2005

84. Prof. Dr. Mohd Khanif Yusop
From the Soil to the Table
1 July 2005

85. Prof. Dr. Annuar Kassim
Materials Science and Technology: Past, Present and the Future
8 July 2005

86. Prof. Dr. Othman Mohamed
Enhancing Career Development Counselling and the Beauty of Career Games
12 August 2005

87. Prof. Ir. Dr. Mohd Amin Mohd Soom
Engineering Agricultural Water Management Towards Precision Framing
26 August 2005

88. Prof. Dr. Mohd Arif Syed
Bioremediation-A Hope Yet for the Environment?
9 September 2005

89. Prof. Dr. Abdul Hamid Abdul Rashid
The Wonder of Our Neuromotor System and the Technological Challenges They Pose
23 December 2005

90. Prof. Dr. Norhani Abdullah
Rumen Microbes and Some of Their Biotechnological Applications
27 January 2006

91. Prof. Dr. Abdul Aziz Saharee
Haemorrhagic Septicaemia in Cattle and Buffaloes: Are We Ready for Freedom?
24 February 2006

92. Prof. Dr. Kamariah Abu Bakar
Activating Teachers' Knowledge and Lifelong Journey in Their Professional Development
3 March 2006

93. Prof. Dr. Borhanuddin Mohd. Ali
Internet Unwired
24 March 2006

94. Prof. Dr. Sundararajan Thilagar
Development and Innovation in the Fracture Management of Animals
31 March 2006

95. Prof. Dr. Zainal Aznam Md. Jelan
Strategic Feeding for a Sustainable Ruminant Farming
19 May 2006

96. Prof. Dr. Mahiran Basri
Green Organic Chemistry: Enzyme at Work
14 July 2006

97. Prof. Dr. Malik Hj. Abu Hassan
Towards Large Scale Unconstrained Optimization
20 April 2007

98. Prof. Dr. Khalid Abdul Rahim
Trade and Sustainable Development: Lessons from Malaysia's Experience
22 June 2007

99. Prof. Dr. Mad Nasir Shamsudin
Econometric Modelling for Agricultural Policy Analysis and Forecasting: Between Theory and Reality
13 July 2007

100. Prof. Dr. Zainal Abidin Mohamed
Managing Change - The Fads and The Realities: A Look at Process Reengineering, Knowledge Management and Blue Ocean Strategy
9 November 2007

101. Prof. Ir. Dr. Mohamed Daud
Expert Systems for Environmental Impacts and Ecotourism Assessments
23 November 2007

102. Prof. Dr. Saleha Abdul Aziz
Pathogens and Residues; How Safe is Our Meat?
30 November 2007

103. Prof. Dr. Jayum A. Jawan
Hubungan Sesama Manusia
7 December 2007

104. Prof. Dr. Zakariah Abdul Rashid
Planning for Equal Income Distribution in Malaysia: A General Equilibrium Approach
28 December 2007

105. Prof. Datin Paduka Dr. Khatijah Yusoff
Newcastle Disease virus: A Journey from Poultry to Cancer
11 January 2008

106. Prof. Dr. Dzulkefly Kuang Abdullah
Palm Oil: Still the Best Choice
1 February 2008

107. Prof. Dr. Elias Saion
Probing the Microscopic Worlds by Lonizing Radiation
22 February 2008

108. Prof. Dr. Mohd Ali Hassan
Waste-to-Wealth Through Biotechnology: For Profit, People and Planet
28 March 2008

109. Prof. Dr. Mohd Maarof H. A. Moksin
Metrology at Nanoscale: Thermal Wave Probe Made It Simple
11 April 2008

110. Prof. Dr. Dzolkhifli Omar
The Future of Pesticides Technology in Agriculture: Maximum Target Kill with Minimum Collateral Damage
25 April 2008

