'EMERGING THEMES IN FUNDAMENTAL AND APPLIED SCIENCES'

MATHEMATICS

VOLUME 2

EDITOR

MOHAMAT AIDIL BIN

MOHAMAT JOHARI

Universiti Putra Malaysia Press

Serdang • 2019

First Print 2019

UPM Press is a member of the Malaysian Book Publishers Association (MABOPA) Membership No.: 9802 and a member of Majlis Penerbitan Ilmiah Malaysia (MAPIM)

Perpustakaan Negara Malaysia Cataloguing-in-Publication Data

EMERGING THEMES IN FUNDAMENTAL AND APPLIED SCIENCES': MATHEMATICS. VOLUME 2 / EDITOR: MOHAMAT AIDIL BIN MOHAMAT JOHARI.
Mode of access: Internet.
eISBN 978-967-344-903-3
1. Mathematics.
2. Statistics.
3. Government publications--Malaysia.
4. Electronic books.
I. Title
510

URL: http://science.upm.edu.my/ebook-3213

E-mail: penerbit@upm.edu.my
Web: www.penerbit.upm.edu.my

EMERGING THEMES IN FUNDAMENTAL AND APPLIED SCIENCES

Edited by

Mohamat Aidil bin Mohamat Johari

Faculty of Science
Universiti Putra Malaysia
Serdang
Selangor
Malaysia

UPM Press
Universiti Putra Malaysia
Serdang
Selangor
Malaysia

CONTENTS

PREFACE

This book is the second volume of research papers presented at the Fundamental Science Congress 2017 at Universiti Putra Malaysia on November 21-22, 2017.The congress served as a platform for researchers from different parts of Malaysia to share their knowledge and initiate collaboration among themselves. This book presents the latest findings in various fields of Statistics.

Chapter 2 propose a statistical method to identify outliers in the response variable of a simple circular regression model with high ratio of contamination. The proposed method depends on the circular distance between circular residuals and its trimmed mean as a measure of identification. The results of the simulation study and real example data show that the proposed method is successful in detecting outliers in the response variable.

Chapter 3 present the comparision between frequentist and Bayesian logistic regression (BLR) for identifying the malaria risk factors in Abuja, Nigeria. The frequentist logistic regression identified gender, family sizes, indoor residual spray and windows and door nets as predictors of malaria in Abuja. Similar findings were found for BLR. However, more concise and better results were found using Bayesian Monte Carlo study via WinBUGS algorithm. Nonetheless, the present study showed that the BLR method was comparable to frequentist logistic model especially when non-informative prior with large was used.

Chapter 4 consists of a paper on Optical tomography. Optical tomography is one of the tomography methods which are non-invasive and non-intrusive system, consisting of emitter with detectors. This research are conducted in order to analyze and proved the capability of laser with Charge Coupled Device in an optical tomography system for measuring object diameter that exist in crystal clear water. Experiments in detecting and capturing static solid rod in crystal clear water are conducted using this hardware and software development.

In Chapter 5, presents Robus Heteroscedasticity consistent covariance matrix (RHCCM) based on modified generalized studentized residuals (MGt) based on DRGP(ISE). The RHCCM estimator is an alternative method in the case of unknown errors structure to remedy both the effect of leverage points and heteroscedasticity.

Chapter 6 discusses a violation of constancy of variance of error terms causes the problem of heteroscedasticity. The OLS estimate is no longer efficient in the presence of heteroscedasticity in a data set, because the OLS estimates will be biased and inconsistent. As an alternative, a weighted residuals (wild bootstrap) may be used to remedy this problem. However, the weakness of wild bootstrap is that, in the presence of outliers the estimates of the standard errors become large. Therefore, a robust wild bootstrap is formulated based on MM-GM6 estimator so that the problems of both

heteroscedasticty and outliers can be rectified. The results show that the proposed method performs better than the existing ones such as OLS, Wu, and Liu.

Chapter 7 introduces a new approach normalization techniques to enhance the K-Means algorithm. This is to remedy the problem of using decimal scaling approach, which has overflow weakness. Hence, the suggested approach is called new approach to decimal scaling (NADS). Furthermore, based on real life datasets, the performance of the suggested method is compared with the existing methods, which evidently indicates that the suggested method outperformed the existing methods with higher average maximum external validity measures, and lower computing time (in minutes). Consequently, the proposed method may be used as data preprocessing methods in distance-based clustering analysis.

Chapter 8 comprises a problem of collinearity among regressors and weighted regressors in the observed Fisher information matrix of maximum likelihood. Under a certain condition, collinearity can reduce variance estimates in the presence of high leverage points.

Chapter 9 presents a study focused on nine accessions of closely related 5 *Passiflora* species; i.e, *Passiflora quadrangularis*, *Passiflora maliformis*, *Passiflora incarnata,* 2 varieties of *Passiflora foetida* and 4 varieties of *Passiflora edulis* as an example aimed to study the purposes of multivariate analyses for species separation. Combination of morphological traits using appropriate set of multivariate analyses and molecular approaches are useful for distinguishing the closely related *Passiflora* species.

CONTRIBUTORS

Preface

Mohamat Aidil bin Mohamat Johari, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang,Selangor, Malaysia.

Chapter 1 Introduction

Mohamat Aidil bin Mohamat Johari, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.

Chapter 2 Detection of Outliers in a Simple Circular Regression Model

Ehab A. Mahmood[1], Habshah Midi[1] , Abdul Ghapor Hussin[2] and Jayanthi Arasan[1]
[1]Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia
[2] National Defence University of Malaysia, 57000, Kem Sungai Besi, Kuala Lumpur, Malaysia
Email: eee.mahmood@gmail.com

Chapter 3 Bayesian and Frequentist Logistic Regression Models on Malaria Risk Factors: A Comparative Study

Emmanuel Segun Oguntade[1], Shamarina Shohaimi[1,2], Meenakshii Nallapan[2], Alaba Ajibola Lamidi-Sarumoh[2]

[1]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Biology, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
Email: shamarina@upm.edu.my

Chapter 4 Statistical Analyses in Validating the Performance of Optical Tomography System in Measuring Object Diameter

J. Jamaludin[1], R.A. Rahim[2,3], M.H.F. Rahiman[4], J.M. Rohani[5], B. Naeem[6]
[1] Faculty of Engineering and Built Environment, Universiti Sains Islam Malaysia, 71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia.
[2] Faculty of Electrical and Electronic Engineering, Universiti Tun Hussien Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
[3]Process Tomography and Instrumentation Engineering Research Group (PROTOM-i), Infocomm Research Alliance, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.
[4] Tomography Imaging Research Group, School of Mechatronic Engineering, Universiti Malaysia Perlis, 02600Arau, Perlis, Malaysia.
[5] Jemmy Mohd Rohani Enterprise, No 43, Jalan Merak ½, Bandar Putra, 81000 Kulai, Johor.
[6]Faculty of Information and Communication Technology, Balochistan University of IT, Engineering and Management Sciences, Quetta 87300, Pakistan.
Email:juliza@usim.edu.my

Chapter 5 New Weighting Method for Robust Heteroscedasticity Consistent Covariance Matrix Estimator in Linear Regression

Habshah Midi[1,2], Muhammad Sani[1,3], Mohd Shafie Mustafa[2] and Jayanthi Arasan[2]
[1]Institute for Mathematical Research, Universiti Putra Malaysia Serdang, Malaysia
[2]Department of Maths, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia.
[3]Department of Mathematical Sciences, Federal University Dutsinma, P.M.B. 5001 Katsina State, Nigeria.
Email: sanimksoro@gmail.com

Chapter 6 On the Performance of Wild Bootstrap based on MM-GM6 Estimator in the Presence of Heteroscedastic Errors and High Leverage Points

O.A.A. Alsattari[1], H. Midi[2]
[1]Department of Mathematics, UPM, 43400, Serdang, Selangor, Malaysia
[2]Institute of Mathematical Research, UPM, 43400, Serdang, Selangor, Malaysia
Email:osama19892014@gmail.com, habshah@upm.edu.my

Chapter 7 New Approach to Normalization Technique in K-Means Clustering Algorithm.

Paul Inuwa Dalatu[1,2,*], Habshah Midi[1], Jayanthi Arasan[1] and Ibragimov Gafurjan[1]
[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.
[2]Department of Mathematics, Faculty of Science, Adamawa State University, Mubi, PMB 25 Mubi, Adamawa State, Nigeria.
Email: dalatup@gmail.com

Chapter 8 The Effect of High Leverage Points on Collinearity Diagnostic in Logistic Regression Model

S.B. Ariffin[1], H. Midi[2] and J. Arasan[3]
[1]Department of Mathematics, UPM, 43400, Serdang, Selangor, Malaysia
[2]Institute of Mathematical Research, UPM, 43400, Serdang, Selangor, Malaysia
Email: syaibabalqish@gmail.com, habshah@upm.edu.my

Chapter 9 Analysis of genetic diversity in closely related plant species using multivariate analyses in comparison with molecular marker evidence

R. Shiamala Devi[1], B. Japar Sidik[2] and Z. Muta Harah[3]
[1]Department of Crop Science, Faculty of Agriculture and Food Sciences, Universiti Putra Malaysia Bintulu Sarawak Campus, Sarawak
[2]Department of Biology, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
[3]Department of Aquaculture, Faculty of Agriculture, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
Email: shiamala_32@yahoo.com

CHAPTER 1

INTRODUCTION

Statistics is a branch of mathematics. Statistics dealing with data collection, organization, analysis, interpretation and presentation. Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments. More over, statistics is concerned on the analysis of data and decision making based upon data. This multidisciplinary book on latest discoveries in various fields of Statistics.

Bayesian statistics is a subfield of statistics based on the Bayesian interpretation of probability where probability expresses a *degree of belief* in an event, which can change as new information is gathered, rather than a fixed value based upon frequency or propensity. The degree of belief may be based on prior knowledge or information about the event, such as the results of previous experiments, or on personal beliefs about the event. This differs from a number of other interpretations of probability, such as the frequentist interpretation that views probability as the limit of the relative frequency of an event after a large number of trials. Bayes' theorem are used to compute and update probabilities after obtaining new data. Bayes' theorem describes the conditional probability of an event based on data as well as prior information or beliefs about the event or conditions related to the event.

Multivariate statistics is also a subfield of statistics. Multivariate statistics encompassing the simultaneous observation and analysis of more than one outcome variable. The application of multivariate statistics named as multivariate analysis. Multivariate statistics is used to understand the different aims and background of each of the different forms of multivariate analysis, and how they relate to each other. The practical application of multivariate statistics to a particular problem may involve several types of univariate and multivariate analyses in order to understand the relationships between variables and their relevance to the problem being studied.

Another subfield of Statistics is robust statistics. Robust statistics is used to provide methods that emulate popular statistical methods, but which are not unduly affected by outliers or other small departures from model assumptions. In statistics, classical estimation methods rely heavily on assumptions which are often not met in practice. In particular, it is often assumed that the data errors are normally distributed, at least approximately, or that the central limit theorem can be relied on to produce normally distributed estimates. However, classical estimators often have very poor performance when there are outliers in the data.

Regression analysis is a subfield of statistical processes. It is used to estimate the relationships among variables. It have many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent

variable and one or more independent variables. In other word, regression analysis helps us to understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Regression analysis is widely used for prediction and forecasting.

This book is very useful not only for statisticians but also to statistics practitioners as a quick reference.

CHAPTER 2

Detection of Outliers in a Simple Circular Regression Model

Abstract

The existence of outliers in any type of data influences the efficiency of the estimators and the study of the results. In the literature, many methods have been proposed to identify outliers in the simple linear regression model. However, few methods have been proposed to identify outliers in the simple circular regression model. Moreover, these proposed methods did not succeed to identify outliers especially with high ratio of contamination. This motivated us to propose a statistical method to identify outliers in the response variable of a simple circular regression model with high ratio of contamination. The proposed method depends on the circular distance between circular residuals and its trimmed mean as a measure of identification. The results of simulation study and real example data show that the proposed method is successful detect outliers in the response variable.

Keywords: Outlier, Circular Regression Model, COVRATIO statistic, von Mises

Introduction

The simple circular regression model is used to represent the relationship between the response and the explanatory variables when both of them are circular. This model can be used in many scientific fields such as Meteorology, Biology, Physics and Medicine. However, the existence of outliers can cause a huge effect of the statistical analysis and the final outcomes. In real data, samples might include noise, or outliers. Outlier is an observation which appears inconsistent (extreme) with the other observations in the statistical data and effect on the results (Barnett and Lewis, 1994). In the literature, few methods are developed to identify outliers in the simple circular regression model. Hassan et al. (2010) suggested the functional relationship model for circular variables and estimated the model parameters by using the maximum likelihood method. Abuzaid et al. (2011) suggested using the *COVRATIO* statistic to detect outliers in the response variable of a simple circular regression model. Rambli (2011) adapted COVRATIO and the mean circular error statistic MCEs that were proposed by Abuzaid (2010) to identify outliers in circular regression model. Abuzaid et al. (2013) proposed the mean circular error statistic DMCEc to identify outliers in the response variable of a simple circular regression model, by using a row deletion approach. Abuzaid (2013) compared the performance of COVRATIO statistic for a simple circular regression model (SC) and a complex linear regression model (CL). He found that the COVRATIO statistic performs better for the SC model than for the CL model. Hussin et al. (2013) proposed a complex linear regression model to fit the circular data, using the complex residuals to detect any possible outliers. However, the problem of outliers detection in a simple circular regression model has not received enough consideration. In this paper, a new approach is proposed to identify outliers in the response variable of a simple circular regression model.

Methodology

The simple circular regression model is given by Hussin et al. (2004) as follows:

$y_i = \alpha + \beta x_i + \varepsilon_i \pmod{2\pi}$

where α and β are the parameters and ε is the circular random error, which follows the von Mises distribution with a circular mean μ and concentration parameter *k*. The probability density function of von Mises distribution with mean direction μ and concentration parameter *k* is given as follows (Mardia and Jupp, 2000):

$$g(\vartheta, \mu, k) = \frac{1}{2\pi I_0(k)} e^{k\cos(\vartheta - \mu)}$$

where I_o denotes the modified Bessel function of the first kind and order zero.
The maximum likelihood estimates of the parameters of the simple circular regression model are given as follows (Hussin et al. 2004) :

$$\hat{\alpha} = \begin{cases} \tan^{-1}(s/c) & \text{if } s > 0 \text{ , } c > 0 \\ \tan^{-1}(s/c) + \pi & \text{if } c < 0 \\ \tan^{-1}(s/c) + 2\pi & \text{if } s < 0 \text{ , } c > 0 \end{cases}$$

where :

$S = \sum \sin(y_i - \hat{\beta} x_i)$, $C = \sum \cos(y_i - \hat{\beta} x_i)$

$$\hat{\beta}_1 \approx \hat{\beta}_o + \frac{\sum x_i \sin(y_i - \hat{\alpha} - \hat{\beta}_o x_i)}{\sum x_i^2 \cos(y_i - \hat{\alpha} - \hat{\beta}_o x_i)}$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i (\mathrm{mod} 2\pi)$$

COVRATIO Statistic

Abuzaid et al. (2011) proposed *COVRATIO* statistic to detect outliers in the response variable of a simple circular regression model. The *COVRATIO* statistic is given by:

$$\mathrm{COVRATIO}_{(-i)} = \frac{|\mathrm{COV}_{(-i)}|}{|\mathrm{COV}|}$$

where: |COV| is the determinant covariance matrix of coefficients for the full data set, $|\mathrm{COV}| = \frac{1}{\hat{k} A(\hat{k})}$ and $|\mathrm{COV}_{(-i)}|$ is the determinant covariance matrix of coefficients for the reduced data set formed by excluding the i-th row, $|\mathrm{COV}_{(-i)}| = \frac{1}{\hat{k}_{(-i)} A(\hat{k}_{(-i)})}$. The cut-off

point represents the maximum value of the statistic $\left|\mathrm{COVRATIO}_{(-\mathrm{i})}-1\right|$ Abuzaid et al. (2011).

$$\text{cut COVRATIO}=\max\left|\mathrm{COVRATIO}_{(-\mathrm{i})}-1\right|$$

The i-th observation is identified as an outlier if $\left|\mathrm{COVRATIO}_{(-\mathrm{i})}-1\right|$ exceeds the cut-off point.

Mean Circular Error Statistic

Abuzaid (2010) and Abuzaid et al. (2013) suggested two statistics, the DMCEs and DMCEc statistics, to identify outliers in the response variable y in a simple circular regression model.

DMCEs Statistic

$$\mathrm{MCEs}=\frac{1}{\mathrm{n}}\sum\sin\left(\frac{\mathrm{d_i}}{2}\right)$$

where $\left(\mathrm{d_i}=\pi-\left|\pi-\left|\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i}\right|\right|\right),\ \mathrm{MCEs}\in[0,1]$

The statistic to detect outliers is given as follows:

DMCEs(i) = | MCEs – MCEs(-i)|

where MCEs(-i) is MCEs with the i-th observation removed.

The cut-off point represents the maximum absolute difference between the value of the statistic for the full data and the reduced data set (formed by excluding the i-th observation).

$$\text{cut DMCEs}=\max\left|\mathrm{MCEs}-\mathrm{MCEs}_{(-\mathrm{i})}\right|$$

The i-th observation is identified as an influential observation if DMCEs(i) is greater than the cut-off point.

DMCEc Statistic

$$\mathrm{MCEc}=1-\frac{1}{\mathrm{n}}\sum\cos(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})$$

where, $\mathrm{MCEc}\in[0,2]$.

The statistic to identify outlier is given as follows:

DMCEc(i) = | MCEc – MCEc(-i)|

where MCEc(-i) is MCEc with the i-th observation removed. The cut-off point is the maximum absolute difference between the value of the statistics for the full data set and the reduced data sets.

$$\text{cut DMCEc}=\max\left|\mathrm{MCEc}-\mathrm{MCEc}_{(-\mathrm{i})}\right|$$

If DMCEc(i) is greater than the cut-off point, the i-th observation is detected as an outlier.

The Proposed Method

A new method is proposed to identify outliers in the response variable of a simple circular regression model based on the circular distance theory between two circular observations. We call this procedure the robust circular distance RCD_y, because it depends on the robust circular distance between any circular error and its mean direction. The proposed method is computed according to the following steps.

Step 1. Calculate the absolute value of the estimated circular error $\hat{e}_i$. We propose to calculate this according to the following:

i- If $(0 \leq \hat{y}_i \leq \pi)$:

$$|\hat{e}_i| = \begin{cases} |y_i - \hat{y}_i| & \text{if } |y_i - \hat{y}_i| \leq \pi \\ 2\pi - y_i + \hat{y}_i & \text{if } |y_i - \hat{y}_i| > \pi \end{cases}$$

ii- If $(\pi < \hat{y}_i \leq 2\pi)$:

$$|\hat{e}_i| = \begin{cases} |y_i - \hat{y}_i| & \text{if } |y_i - \hat{y}_i| \leq \pi \\ 2\pi - \hat{y}_i + y_i & \text{if } |y_i - \hat{y}_i| > \pi \end{cases}$$

where $[0 \leq |\hat{e}_i| \leq \pi]$

Step 2. Compute the trimmed mean.
Step 3. Compute the circular distance $[dist_{(i)}]_y$ between $|\hat{e}_i|$ and its trimmed mean as follows:

- If $(0 \leq \mu_t \leq \pi)$:

$$[\text{dist}_{(i)}]_y = \begin{cases} \big||\hat{e}_i| - \mu_t\big| & \text{if } \big||\hat{e}_i| - \mu_t\big| \leq \pi \\ 2\pi - |\hat{e}_i| + \mu_t & \text{if } \big||\hat{e}_i| - \mu_t\big| > \pi \end{cases}$$

- If $(\pi < \mu_t \leq 2\pi)$:

$$[\text{dist}_{(i)}]_y = \begin{cases} \big||\hat{e}_i| - \mu_t\big| & \text{if } \big||\hat{e}_i| - \mu_t\big| \leq \pi \\ 2\pi - \mu_t + |\hat{e}_i| & \text{if } \big||\hat{e}_i| - \mu_t\big| > \pi \end{cases}$$

$[0 \leq [\text{dist}_{(i)}]_y \leq \pi]$

Finally, if $y_{(i)}$ is an outlier then $[dist_{(i)}]_y$ is expected to be relatively large. Therefore, the cut-off point should be the following:

$$RCD_y = \max[\text{dist}]_y$$

Consequently, $y_{(i)}$ is identified as an outlier if $[dist_{(i)}]_y$ exceeds the cut-off point.

Results

Simulation Study

The performance of the RCD_y statistic is investigated by comparing with the *COVRATIO*, *DMCEs* and *DMCEc* methods using Monte Carlo simulations. We compare the results by using four sample sizes, n =10, 50, 100, and 150, and six concentration parameters, k = 3, 5, 8, 10, 20, and 30. Following Abuzaid et al. (2013), the response variable y is contaminated according to the following formulas:

$$yc_{[i]} = y_{[i]} + \lambda\pi \bmod(2\pi)$$

where:
$yc_{[i]}$ is the contaminated circular observation.
λ is the degree of contamination, such that ($0 \leq \lambda \leq 1$).

For all combinations of sample sizes and concentration parameters, we generate 20% and 30% contaminated data with $\lambda = 0.8$. We replicate these processes 5000 times for each combination of sample size and concentration parameter. The values of cut-off points of RCD_y statistic with 5% upper percentile are given in Table 1.
To evaluate the performance of all the statistics, the proportion of outliers detected is computed. The results are shown in Figures 1–2.

Table 1: Cut-off points of the RCD_y statistic with 5% upper percentile

n \ *k*	2	3	5	6	8	10	15	20	30
10	2.26	1.72	1.45	1.39	1.12	0.703	0.494	0.409	0.325
20	2.33	2.02	1.18	1.01	0.878	0.814	0.746	0.561	0.430
30	2.36	1.52	1.23	1.10	0.892	0.772	0.621	0.567	0.517
40	2.39	2.20	1.27	1.11	0.910	0.822	0.637	0.534	0.457
50	2.40	1.75	1.36	1.17	0.959	0.819	0.668	0.569	0.448
60	2.42	2.30	1.39	1.18	0.982	0.831	0.672	0.577	0.457
70	2.42	2.35	1.45	1.22	0.990	0.853	0.677	0.582	0.465
80	2.43	2.37	1.49	1.25	1.00	0.888	0.688	0.587	0.474
90	2.43	2.40	1.50	1.27	1.02	0.896	0.700	0.596	0.480
100	2.44	2.43	1.53	1.29	1.03	0.905	0.718	0.603	0.488
110	2.45	2.45	1.57	1.30	1.05	0.911	0.719	0.604	0.494
120	2.45	2.45	1.58	1.31	1.05	0.920	0.725	0.610	0.496
130	2.45	2.46	1.59	1.32	1.06	0.938	0.734	0.616	0.498
140	2.45	2.47	1.60	1.33	1.07	0.944	0.735	0.625	0.500
150	2.45	2.49	1.61	1.35	1.08	0.948	0.737	0.635	0.504