111. Prof. Dr. Mohd. Yazid Abd. Manap
Probiotics: Your Friendly Gut Bacteria
9 May 2008

112. Prof. Dr. Hamami Sahri
Sustainable Supply of Wood and Fibre: Does Malaysia have Enough?
23 May 2008

113. Prof. Dato' Dr. Makhdzir Mardan
Connecting the Bee Dots
20 June 2008

114. Prof. Dr. Maimunah Ismail
Gender & Career: Realities and Challenges
25 July 2008

115. Prof. Dr. Nor Aripin Shamaan
Biochemistry of Xenobiotics: Towards a Healthy Lifestyle and Safe Environment
1 August 2008

116. Prof. Dr. Mohd Yunus Abdullah
Penjagaan Kesihatan Primer di Malaysia: Cabaran Prospek dan Implikasi dalam Latihan dan Penyelidikan Perubatan serta Sains Kesihatan di Universiti Putra Malaysia
8 August 2008

117. Prof. Dr. Musa Abu Hassan
Memanfaatkan Teknologi Maklumat & Komunikasi ICT untuk Semua
15 August 2008

118. Prof. Dr. Md. Salleh Hj. Hassan
Role of Media in Development: Strategies, Issues & Challenges
22 August 2008

119. Prof. Dr. Jariah Masud
Gender in Everyday Life
10 October 2008

120 Prof. Dr. Mohd Shahwahid Haji Othman
Mainstreaming Environment: Incorporating Economic Valuation and Market-Based Instruments in Decision Making
24 October 2008

121. Prof. Dr. Son Radu
Big Questions Small Worlds: Following Diverse Vistas
31 October 2008

122. Prof. Dr. Russly Abdul Rahman
Responding to Changing Lifestyles: Engineering the Convenience Foods
28 November 2008

123. Prof. Dr. Mustafa Kamal Mohd Shariff
Aesthetics in the Environment an Exploration of Environmental: Perception Through Landscape Preference
9 January 2009

124. Prof. Dr. Abu Daud Silong
Leadership Theories, Research & Practices: Farming Future Leadership Thinking
16 January 2009

125. Prof. Dr. Azni Idris
Waste Management, What is the Choice: Land Disposal or Biofuel?
23 January 2009

126. Prof. Dr. Jamilah Bakar
Freshwater Fish: The Overlooked Alternative
30 January 2009

127. Prof. Dr. Mohd. Zobir Hussein
The Chemistry of Nanomaterial and Nanobiomaterial
6 February 2009

128. Prof. Ir. Dr. Lee Teang Shui
Engineering Agricultural: Water Resources
20 February 2009

129. Prof. Dr. Ghizan Saleh
Crop Breeding: Exploiting Genes for Food and Feed
6 March 2009

130. Prof. Dr. Muzafar Shah Habibullah
Money Demand
27 March 2009

131. Prof. Dr. Karen Anne Crouse
In Search of Small Active Molecules
3 April 2009

132. Prof. Dr. Turiman Suandi
Volunteerism: Expanding the Frontiers of Youth Development
17 April 2009

133. Prof. Dr. Arbakariya Ariff
Industrializing Biotechnology: Roles of Fermentation and Bioprocess Technology
8 May 2009

134. Prof. Ir. Dr. Desa Ahmad
Mechanics of Tillage Implements
12 June 2009

135. Prof. Dr. W. Mahmood Mat Yunus
Photothermal and Photoacoustic: From Basic Research to Industrial Applications
10 July 2009