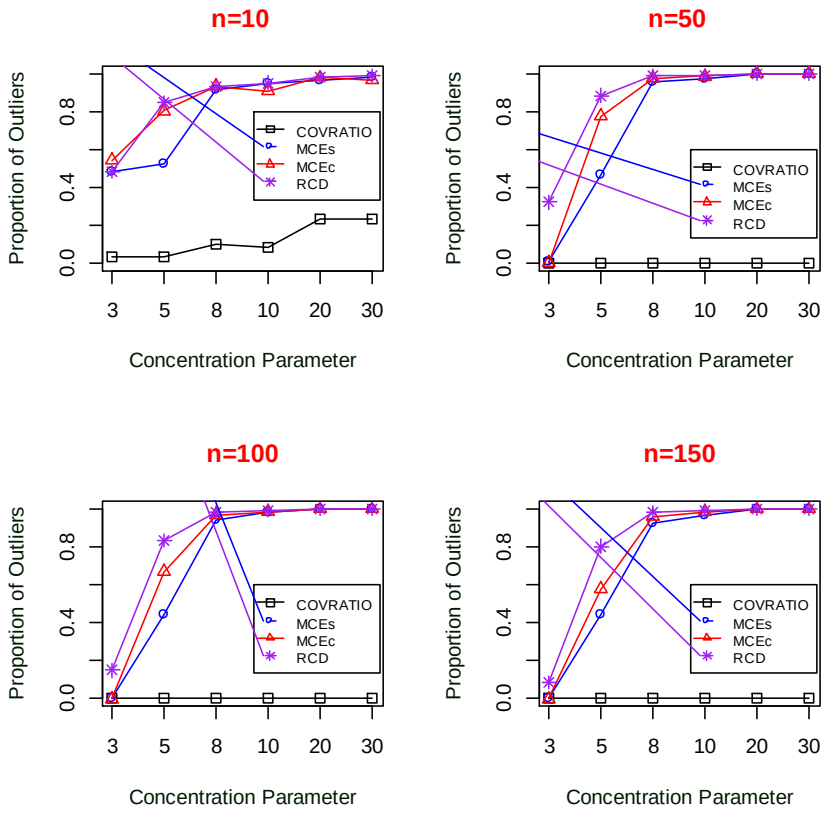


Figure 1: Proportion of outliers detected with 20% contamination

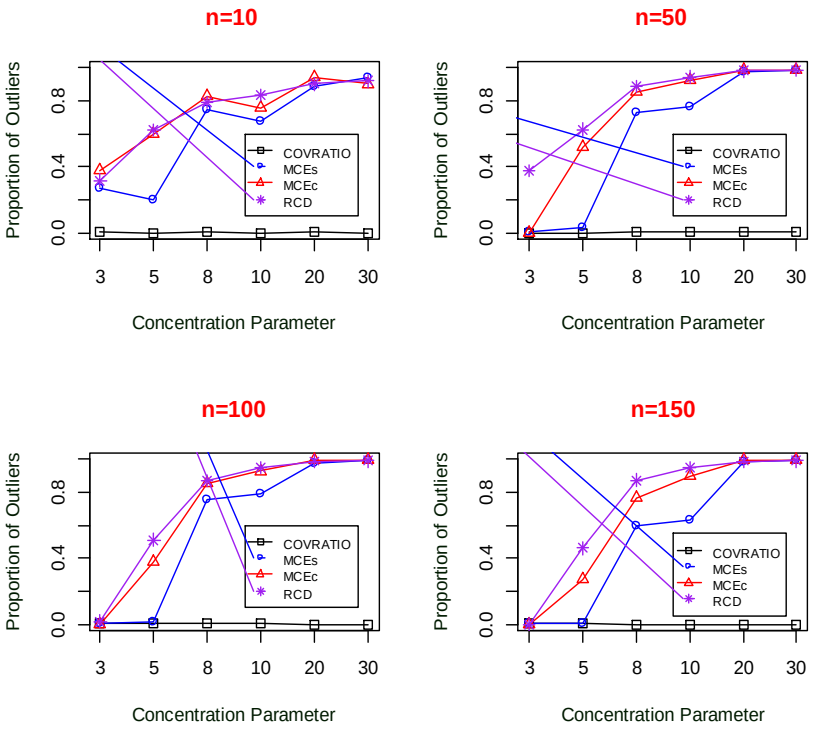


Figure 2: Proportion of outliers detected with 30% contamination

We can notice that *COVRATIO* statistic not true for different sample sizes with 20% and 30% of contaminations. The MCEs and MCEc gave relatively low proportion of detection. In contrast, the RCD_y statistic gives the greatest proportion of outliers detected than the other methods. The proportion is an increasing function of the concentration parameter, and increases to 100% for values of the concentration parameter greater than 10.

Practical Example

We study the data set that were collected along the Holderness coastline (the Humberside coast of the North Sea, United Kingdom) Hussin, 1997. There were 78 measurements recorded by HF radar system (OSCR) and anchored wave buoy. The deployment began in October 1994. The observations 60, 70 and 71 are identified as outliers (Hussin, 1997). The RCD_y statistic is calculated and the results are plotted in Figure 3. The $[RCD60]_y$, $[RCD70]_y$ and $[RCD71]_y$ exceed the cut-off point, so the observations number 60, 70 and 71 are classified as outliers. These detections correspond with those given by (Hussin, 1997). The RCD_y statistic is very successful to identify outliers

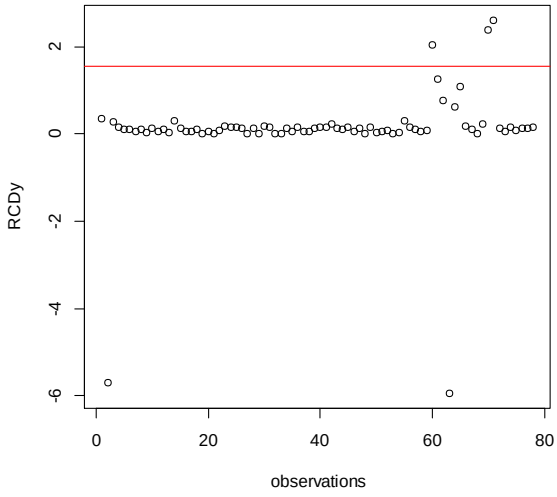


Figure 3: RCD_y statistic of the wind direction data (n=78)

Conclusion

In this article, a robust method is proposed to identify outliers in the response variable of a simple circular regression model. The circular distance between circular residuals and its trimmed mean is proposed as a measure to detect outliers. The results show that the proposed method is successful identifying outliers.

References

1. Abuzaid, A. (2010) Some problems of outliers in circular data. Ph.D. thesis, Faculty of Science, University Malaya.
2. Abuzaid, A. (2013) On the influential points in the functional circular relationship models. Pakistan Journal Statistics Operation Research, IX(3), 333-342.
3. Abuzaid, A., Mohamed, I.B., Hussin, A.G. and Rambli, A. (2011) COVRATIO statistic for simple circular regression model. *Chiang Mai Journal of Science*, 38(3) 321-330.
4. Abuzaid A., Hussin, A.G. and Mohamed, I.B. (2013) Detection of outliers in simple circular regression models using the mean circular error statistic. *Journal of Statistical Computation and Simul*ation, 83(2):269-277.
5. Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data* (3rd edition). Wiley, New York.
6. Hassan S. F., Hussin, A.G. and Zubairi, Y.Z. (2010) Estimation of functional relationship model for circular variables and its application in measurement problems. *Chiang Mai Journal Sciences* 37(2), 195-205.
7. Hussin, A. G. (1997). *Pseudo-Replication in Functional Relationships with Environmental Applications*. Unpublished PhD. thesis, University of Sheffield,UK.
8. Hussin, A.G., Fieller, N.R.J. and Stillman, E.C. (2004) Linear Regression for Circular Variables with Application to Directional Data. *Journal of Applied Science and Technology*, 8(1-2) ,1-6.

9. Hussin, A.G., Abuzaid A.H., Ibrahim, A.I.N. and Rambli, A. (2013) Detection of outliers in the complex linear regression model. *Sains Malaysiana* 42(6), 869-874.
10. Mardia, K. V. and Jupp, P. (2000) *Directional Statistics*, Wiley, London.
11. Rambli, A. (2011) *Outlier detection in circular data and circular-circular regression model*. Unpublished Master's Thesis, University of Malaya, Malaysia.

CHAPTER 3

Bayesian and Frequentist Logistic Regression Models on Malaria Risk Factors: A Comparative Study

Abstract

Despite numerous interventions against malaria cases, the number of malaria cases is still on the increase, particularly in the tropics. The present study aimed to compare frequentist and Bayesian logistic regression (BLR) for identifying the malaria risk factors in Abuja, Nigeria. The study design was a cross-sectional study on 384 participants selected randomly from the four strata (cardinal points) of Gwagwalada Area Council Abuja. The data were collected from the month of March to September 2016 using a validated structured questionnaire. Results from multivariable logistic regression (optimal number of parameters chosen via MASS and BMA in R packages) and BLR analyses were compared. The frequentist logistic regression identified gender, family sizes, indoor residual spray and windows and door nets as predictors of malaria in Abuja. Similar findings were found for BLR. However, more concise and better results were found using Bayesian Monte Carlo study via WinBUGS algorithm. Nonetheless, the present study showed that the BLR method was comparable to frequentist logistic model especially when non-informative prior with large was used. The higher the precision assumed in the prior probability, the better the results especially with larger sample sizes.

Keywords: Bayesian, Frequentist, Prior, Likelihood, Malaria, Nigeria.

1.0 Introduction

Malaria is caused by the bite of the infected female *Anopheles* mosquito which is active from dusk till dawn as it seeks blood for its eggs. *Plasmodium* parasites causes malaria in mammals, reptiles and even birds and the main source of transmissions are basically through mosquito of the genus Anopheles (Abdullahi et al., 2015). Other modes of transmission include organ transplant, congenital transmission through birth, blood transfusion as well as using unsterilized objects like syringe, needles, blades etc. Malaria symptoms are fever, high temperature, pain, and weakness of the body. It is curable but can relapse if not properly treated and possible recurrence as a result of dormant parasites in the liver cells (NIH, 2007). The most vulnerable individuals are those with lower level of immunity, for example children aged five or less, pregnant women and people with other forms of diseases (Woyessa et al., 2013).

Malaria is a global phenomenon that has affected tropical and sub-tropical countries of the world where the breeding of host vectors is favoured by the prevailing environmental conditions, human sanitation, irrigation and agricultural practices (Babajide et al., 2015). Despite many control interventions in place malaria is still widespread. Some of the control interventions in sub-Sahara Africa (SSA) are vector control mechanism through insecticide treated nets (ITNs), windows and door nets (WDNs) and indoor residual spray (IRS) which have been found effective through proper ownership and usage; health promotions through health education interventions; prompt and adequate case management through artemisinin combination therapy for uncomplicated malaria as

recommended by World Health Organization and intermittent preventive therapy for women in pregnancy; cross border surveillance among others (Moonasar et al., 2012).

Many models have been proposed in literature for identifying basic factors of malaria outbreaks especially from frequentist point of views, but such method suffers flexibility and accuracy as parameters of interest are assumed fixed and unknown quantities (Stauffer, 2008). Such quantities can only be estimated with maximum likelihood estimation (MLE) to maximize the probability of achieving the desired results within an estimated interval (confidence interval (CI)) with a fixed probability. However, parameters need not be fixed and can be viewed as a random variable incorporating probability inform of belief about such parameters which are specified based on available information before observations were made. Such beliefs when updated and improved upon gives better understanding of the process.

The process of updating beliefs with the likelihood of the parameter given the data set is the idea behind Bayesian theory. The Bayesian statistical inference (BSI) utilizes basically the prior, posterior distributions and the likelihood functions from the data and models to estimate parameters. This can be achieved through Monte Carlo simulation from the posterior distribution via WinBUGS software. The conventional frequentist methods suffer some setbacks to pave way for general adoption and application of Bayesian analysis with advent of modern computer that makes simulations and data generation mechanisms much easier. Visualization of repeated sampling is difficult in frequentist statistical inference and sampled observations may not always be random, non-intuitive and confusing way of estimating parameter of interest without assignment of probability to parameters (Stauffer, 2008). In essence, the frequentist method assumed fixed values for the parameter within an estimated interval while prior information is ignored. BSI overcomes some of the short comings assuming parameters to be random, incorporation of prior knowledge (as known probability distribution) and likelihood based on the observed dataset.

Studies have been done to compare estimations from both the Bayesian and logistic regression models. For example, Tsai (2004) applied binary logistic to Taipei mayoral election with Bayesian inference incorporating probability inform of vague prior for the election parameters for desired interest to be achieved while similar model was used to explore combinations of risk factors of type 2 diabetes mellitus among selected men and women in Malaysia (Chiaka et al., 2015).

Several authors have used Bayesian statistics in identifying malaria risk factors across SSA (Diboulo et al., 2015; Pullan et al., 2010). However, there are limited studies on malaria risk factors in Nigeria that employed the Bayesian logistic regression (BLR) method using different non-informative priors. Thus, this study explored various priors for the estimation of the model parameters for the malaria risk factors. Also, the performances of BLR in identifying the malaria risk factors were compared in relation to frequentist logistic regression (FLR).

2.0 Materials and Methods

2.1 Data

The data for this study were obtained from a cross-sectional study conducted in Gwagwalada-Abuja, Nigeria in 2016. The subjects were recruited into the study voluntary based on desire to participate. Health status of subjects in relation to malaria in the last 15 days (a criterion based on NPC (2012)) were considered as outcome variable for logistic regression model (LRM) with eleven risk factors. These are gender, age group, marital status, occupation, education, family size, socio-economic status, residential area, ITNs, IRS and WDNs.

2.2 Logistic regression

Logistic regression is a statistical tool for modelling a binary dependent variable with one or more independent variables. It is a generalized linear regression model which uses logit as link function for the transformation of the model components. Let $\lambda(x_i)$ represents the probability of event for individual i, x_i are the vectors of risk factors and criterion variable denoted by y, which assumes a value 1 for the probability of occurrence and 0 if otherwise. The logistic function is represented by (1)

$$\lambda(x_i)=\frac{\ell^{\alpha+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k}}{1+\ell^{\alpha+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k}};\quad \frac{\lambda(x_i)}{1-\lambda(x_i)}=\frac{\ell^{\alpha+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k}}{1} \tag{1}$$

where: $\lambda(x_i)$ is a probability of having malaria infection

β_i are the slope parameters

α is the intercept

x_i are the independent variables (malaria risk factors)

The outcome variable y depends on the independent variables (x_i), then, the logarithm transformation of the odds gives (2)

$$\text{logit}\,\lambda(\mathrm{x_i}) = log\left(\frac{\lambda(x_i)}{1-\lambda(x_i)}\right)=\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \tag{2}$$

where $y_i = bern(1,\lambda_i)$, $y_i|x^{|}\beta = \lambda^{y_i}(1-\lambda)^{1-y_i}$

2.3 Bayes rule

$$P(\lambda\,|\,Data)=\frac{p(\lambda)\,p(Data\,|\,\lambda)}{P(Data)} \tag{3}$$

Where: $p(\lambda)$ is the prior probability

$p(Data|\lambda)$ is the likelihood of Data given λ

$p(Data)$ is the marginal or the total probability

2.4 Likelihood

Let $x_1, x_2, \ldots, x_k$ be a set of independent Bernoulli distributed random variables, then, the likelihood function is given by (4):

$$L(x_1, x_2 \ldots x_k) = \Pi_i P(x_i / \lambda) \; ; L(x_1, x_2 \ldots x_k) = \Pi_i \{\lambda^{x_i}(1-\lambda)^{1-x_i}\} = \lambda^{\sum x_i}(1-\lambda)^{k-\sum x_i}$$

$$L(x_1, x_2 \ldots x_k) = \lambda^{\Omega}(1-\lambda)^{k-\Omega} \quad \forall \Omega = \sum x_i \tag{4}$$

From (1), it follows that:

$$L(x_1, x_2 \ldots x_k) = \left(\frac{\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}\right)^{\Omega}\left(1-\frac{\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}\right)^{k-\Omega} \quad \forall \Omega = \sum x_i \tag{5}$$

2.5 Prior probability distribution

We assumed normal prior for our coefficients with zero mean and large variance to make it non-informative prior. Likewise, we also assumed a prior that is not perfectly flat with mean zero and a small variance. The parameters were assigned zero mean and precision 1, 0.001, and 0.000001. The choice of the priors with large variance was to minimize the influence of prior on the likelihood and also to give every value of the parameters the same likelihood. This was done in line with literature on malaria risk factors (Onyiri, 2015) and other related study (Chiaka et al., 2015).

2.6 Posterior distribution

With the aid of Bayesian theorem, the kernel from the posterior is given by the product of the likelihood and the prior probability assumed for the parameters of interest. Thus, using equation (5) and a normal prior give the form of the posterior distribution (6).

$$Posterior \propto \left(\frac{\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}\right)^{\Omega}\left(1-\frac{\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+\ell^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}\right)^{k-\Omega} . \prod_{i=0} \{\sqrt{2\Pi\delta^2}\}^{-1} \exp\{0-0.5\left(\frac{\beta_i-\mu_i}{\delta}\right)^2\} \tag{6}$$

2.7 Markov Chain Monte Carlo analysis via WinBUGS

As in other Bayesian analysis, the BLR requires stating the following steps: the joint distribution of the outcome variable and all the model parameters, the likelihood function and posterior density in the regression. Based on the WinBUGS algorithms, the following assumptions were made for the logit model in line with related studies (Chiaka et al., 2015; Onyiri, 2015).

(i) $y_i \sim dbern$, the outcome variable sis Bernoulli distributed

(ii) $b_i \sim dnorm(0, k)$, a normal prior assumed for each of the slope parameters

(iii) $\alpha = 0. \beta_i = 1$, the initial values for each of the parameters

(iv) Iterations = 5,000; 50,000; 100,000; 150,000.

3.0 Results and Discussion

3.1 Demographic characteristics

Table 1 displays the participants' demographic characteristics. The average age group of respondents was 32± 9 years. The unmarried were about one-third of the respondents are unmarried, while the households were characterized by large family sizes with over 60% having a family size greater than or equals 4 persons. Of the 384 respondents, only 134 (34.9%) had a tertiary education while the remaining had at most secondary. Refer to Table 1, the goodness-of-fit test revealed that there is a significant difference for each level of the socio-demographic variables we considered ($p<0.05$).

Table 1: Socio-demographic characteristics

Variables	n (%)	χ^2	*p*-value
Age groups			
18-25	136(35.4)		
26-35	123(32.1)		
36-45	88(22.9)		
46 and above	37(9.6)	61.188	<0.001*
Gender			
Male	234(60.9)		
Female	150(39.1)	18.375	<0.001*
Marital status			
Married	203(52.9)		
Divorced	25(6.5)		
Widow	19(5.0)		
Single	137(35.6)	2.510E2	<0.001*
Education			
Tertiary	134(34.9)		
Secondary	127(33.1)		
Elementary	58(15.1)		
Illiterate	65(16.9)	50.104	0.001*
Social status			
Poor	124(32.3)		
Average	173(45.0)		
Rich	87(22.7)	29.078	<0.001*
Family size			
1	31(8.1)		
2	35(9.1)		
3	45(11.7)		
4	100(26.0)		
More than 4	173(45.1)	190.740	<0.001*

χ^2: Pearson chi-squared goodness-of-fit test for significant difference in observed scores/frequency.
* Statistical significance at $p<0.05$.

3.2 Logistic regression model results

Based on Table 2, the final multiple LRM included only four significant predictors (p<0.05). The female gender, smaller family size, non-usage of IRS and WDNs showed susceptibility towards malaria as odds of the disease were more likely compared to the reference categories. The overall fit of the final model over reduced model was assessed with the likelihood ratio test (-2LL=180.51, p≤0.001).

Table 2: Multivariate logistic regression parameter estimates

Model Predictor	Model Coefficient	Std Error	Wald's Chi-Sq.	OR	95% CI	P-Value
Gender						
Male	0[a]	-	-	1.000	-	-
Female	1.238	0.303	16.639	3.448	1.904, 6.245	<0.001*
Family Size						
<-4 Persons.	1.003	0.294	11.602	2.726	1.532, 4.851	0.001*
>4 Persons.	0[a]	-	-	1.000	-	-
IRS						
No	0.727	.246	8.702	2.068	1.277, 3.350	0.003*
Yes	0[a]	-	-	1.000	-	-
WDNs						
No	0.826	0.251	10.800	2.284	1.396, 3.735	0.001*
Yes	0[a]	-	-	1.000	-	-
Intercept	0.431	0.177	5.968	1.538	1.089, 2.176	0.015*

[a]Reference category, *Significant, OR=Odds Ratio; WDNs: Window and door nets, IRS: Indoor residual spray

3.3 Bayesian results

There were variations in the output of the simulation, especially with lower sample sizes using different values of precision. However, the means of the posterior distribution of the model coefficients are similar to FLR model when non-informative prior with large variance were used and at larger sample sizes (Tables 3). It thus allows the data to dictate the output of the analysis due to minimal influence of the diffuse prior. Considering the standard deviations of the estimated coefficients, the results revealed that the standard deviations of the non-informative prior with small variance (precision=1) (Table 4) were smaller than that of the non-informative prior with large variance (precision=0.000001) and the FLR. This implies that the BLR with the smaller standard error is better. Therefore, the higher the precision the better the results especially when the sample size is 150,000 Figure 1 reveals the time series plot of the value of the parameter against the iteration number. It showed that there was no bad mixing as the chain was not stuck in the parameter space.

Table 3: Bayesian Monte Carlo output (Precision=0.000001, *n*=150,000)

Node	Mean	SD	MC Error	Credible Interval	
				2.5%	97.5%
Alpha	3.4420	0.4598	0.0051	2.5640	4.3720
Gender	-1.2730	0.3085	0.0026	-1.9010	-0.6893
Family size	-1.0280	0.2999	0.0024	-1.6320	-0.4530
WDNs	-0.8405	0.2564	0.0017	-1.3530	-0.3459
IRS	-0.7399	0.2485	0.0014	-1.2340	-0.2581

SD: Standard deviation, MC: Monte Carlo, WDNs: Window and door nets, IRS: Indoor residual spray

Table 4. Bayesian Monte Carlo Output (Precision=1, *n*=150,000)

Node	Mean	SD	MC Error	Credible Interval	
				2.5%	97.5%
Alpha	2.7370	0.3691	0.0035	2.0300	3.4810
Gender	-0.9729	0.2685	0.0019	-1.5100	-0.4564
Family size	-0.7510	0.2625	0.0018	-1.2780	-0.2478
WDNs	-0.6709	0.2371	0.0013	-1.1410	-0.2089
IRS	-0.6186	0.2325	0.0012	-1.0790	-0.1658

SD: Standard deviation, MC: Monte Carlo, WDNs: Window and door nets, IRS: Indoor residual spray

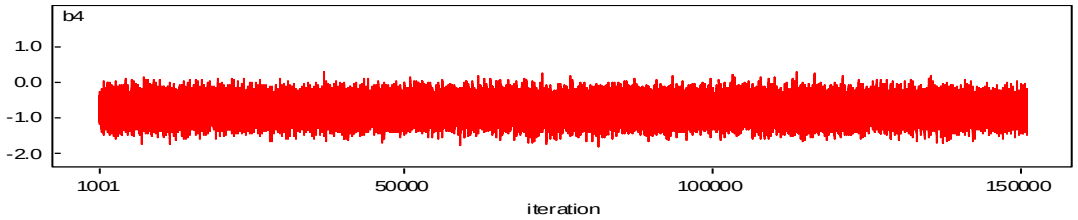


Figure 1: Trace plots for posterior distribution.

3.4 Discussion

The multivariate logistic regression analysis based in this study was conducted to identify significant factors of malaria epidemic in Abuja, Nigeria.