136. Prof. Dr. Taufiq Yap Yun Hin
Catalysis for a Sustainable World
7 August 2009

137 Prof. Dr. Raja Noor Zaliha Raja Abd. Rahman
Microbial Enzymes: From Earth to Space
9 October 2009

138 Prof. Ir. Dr. Barkawi Sahari
Materials, Energy and CNGDI Vehicle Engineering
6 November 2009

139. Prof. Dr. Zulkifli Idrus
Poultry Welfare in Modern Agriculture: Opportunity or Threat?
13 November 2009

140. Prof. Dr. Mohamed Hanafi Musa
Managing Phosphorus: Under Acid Soils Environment
8 January 2010

141. Prof. Dr. Abdul Manan Mat Jais
Haruan Channa striatus a Drug Discovery in an Agro-Industry Setting
12 March 2010

142. Prof. Dr. Bujang bin Kim Huat
Problematic Soils: In Search for Solution
19 March 2010

143. Prof. Dr. Samsinar Md Sidin
Family Purchase Decision Making: Current Issues & Future Challenges
16 April 2010

144. Prof. Dr. Mohd Adzir Mahdi
Lightspeed: Catch Me If You Can
4 June 2010

145. Prof. Dr. Raha Hj. Abdul Rahim
Designer Genes: Fashioning Mission Purposed Microbes
18 June 2010

146. Prof. Dr. Hj. Hamidon Hj. Basri
A Stroke of Hope, A New Beginning
2 July 2010

147. Prof. Dr. Hj. Kamaruzaman Jusoff
Going Hyperspectral: The "Unseen" Captured?
16 July 2010

148. Prof. Dr. Mohd Sapuan Salit
Concurrent Engineering for Composites
30 July 2010

149. Prof. Dr. Shattri Mansor
Google the Earth: What's Next?
15 October 2010

150. Prof. Dr. Mohd Basyaruddin Abdul Rahman
Haute Couture: Molecules & Biocatalysts
29 October 2010

151. Prof. Dr. Mohd. Hair Bejo
Poultry Vaccines: An Innovation for Food Safety and Security
12 November 2010

152. Prof. Dr. Umi Kalsom Yusuf
Fern of Malaysian Rain Forest
3 December 2010

153. Prof. Dr. Ab. Rahim Bakar
Preparing Malaysian Youths for The World of Work: Roles of Technical and Vocational Education and Training (TVET)
14 January 2011

154. Prof. Dr. Seow Heng Fong
Are there "Magic Bullets" for Cancer Therapy?
11 February 2011

155. Prof. Dr. Mohd Azmi Mohd Lila
Biopharmaceuticals: Protection, Cure and the Real Winner
18 February 2011

156. Prof. Dr. Siti Shapor Siraj
Genetic Manipulation in Farmed Fish: Enhancing Aquaculture Production
25 March 2011

157. Prof. Dr. Ahmad Ismail
Coastal Biodiversity and Pollution: A Continuous Conflict
22 April 2011

158. Prof. Ir. Dr. Norman Mariun
Energy Crisis 2050? Global Scenario and Way Forward for Malaysia
10 June 2011

159. Prof. Dr. Mohd Razi Ismail
Managing Plant Under Stress: A Challenge for Food Security
15 July 2011

160. Prof. Dr. Patimah Ismail
Does Genetic Polymorphisms Affect Health?
23 September 2011

161. Prof. Dr. Sidek Ab. Aziz
Wonders of Glass: Synthesis, Elasticity and Application
7 October 2011

162. Prof. Dr. Azizah Osman
Fruits: Nutritious, Colourful, Yet Fragile Gifts of Nature
14 October 2011

163. Prof. Dr. Mohd. Fauzi Ramlan
Climate Change: Crop Performance and Potential
11 November 2011

164. Prof. Dr. Adem Kiliçman
Mathematical Modeling with Generalized Function
25 November 2011

165. Prof. Dr. Fauziah Othman
My Small World: In Biomedical Research
23 December 2011

166. Prof. Dr. Japar Sidik Bujang
The Marine Angiosperms, Seagrass
23 March 2012

167. Prof. Dr. Zailina Hashim
Air Quality and Children's Environmental Health: Is Our Future Generation at Risk?
30 March 2012

168. Prof. Dr. Zainal Abidin Mohamed
Where is the Beef? Vantage Point form the Livestock Supply Chain
27 April 2012

169. Prof. Dr. Jothi Malar Panandam
Genetic Characterisation of Animal Genetic Resources for Sustaninable Utilisation and Development
30 November 2012

170. Prof. Dr. Fatimah Abu Bakar
The Good The Bad & Ugly of Food Safety: From Molecules to Microbes
7 December 2012

171. Prof. Dr. Abdul Jalil Nordin
My Colourful Sketches from Scratch: Molecular Imaging
5 April 2013

172. Prof. Dr. Norlijah Othman
Lower Respiratory Infections in Children: New Pathogens, Old Pathogens and The Way Forward
19 April 2013