As stated by Fayehun and Salami (2014), the gender of the household heads was found to be associated with malaria cases, as in other study This study also showed that the odds of malaria cases were 3.45 times higher among households headed by a female than households without a female head. This implies that malaria is more likely among the household headed by a female with a chance of about 3 times greater than those headed by a male. The findings of the current study suggested that the possibility of the households being headed by a woman or presence of many children in the homes; as such suppression in immunity is expected in their body systems and make them more susceptible to malaria. This may likely increase the likelihood of more malaria cases reported.

The result also revealed that more likely cases of malaria with family sizes less than four persons. This finding negates intuitive and insightful reasoning, as less malaria cases are expected in a low densely populated areas or houses. This might be due to the fact that

man biting rate (MBR) increases with reduction in number of occupants. Based on this result, the households with lower family sizes are at a higher risk of stable and high-intensity malaria parasite transmissions as entomological inoculation rate might occasionally exceeds the threshold of 1.5 (Ebenezer et al., 2016). The higher the number of sporozoite-infected female *Anopheles* mosquitoes in a room or house relative to the number of occupants thereof, the higher the MBR and sporozoite rate. Hence, a higher disease transmission and more likely the disease cases reported. This finding was consistent with a study in SSA (Krefis et al., 2010). However, it contradicted with the findings of other related studies in Nigeria (Ajadi et al., 2012). That is, the higher likelihood of disease cases for larger family sizes.

The use of intervention measures had been appraised and well documented in the literature (Baume & Franca-Koh, 2011). The study revealed that the higher the level of usage of these measures, the lower the disease prevalence. The finding we obtain was supported by earlier research on poor usage of IRS and WDNs (Amoran, 2013), and it is incongruent with that detailed in Olayemi et al. (2012).

Based on the standard deviation and credible interval of the estimated coefficients, the results of BLR were comparable to FLR model with sufficient sample size simulated and at lower precision levels (Chiaka et al., 2015). However, it requires great caution as the choice of initial parameters and prior probabilities may affect the output of the simulation.

4.0 Conclusion

This study identified the risk factors of malaria in Abuja, Nigeria and also revealed the performances of both FLR and BLR. The significant factors are gender, family size, IRS and WDNs. Both methods yielded similar results when non-informative prior with large variance is assumed in BLR. The right choice of prior and large number of samples generated offer additional advantage for BLR.

References

Abdullahi, M. B., Hasan, Y. A., & Abdullah, F. A. (2015). A mathematical analysis of the effects of control Plasmodium knowlesi malaria. *Pakistan Journal of Statistics*, *31*(5), 483–514.

Ajadi, K. O., Olaniran, H. D., Alabi, F. M., & Adejumobi, D. O. (2012). Incidence of malaria among various rural socioeconomic households. *Greener Journal of Medical Sciences*, *2*(3), 051–063.

Amoran, O. E. (2013). Impact of health education intervention on malaria prevention practices among nursing mothers in rural communities in Nigeria. *Nigeria Medical Journal*, *54*(2), 115–122.

AMR. (2003). The burden of malaria in Africa. The Africa Malaria Report.

Babajide, S., Perry, B., Huffer, F. W., Onubogu, O., Dutton, M., Becker, A., & Saleh, R. (2015). Effect of meteorological variables on malaria incidence in Ogun State, Nigeria. *International Journal of Public Health and Epidemiology*, *4*(10), 205–215.

Baume, C. A., & Franca-koh, A. C. (2011). Predictors of mosquito net use in Ghana.

Malaria Journal, *10*(265), 1–6.
Baume, C. A., Reithinger, R., & Woldehanna, S. (2009). Factors associated with use and non-use of mosquito nets owned in Oromia and Amhara Regional States, Ethiopia. *Malaria Journal*, *8*(1), 264.
Chiaka, S. E., Adam, M. B., Krishnarajah, I., Shohaimi, S., & Guure, C. B. (2015). Bayesian logistic regression model on risk factors of Type 2 Diabetes Mellitus. *Mathematical Theory and Modelling*, *5*(1), 113–123.
Diboulo, E., Sié, A., Diadier, D. A., Voules, D. A. K., Yé, Y., & Vounatsou, P. (2015). Bayesian variable selection in modelling geographical heterogeneity in malaria transmission from sparse data : an application to Nouna Health and Demographic Surveillance System (HDSS) data , Burkina Faso. *Parasites and Vectors*, *8*(118), 1–14.
Ebenezer, A., Noutcha, A. E. M., & Okiwelu, S. N. (2016). Relationship of annual entomological inoculation rates to malaria transmission indices , Bayelsa State, Nigeria. *Journal of Vector Borne Disease*, *53*(March), 46–53.
Fayehun, O. A., & Salami, K. K. (2014). Older persons and malaria treatment in Nigeria. *Etude de La Population Africaine*, *27*(2 SUPPL.), 424–433.
Krefis, A. C., Schwarz, N. G., Nkrumah, B., Acquah, S., Loag, W., Sarpong, N., … May, J. (2010). Principal component analysis of socioeconomic factors and their association with malaria in children from the Ashanti region, Ghana. *Malaria Journal*, *9*, 201.
Loha, E., Lunde, T. M., & Lindjorn, B. (2012). Effect of bednets and indoor residual spraying on spatio-temporal clustering of malaria in a village in South Ethiopia: A longitudinal study. *PLoS ONE*, *7*(10).
Moonasar, D., Nuthulaganti, T., Kruger, P. S., Mabuza, A., Rasiswi, E. S., F.G., B., & Maharaj, R. (2012). Malaria control in South Africa 2000–2010: Beyond MDG6. *Malaria Journal*, *11*(294).
NIH. (2007). Understanding malaria: Fighting an ancient scourge. *NIH Publications*. USA: Department of Health and Human Services. Retrieved from www.niaid.nih.gov.
NPC. (2012). National Malaria Control Programme (NMCP) [Nigeria], and ICF International: 2010 Nigeria Malaria Indicator Survey. Abuja, Nigeria: NPC, NMCP, and ICF International. Retrieved from The 2010 Nigeria Malaria Indicator Survey,.pdf.crdownload
Olayemi, I. K., Omalu, I. C. J., Abolarinwa, S. O., Mustapha, O. M., Ayanwale, V. A., Mohammed, A. Z., … Chuckwuemeka, V. I. (2012). Knowledge of malaria and implication for control in an endemic urban area of North Nigeria. *Asian Journal of Epidemiology*, *5*(2), 42–49.
Onyiri, N. (2015). Estimating malaria burden in Nigeria: A geostatistical modelling approach. *Geospatial Health*, *10*(2), 163–170.
Pullan, R. L., Bukirwa, H., Staedke, S. G., Snow, R. W., & Brooker, S. (2010). Plasmodium infection and its risk factors in eastern Uganda. *Malaria Journal*, *9*(2), 1–11.
Siraj, A. S., Santos-Vega, M., Bouma, M. J., Yadeta, D., Ruiz-Carrascal, D., & Pascual, M. (2014). Altitudinal changes in malaria. *Science*, *1154*(March), 1154–1159.
Stauffer, H. B. (2008). *Contemporary Bayesian and Frequentist statistical research*

methods for natural resource scientists. Canada: John Wiley & Sons, Inc.
Tsai, C. (2004). Bayesian inference in Binomial logistic regression: A case study of the 2002 Tapei Moyoral Election. Taipei: Academia Sinica Taipei.
Woyessa, K., Deressa, W., Ali, A., & Lindtjørn, B. (2013). Malaria risk factors in Butajira area, south-central Ethiopia: a multilevel analysis. *Malaria Journal*, *12*, 273.
Yusuf, O. B., Adeoye, B. W., Oladepo, O. O., Peters, D. H., & Bishai, D. (2010). Poverty and fever vulnerability in Nigeria: a multilevel analysis. *Malaria Journal*, *9*(1), 235.

CHAPTER 4

Statistical Analyses in Validating the Performance of Optical Tomography System in Measuring Object Diameter

Abstract

Optical tomography is one of the tomography methods which are non-invasive and non-intrusive system, consisting of emitter with detectors. This research are conducted in order to analyze and proved the capability of laser with Charge Coupled Device in an optical tomography system for measuring object diameter that exist in crystal clear water. Experiments in detecting and capturing static solid rod in crystal clear water are conducted using this hardware and software development. T-test and Analysis of Variance (ANOVA) were used to analyze and validate the experiments data with the help of Minitab 16 software. This software helped to solve the statistical calculation and analyses. As a conclusion, this research has successfully developed an optical tomography system that capable to measure the diameter of solid rod in non-flowing crystal clear water. The performance of the system are validated by statistical analyses results.

Keywords: Statistical analyses; measurement; T-Test; Analysis of Variance (ANOVA); Minitab 16 software; optical tomography system

INTRODUCTION

Multiphase detectors become an important instrumentation in industries for the purpose of monitoring and analysis of objects behavior in process system. Multiphase flow consists of two or three phases in one flow system. Gas, liquid and solid have different physical properties and move in different velocities. Process industries applications can be dividing into four group of multiphase flow as listed below.

i. Solid and gas

ii. Solid and liquid

iii. Gas and liquid

iv. Solid, liquid and gas

Sediment flow is one of the two phase flow application system to monitor the mixing of solid and liquid. This sediment flow system usually applied for monitoring and controlling the rivers contents especially to avoid pollution occurred. Controlling rivers quality is important for good health environment. Rivers that near with industries especially need to apply this sediment flow measurement system to control the quality of water. This system is focus on mass and momentum exchange between the sedimentary and carrier fluid (Michaelides, 2006). Sedimentary particles usually contents molecules of heavy metals, organic and inorganic pollutants.

There are two types of two-phase flow detector, intrusive and invasive technique or non-intrusive and non-invasive technique. Intrusive technique is a technique that directly contact to the flow regime. There are variety types of flow regime technique such as impedance probe, electrical resistance probe, optical fiber probe, ultrasound technique, endoscope probe and hot film anemometry (Kumar, Dudukovic, & Toseland, 1996).

But, non-intrusive techniques become popular in process industries because this application does not disturb the flow of liquid and can give more accurate data measurement. The examples of non-intrusive technique are; pressure transducer, visualization technique, Gamma-ray density gauge, laser technique, X-Ray technique, Positron Emission Tomography (PET), Magnetic Resonance Tomography (MRI), Computed Tomography (CT), Electrical Capacitance Tomography (ECT) and Optical Tomography (Yang, Du, & Fan, 2007) (Kumar, Dudukovic, & Toseland, 1996).

Tomography method has been used since 1950 in medical fields and being spread into industry by 1990 (M.S.Beck, 1996). Tomography system is suitable to apply for non-invasive and non-intrusive monitoring system, especially for the industries that deal with the multiphase flow. Optical tomography (OPT) is the best approach because this method consists of hard field sensors (Rahim, Optical Tomography System: Principles, Technique and Applications, 2011) where the sensor does not depend on the changes of conductivity or permittivity of subjects that are being analyzed. OPT system provide a good spatial resolution where it can capture a very detailed image without making the pixels visible. OPT also provides a high speed data capturing system and it is suitable for online monitoring system applications (Spring, Fellers, & Davidson, 2013).

The aim of this research project is to build an OPT system using the combination of Charge Coupled Device (CCD) linear sensor and laser diode with LabVIEW software to detect multiphase flow. Qualitative and quantitative analyses were done using the LabVIEW and Minitab software. Minitab software are used for statistical analysis, while, LabVIEW programming are developed to measure the object diameter and to produce a cross-sectional pipeline image for online data.

RESEARCH METHODOLOGY

All the experiments are conducted at room temperature between 25^{o}C to 33^{o}C and relative humidity is within 65% to 85%. Light scatterion and diffraction effect are minimal and it is ignored from the calculations. The luminosity for lasers is maintained at 0.3 Lux value for full non-flowing crystal clear water experiments. The laser Lux values are measured using UNI-T UT381 Lux meter. Figure 1 shows the mechanical diagram of the suggested OPT system using CCD linear sensors and laser diodes.

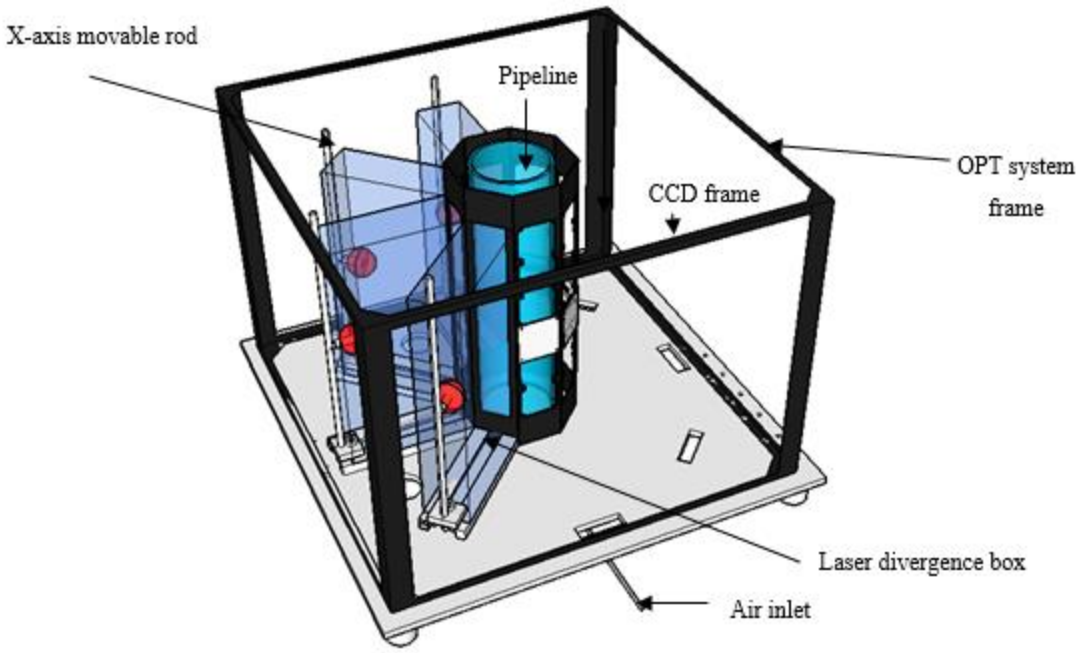


Figure 1: Mechanical diagram of OPT and pipeline system (Jamaludin, J & Rahim, R.A, 2016).

Meanwhile, Figure 2 shows the illustration of cross-sectional image of pipeline and OPT system (upper and lower plane) from side view.

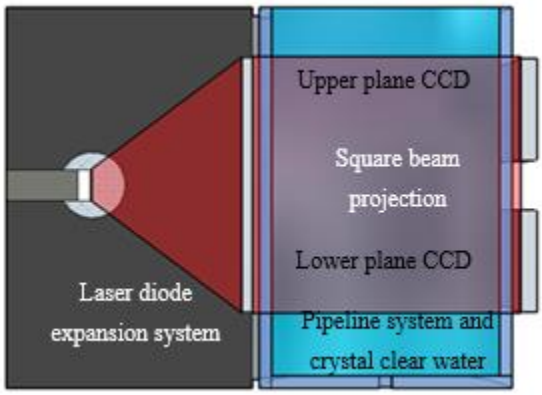


Figure 2: Illustration of cross-sectional image of pipeline and OPT system from side view

Experiments that involved in this research study are to validate the offline OPT system LabVIEW programming in detecting and measuring diameter of a solid rod. Results and discussions in this research are focusing on analyzing the capability of online CCD OPT system in capturing solid rod diameter in static crystal clear water. T-test and ANOVA were applied to analyze and validate the experiments data with the help of Minitab 16 software to do the statistical calculation.

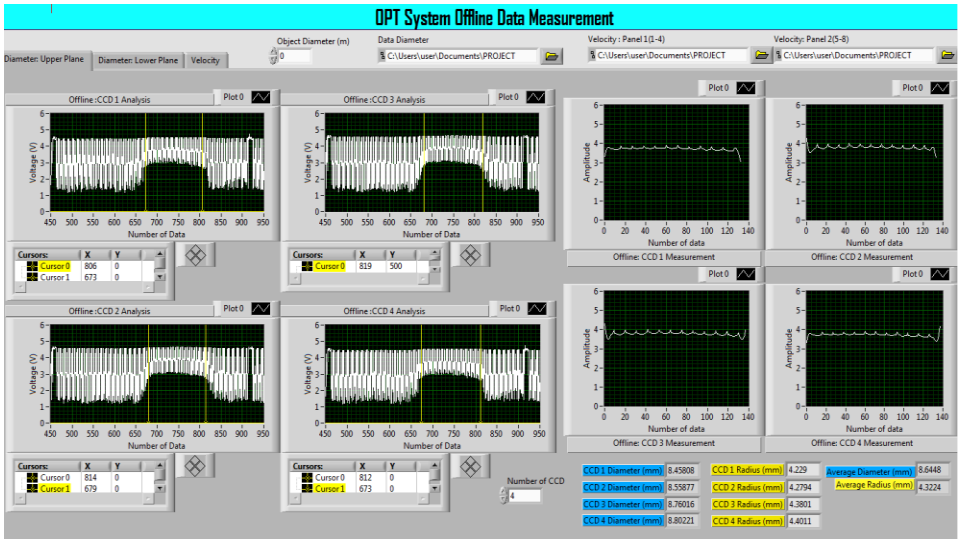
RESULTS

A series of experiments were conducted to evaluate the capability of this system in capturing and measuring objects diameter exist in crystal clear water. The objects that involved is solid rod.

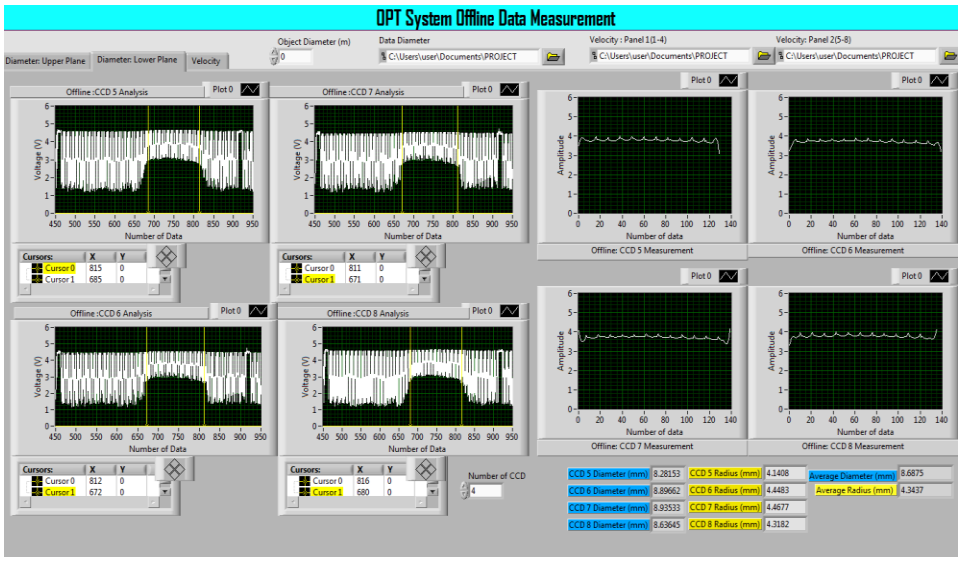
In order to have a targeted diameter value, the objects must first be measured by a Vernier Caliper. The known accuracy of the Vernier Caliper are at ±0.01 mm, while the OPT system accuracy are at ±0.0001 mm. The solid rod diameter is 8.54 mm based on the Venier Caliper reading as shown in Figure 3. This values will be used as the targeted mean value.

The totals of 50 diameter measurement of solid rod were observed using OPT system. Figure 4 below shows the example of LabVIEW programming front panel for solid rod diameter measurement. Figure 4 (a) are showing readings of CCD sensor 1 to 4 (upper plane) and Figure 4 (b) are showing readings of CCD sensor 5 to 8 (lower plane).

Figure 3: Diameter value for solid rod using Vernier calliper



(a)



(b)

Figure 4: Offline LabVIEW programming in diameter measurement of solid rod from (a) upper and (b) lower plane point of view (Jamaludin, J & Rahim, R.A, 2016).

DISCUSSION

T-test was used to validate the ability of OPT system in measuring diameter of static objects. Fifty data of the diameter measurement was obtained for this evaluation. The main objective of this analysis are to compare the measured diameter captured by the OPT system with the actual diameter value which measured by a Vernier Caliper. The two hypotheses involved in this analysis are;

H_0: Mean of experiment static object diameter data = Caliper measured static object diameter data

H_1: Mean of experiment static object diameter data ≠ Caliper measured static object diameter data

Minitab 16 software will generate a graphical chart for a visual understanding of each evaluation as shown in Figure 5. The charts are known as individual value plot graph where the red dotted presented the data samples distribution, x bar equal to samples mean, blue line are acceptable mean range and light red circle represent the targeted diameter value of solid rod. For P-value greater than 0.05, the null hypothesis

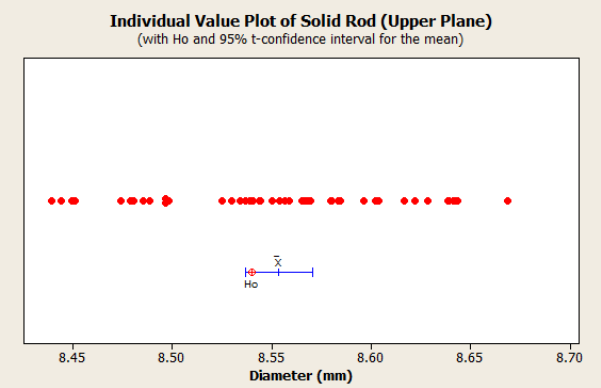
could not be rejected. In other words, the sample mean value is equal to the targeted mean value (Ross, Analysis of Variance, 2009).

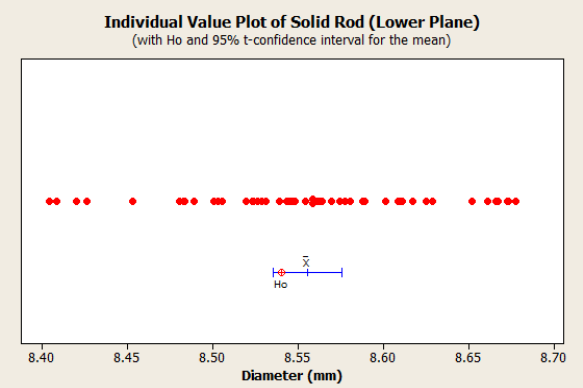
a) Summary of solid rod upper plane data measurement (T-test, P-value = 0.112)

P-value for OPT system upper plane data of solid rod are more than 0.05, thus we fail to reject the null hypothesis. This shows that based on the one sample T-test, the mean value of upper plane OPT system solid rod diameter measurements are not statistically different with the solid rod diameter value measured by Vernier Caliper.

b) Summary of solid rod lower plane data measurement (T-test P-value = 0.130)

P-value for OPT system lower plane data of solid rod are more than 0.05, thus we fail to reject the null hypothesis. This shows that based on the one sample T-test, the mean value of lower plane OPT system solid rod diameter measurements not statistically different with the solid rod diameter value measured by Vernier Caliper.





One-Sample T: Solid Rod (Upper Plane)

Test of mu = 8.54 vs not = 8.54

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Upper Plane	50	8.55359	0.05935	0.00839	(8.53672, 8.57046)	1.62	0.112

One-Sample T: Solid Rod (Lower Plane)

Test of mu = 8.54 vs not = 8.54

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Upper Plane	50	8.55532	0.07037	0.00995	(8.53532, 8.57532)	1.54	0.130

Figure 5: T-test graph result for solid rod upper and lower plane diameter measurement (Jamaludin, J & Rahim, R.A, 2016).