173. Prof. Dr. Jayakaran Mukundan
Steroid-like Prescriptions English Language Teaching Can Ill-afford
26 April 2013

174. Prof. Dr. Azmi Zakaria
Photothermals Affect Our Lives
7 June 2013

175. Prof. Dr. Rahinah Ibrahim
Design Informatics
21 June 2013

176. Prof. Dr. Gwendoline Ee Cheng
Natural Products from Malaysian Rainforests
1 November 2013

177. Prof. Dr. Noor Akma Ibrahim
The Many Facets of Statistical Modeling
22 November 2013

178. Prof. Dr. Paridah Md. Tahir
Bonding with Natural Fibres
6 December 2013

179. Prof. Dr. Abd. Wahid Haron
Livestock Breeding: The Past, The Present and The Future
9 December 2013

180. Prof. Dr. Aziz Arshad
Exploring Biodiversity & Fisheries Biology: A Fundamental Knowledge for Sustainabale Fish Production
24 January 2014

181. Prof. Dr. Mohd Mansor Ismail
Competitiveness of Beekeeping Industry in Malaysia
21 March 2014

182. Prof. Dato' Dr. Tai Shzee Yew
Food and Wealth from the Seas: Health Check for the Marine Fisheries of Malaysia
25 April 2014

183. Prof. Datin Dr. Rosenani Abu Bakar
Waste to Health: Organic Waste Management for Sustainable Soil Management and Crop Production
9 May 2014

184. Prof. Dr. Abdul Rahman Omar
Poultry Viruses: From Threat to Therapy
23 May 2014

185. Prof. Dr. Mohamad Pauzi Zakaria
Tracing the Untraceable: Fingerprinting Pollutants through Environmental Forensics
13 June 2014

186. Prof. Dr. -Ing. Ir. Renuganth Varatharajoo
Space System Trade-offs: Towards Spacecraft Synergisms
15 August 2014

187. Prof. Dr. Latiffah A. Latiff
Tranformasi Kesihatan Wanita ke Arah Kesejahteraan Komuniti
7 November 2014

188. Prof. Dr. Tan Chin Ping
Fat and Oils for a Healthier Future: Macro, Micro and Nanoscales
21 November 2014

189. Prof. Dr. Suraini Abd. Aziz
Lignocellulosic Biofuel: A Way Forward
28 November 2014

190. Prof. Dr. Robiah Yunus
Biobased Lubricants: Harnessing the Richness of Agriculture Resources
30 January 2015

190. Prof. Dr. Khozirah Shaari
Discovering Future Cures from Phytochemistry to Metabolomics
13 February 2015

191. Prof. Dr. Tengku Aizan Tengku Abdul Hamid
Population Ageing in Malaysia: A Mosaic of Issues, Challenges and Prospects
13 March 2015

192. Prof. Datin Dr. Faridah Hanum Ibrahim
Forest Biodiversity: Importance of Species Composition Studies
27 March 2015

192. Prof. Dr. Mohd Salleh Kamarudin
Feeding & Nutritional Requirements of Young Fish
10 April 2015

193. Prof. Dato' Dr. Mohammad Shatar Sabran
Money Boy: Masalah Sosial Era Generasi Y
8 Mei 2015

194. Prof. Dr. Aida Suraya Md. Yunus
Developing Students' Mathematical Thinking: How Far Have We Come?
5 June 2015

195. Prof. Dr. Amin Ismail
Malaysian Cocoa or Chocolates: A Story of Antioxidants and More...
14 August 2015

196. Prof. Dr. Shamsuddin Sulaiman
Casting Technology: Sustainable Metal Forming Process
21 August 2015

197. Prof. Dr. Rozita Rosli
Journey into Genetic: Taking the Twist and Turns of Life
23 October 2015

198. Prof. Dr. Nor Aini Ab Shukor
The (Un)Straight Truth About Trees
6 November 2015

198. Prof. Dr. Maznah Ismail
Germinated Brown Rice and Bioactive Rich Fractions: On Going Journey form R&D to Commercialisation
29 April 2016