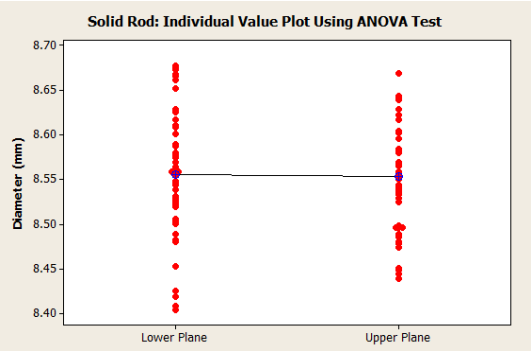
ANOVA test was applied to study the capability of upper and lower plane OPT system in measuring the same static object diameter. This statistical analysis is to validate the variation of the solid rod diameter measured at the upper plane with the variation of the solid rod diameter measured at lower plane.

According to ANOVA test, the statistical significant value or P-value below than 0.05 mean that the null hypotheses must be rejected and accepted the alternative hypothesis (Ross, Analysis of Variance, 2009). The hypotheses for upper and lower plane static object diameter measurement are as below;

H_0: Mean of upper plane single static object diameter measurements = Mean of lower plane single static object diameter measurements

H_1: Mean of upper plane single static object diameter measurements plane ≠ Mean of lower plane single static object diameter measurements

Figure 6 show the ANOVA test results for solid rod samples. In Figures 6, it shows the individual value plot for solid rod upper and lower plane samples data measurement. The P-value obtained from the ANOVA test was equal to 0.895 which greater than 0.05. The mean values of solid rod observed by upper plane are equal to 8.5536 mm and lower plane are equal to 8.5553 mm with standard deviation equal to 0.0594 and 0.0704. At 95% confident interval, both groups mean and standard deviations are overlapped between one and another. This lead to the acceptation to null hypothesis for this solid rod experiment. It is concluded that both upper and lower plane OPT system are indeed measuring the same solid rod.



Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
Lower Plane	50	8.5553	0.0704	(----------*----------)
Upper Plane	50	8.5536	0.0594	(----------*----------)
				8.540 8.550 8.560 8.570

Pooled StDev = 0.0651

One-way ANOVA: Solid Rod: Upper Plane, Lower Plane

Source	DF	SS	MS	F	P
Factor	1	0.00007	0.00007	0.02	0.895

Error 98 0.41525 0.00424

Total 99 0.41533

S = 0.06509 R-Sq = 0.02% R-Sq(adj) = 0.00%

Figure 6: ANOVA test graph results and data summarization for solid rod (Jamaludin, J & Rahim, R.A, 2016).

CONCLUSION

The OPT system developed is a non-intrusive and non-invasive technique for two-phase flow measurement. This system is safe and it does not pollute the environment with any chemicals or hazardous radiation. Since no probes or sensors need to be fitted in the liquid medium, this system offers more accurate data due to the absence of any interference of the OPT system in the pipeline processes.

Also, it is concluded that the OPT system consisting of laser and CCD as it transmitter and receiver, with LabVIEW as the programming platform are proven to have the capabilities in measuring objects diameter. In single static object experiments, it is determined that OPT system are indeed measuring the diameters of solid rod. Based on the T-test and ANOVA test results, it is proved that CCD linear sensor OPT system are reliable system in measuring objects diameter in static crystal clear water.

ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Higher Education and University Sains Islam Malaysia for funding the study. Very special thanks go to Universiti Teknologi Malaysia and PROTOM research group for their generous support and cooperation.

REFERENCES

Jamaludin, J & Rahim, R.A (2016). Online Optical Tomography System for Detecting and Measuring the Diameters of Solid and Transparent Objects. *IEEE Sensors Journal*, 16(16), 6175-6183.

Kumar, S., Dudukovic, M., & Toseland, B. (1996). Non Invasive Monitoring of Multiphase Flows. ElsevierScience.

M.S.Beck, R. (1996). Process Tomography : A European Innovation and Its Applications. *Meas. Sci. Technology., 7*, 215-224.

Michaelides, E. (2006). Particles, bubbles & drops: their motion, heat and mass transfer. World Scientific.

Rahim, R. A. (2011). *Optical Tomography: Principals, Technique and Applications.* Malaysia: Penerbit UTM.

Ross, S. M. (2009). Analysis of Variance. In *Probability and Statistic for Engineers and Scientist* (pp. 441-473). Burlington, USA: Elsevier Academic Press.

Spring, K. R., Fellers, T. J., & Davidson, M. W. (2013). *Nikon: The Source for Microscopy Education*, from https://www.microscopyu.com/articles/digitalimaging/ccdintro.html

Yang, G., Du, B., & Fan, L. (2007). A Review- Bubble Formation and Dynamics in Gas, Liquid, and Solid Fluidization. Chemical Engineering Science, 62, 2-27

CHAPTER 5

New Weighting Method for Robust Heteroscedasticity Consistent Covariance Matrix Estimator in Linear Regression

Abstract

The presence of high leverage points (HLPs) and heteroscedasticity are very common in empirical analyses using regression model. Weighted least squares are usually used to remedy the problem of heteroscedasticity if the heteroscedastic error structures are known. Heteroscedasticity consistent covariance matrix (HCCM) estimator is an alternative method in the case of unknown errors structure to remedy both the effect of leverage points and heteroscedasticity. However, the HCCM suffers tremendous effect due to the effect of masking and swamping. We proposed a HCCM based on modified generalized studentized residuals (MGt) based on DRGP(ISE). The results obtained from real data sets indicate that MGt weighting method outperformed the existing weighting method.

Keywords: *ordinary least squares, robust HCCM estimator, weighted least squares, high leverage points.*

1.0 Introduction

A linear regression model is normally analyzed by ordinary least squares (OLS) method. The homoscedasticity (equal variances of the errors) assumption is often violated in most empirical analyses which lead to heteroscedastic errors (unequal variances of the errors). In that case OLS provides inefficient parameter estimates and the inference becomes unreliable due to the inconsistency of the variance-covariance matrix estimator.

The most widely used estimation strategy for a heteroscedasticity of unknown form is to perform OLS estimation, and then employ a heteroscedasticity consistent covariance matrix (HCCM) estimator denoted by HC0 (see White, 1980). It is consistent under both homoscedasticity and heteroscedasticity of unknown form. MacKinnon and White (1985) proposed another HCCM estimator namely the HC1 and HC2. Davidson and MacKinnon (1993) modified HC2 and named it HC3 which is closely approximated to Jackknife estimator. Cribari-Neto (2004) proposed another HCCM estimator and called it HC4 where he adjusted the residuals by a leverage factor. Cribari-Neto et al. (2007) then proposed HC5 estimator, whereby they modified the exponent used in HC4 in order to consider the effect of maximal leverage.

It is important to note that HCCM estimators are constructed using OLS residuals vector. In the presence of high leverage points (HLPs), the coefficient estimates and residuals are biased. As a consequence, the inference becomes misleading. Furno (1996) proposed robust heteroscedasticity consistent covariance matrix (RHCCM) in order to reduce the biased caused by leverage points. He employed residuals of weighted least squares (WLS) regression where the weights are determined by the leverage measures (hat matrix) of the different observations. The shortcoming of Furno's method is that, in the presence of HLPs, the variances tend to be large resulting to unreliable parameter estimates which is due to the effect of swamping and masking of HLPs. The main reason for this weakness is the used of hat matrix in determining the weight of the RHCCM

algorithms of Furno (1996). It is evident that hat matrix is not very successful in detecting HLPs (Habshah et al. 2009). Consequently, less efficient estimates are obtained by employing unreliable method of detecting HLPs. His work has motivated us to use weight function based on more reliable diagnostic measure for the identification of HLPs. We proposed a new robust weighting method based on HLPs detection measures termed FMGt-DRGP The weights determined by FMGt-DRGP are expected to successfully down weight all bad HLPs.

2.0 Methodology

2.1 Heteroscedasticity Consistent Covariance Matrix (HCCM) Estimators

The regression model given by:

$$y = X\beta + \varepsilon \quad (1)$$

where, y is an $n \times 1$ vector of responses, X is an $n \times p$ matrix of independent variables, β is a vector of regression parameters, and ε is the n-vector of random errors. For heteroscedasticity the errors are such that $E(\varepsilon_i) = 0$, $var(\varepsilon_i) = \sigma_i^2$ for $i = 1, \dots, n$ and, $E(\varepsilon_i\, \varepsilon_s) = 0$ for all $i \neq s$. Covariance matrix of ε is given as $\Phi = \mathrm{diag}\{\sigma_i^2\}$. The ordinary least squares (OLS) estimator of β is $\hat{\beta} = (X'X)^{-1}X'y$ which is unbiased, with the covariance matrix given by:

$$cov(\hat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1} \quad (2)$$

However, under homoscedasticity $\sigma_i^2 = \sigma^2$ which implies $\Phi = \sigma^2 I_n$, where I_n is an $n \times n$ identity matrix. The covariance matrix $cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$ is estimated by $\hat{\sigma}^2(X'X)^{-1}$ (which is inconsistent and biased under heteroscedasticity) and $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/n - p$, $\hat{\varepsilon} = (I_n - H)y$, where H is an idempotent and symmetric matrix known as hat matrix or leverage matrix or weight matrix as named by different authors. The hat matrix (H) is defined as $H = X(X'X)^{-1}X'$, and it plays great role in determining the HLPs in regression model. The diagonal elements $\mathrm{h_i} = x_i(x'x)^{-1}x_i'$ for $i = 1, \dots, n$ of the hat matrix are the values for leverage of the i[th] observations.

White (1980) proposed the most popular HCCM estimator known as HC0 where he replaced the σ_i^2 with $\hat{\varepsilon}_i^2$ in covariance matrix of $\hat{\beta}$ as:

$$HC0 = (X'X)^{-1}X'\widehat{\Phi}_0 X(X'X)^{-1} \quad (3)$$

where, $\widehat{\Phi}_0 = \mathrm{diag}\{\hat{\varepsilon}_i^2\}$. HC0, HC1, HC2, and HC3 are generally biased for small sample size (see Furno 1997; Lima et al. 2009; Hausman and Palmer 2011). This research will focus on HC5.

Cribari-Neto et al. (2007) modify of the exponent of HC4 in order to control the level of maximal leverage and named it HC5 defined as:

$$HC5 = (X'X)^{-1}X'\widehat{\Phi}_5 X(X'X)^{-1} \quad (4)$$

where, $\widehat{\Phi}_5 = \mathrm{diag}\left\{\frac{\hat{\varepsilon}_i^2}{\sqrt{(1-\mathrm{h}_i)^{\alpha_\mathrm{i}}}}\right\}$ for $i = 1, \dots, n$ with $\alpha_\mathrm{i} = \min\left\{\frac{\mathrm{h_i}}{\mathrm{h}}, \max\left\{4, \frac{\mathrm{kh_{max}}}{\mathrm{h}}\right\}\right\}$, which determine how much the *i[th]* squared residual should be inflated, given by the ratio between $\mathrm{h_{max}}$ (maximal leverage) and *h* (mean leverage value of $\mathrm{h_i}$'s). when $\frac{\mathrm{h_i}}{\mathrm{h}} \leq 4$ it

follows that $\alpha_i = \frac{\mathrm{h_i}}{\mathrm{h}}$. Also, since $0 < 1 - \mathrm{h_i} < 1$ and $\alpha_i > 0$, it similarly follows that $0 < (1 - \mathrm{h_i})^{\alpha_i} < 1$ and k is a constant ranges between $0 < \mathrm{k} < 1$ and was suggested to be 0.7 by Cribari-Neto et al. (2007).

2.2 Robust HCCM Estimators

The problems of heteroscedasticity and high leverage points was addressed by Furno (1996) in order to reduce the bias caused by the effect of leverage points in the presence of heteroscedasticity. He suggested using WLS regression residuals instead of OLS residuals used by White (1980) in HCCM estimator. The weight is based on the hat matrix ($\mathrm{h_i}$) and the robust (weighted) version of HC0 is defined as:

$$HC0_W = (X'WX)^{-1}X'W\widehat{\Phi}_{0\mathrm{w}}WX(X'WX)^{-1} \quad (5)$$

where, W is an $n \times n$ diagonal matrix with,

$$w_i = min(1,\ c/\mathrm{h}_i), \quad (6)$$

and *c* is the cutoff point, $c = 1.5p/n$, p being the number of parameters in a model including the intercept and n is the sample size, $\widehat{\Phi}_{0\mathrm{w}} = \mathrm{diag}\ \{\tilde{\varepsilon}_i^2\}$ with $\tilde{\varepsilon}_i$ being the *i*th residuals from weighted least squares. Note that, non-leveraged observations are weighted by 1 and leveraged observations are weighted by ($c/\mathrm{h_i}$) to reduce their intensity and w_i is considered as the weight in this weighted least squares (WLS) regression, so that the WLS estimator of β is:

$$\tilde{\beta} = (X'WX)^{-1}X'Wy. \quad (7)$$

The robust HCCM estimator for the HC5 based on Furno's weighting method is defined as:

$$HC5_W = (X'WX)^{-1}X'W\widehat{\Phi}_{5\mathrm{w}}WX(X'WX)^{-1} \quad (8)$$

where, $\widehat{\Phi}_{5\mathrm{w}} = \mathrm{diag}\left\{\frac{\tilde{\varepsilon}_i^2}{\sqrt{(1-\mathrm{h}_i^*)^{\alpha_i^*}}}\right\}$ for $i = 1, \ldots, n$ with $\alpha_i^* = \min\left\{\frac{\mathrm{h}_i^*}{\mathrm{h}^*}, \max\left\{4, \frac{\mathrm{kh}^*_{\max}}{\mathrm{h}^*}\right\}\right\}$, the *i*th diagonal elements of the weighted hat matrix $H_w = \sqrt{W}X(X'WX)^{-1}X'\sqrt{W}$. In this paper the Furno's WLS for RHCCM estimation method is denoted by $\mathrm{WLS_F}$.

2.3 New proposed Robust HCCM estimator

In this study, we employed the idea of Furno's RHCCM estimation on new weighting method based on modified generalized studentized residuals (MGt) and diagnostic robust generalized potential based on index set equality (DRGP (ISE)) (Lim and Habshah 2016) in order to detect good and bad HLPs. The DRGP(ISE) consist of two steps, whereby in the first step, the suspected HLPs are determined using RMD based on ISE. The suspected HLPs will be placed in the 'D' set and the remaining in the 'R' set. The generalized potential ($\hat{p}_i$) is employed in the second step to check all the suspected HLPs, those possess a low leverage point will be put back to the 'R' group. This technique continued until all points of the 'D' group has been checked to confirm whether they can be referred as HLPs. The generalized potential is defined as follows:

$$\hat{p_i} = \begin{cases} h_i^{(-D)} & \text{for } i \in D \\ \dfrac{h_i^{(-D)}}{1-h_i^{(-D)}} & \text{for } i \in R \end{cases} \quad (9)$$

The cut-off point for DRGP is given by,

$$C_{DRGP} = median\,(\hat{p}_i) + 3\,Q_n\,(\hat{p}_i) \quad (10)$$

Q_n is employed to improve the accuracy of the identification of HLPs. $Q_n = c\{|x_i - xj\,|; < j\}_{(k)}$ is a pair wise order statistic for all distance proposed by Rousseeuw and Croux (1993) where $k = {}^{h}C_2 \approx {}^{h}C_2/4$ and $h = [n/2] + 1$. They make used of c = 2.2219, as this value will provides Q_n a consistent estimator for gaussian data. If some identified $\hat{p}_i$ did not exceed cdi then, the case with the least $\hat{p}_i$ will be returned to the estimation subset for re-computation of $\hat{p}_i$. The values of generalized potential based on final 'D' set is the DRGP(ISE) represented by $\hat{p}_i$ and the 'D' points will be declared as HLPs. The modified generalized studentized residuals (MGt) (Mohammed and Habshah 2015) is given by,

$$\mathrm{MGt}_i = \begin{cases} \dfrac{\hat{\varepsilon}_{i(R^*)}}{\hat{\sigma}_{R^*-i}\sqrt{1-h^{***}_{ii(R^*)}}}, & \text{for } i \in R^* \\ \dfrac{\hat{\varepsilon}_{i(R^*)}}{\hat{\sigma}_{R^*}\sqrt{1+h^{***}_{ii(R^*)}}}, & \text{for } i \notin R^* \end{cases} \quad (11)$$

where $\hat{\varepsilon}_{i(R^*)}, \hat{\sigma}_{(R^*)}$ are the OLS residuals and residuals standard error for remaining set *R*, respectively. The observations are called influential observation when their values of MGti greater than its cut-off point (C_{MGti}). The C_{MGti} is calculated as follows:

$$C_{MGTi} = median(MGti) + cMAD(MGti) \quad (12)$$

To classify HLPs we plot MGt versus DRGP(ISE) and follows the procedure given by Mohammed and Habshah (2015) of classification of HLPs.

i. Regular observation (RO): An observation is declared as regular observation if $|MGti| \leq C_{MGti}$ and $|DRGPi| \leq C_{DRGPi}$
ii. Vertical outlying observation (VO): Any observation is declared as VO if $|MGti| > C_{MGti}$ and $|DRGPi| \leq C_{DRGPi}$
iii. Good leverage observation (GLO): Any observation is declared GLO if $|MGti| \leq C_{MGti}$ and $|DRGPi| > C_{DRGPi}$
iv. Bad leverage observation (BLO): Any observation is declared BLO if $|MGti| > C_{MGti}$ and $|DRGPi| > C_{DRGPi}$

So, we down weight only VO and BLO and employed RHCCM estimation methods discussed in Section (2) to obtain the RHCCM estimator based on MGt-DRGP(ISE) weighting method denoted by WLS_{FMGt}.

3.0 Method Applications

3.1 Data Used

We employed education expenditure data used by Chatterjee and Hadi (2006) that represents the relationship between per capita income on education project for 1975 and three independent variables namely, per capita income in 1973 (x_1), number of residents per thousands under 18years of age (x_2) and number of residents per thousands under 18years of age in 1974 (x_3). The new proposed method (WLS_{FMGt}) and existing methods (OLS and WLS_F) were applied to the data.

4.0 Results and Discussions

Figure 1(a) – 1(c) shows the plot of residuals vs fitted values of OLS, WLS_F and WLS_{FMGt} respectively, which indicate the presence of heteroscedasticity due the systematic pattern produce by the variances of the error terms. Figure 1(d) shows the classification of observations where the 49th observation is declared as BIO, 3rd , 4th, 5th and 42nd were declared as GIO, the rest of the observations are RO.

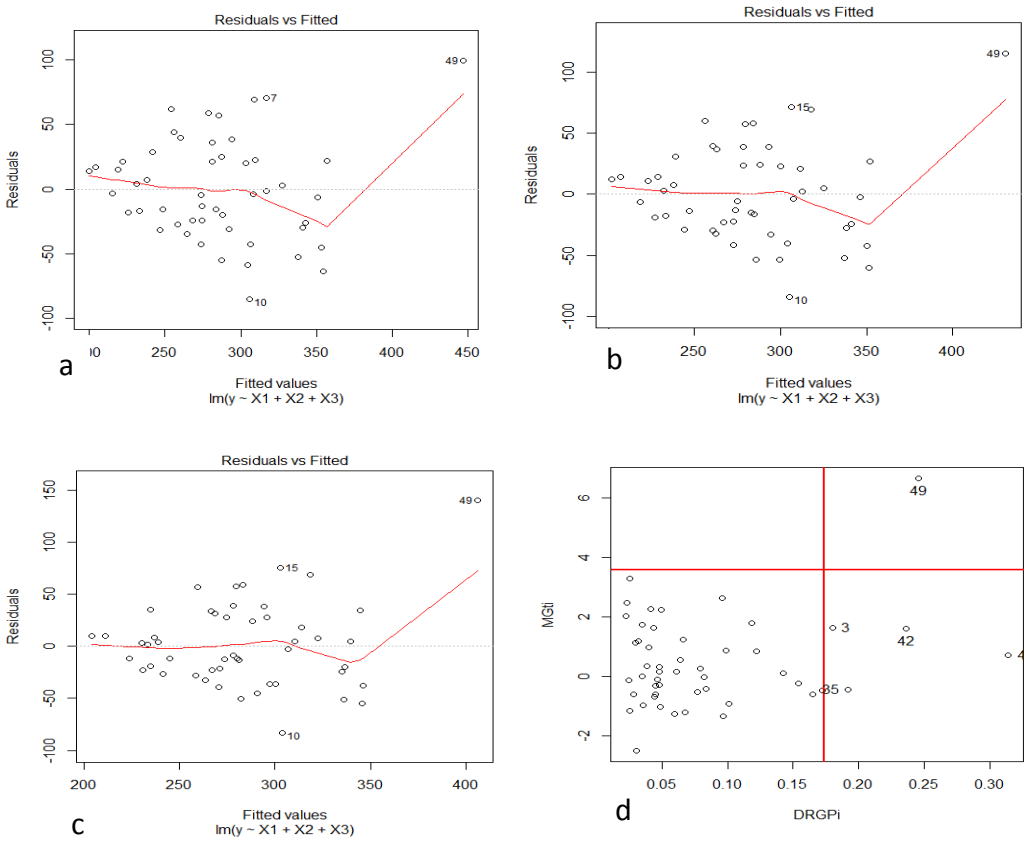


Figure 1. Plot of OLS residuals versus fitted values (a – c) and MGti versus DRGPi (d) for education expenditure data

Table 1: Regression estimates for the education expenditure data set**.**

Estimator		Coeff. of Estimates	Standard Errorof Estimates	Standard Error of HC5
OLS	b_0	-556.5680	123.1953	0.0003
	b_1	0.0724	0.0116	0.0131
	b_2	1.5521	0.3147	0.0948
	b_3	-0.0043	0.0514	0.0764
WLS_F	b_0	-496.6444	128.6804	0.0002
	b_1	0.06811		
	b_2	1.4028	0.01193	
	b_3	0.0076	0.0121	
			0.3318	0.0939
			0.0525	0.0663
WLS_{FMGti}	b_0	-412.2189	128.9292	0.0002
	b_1	0.0599	0.0119	0.0098
	b_2	1.2079	0.3251	0.0809
	b_3	0.0325	0.0503	0.0507

We modified the data by introducing HLPs contamination, in which the 2nd, 27th and 40th observations were replaced by 1323, 817 and 1605 for x_2, x_1, x_3 respectively. Figures 2 shows the classification of the observations.

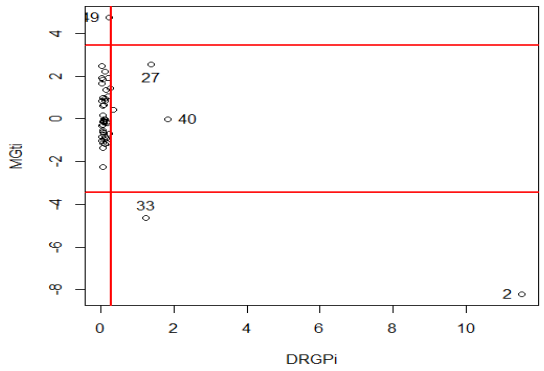


Figure 2. Plot of MGti versus DRGPi for modified education expenditure data.

Table 2: Regression estimates for the modified education expenditure data set**.**

Estimator		Coeff. of Estimates	Standard Error Estimates	Standard Error of HC5
OLS	b_0	114.6463	55.0662	0.0085
	b_1	0.0372	0.0100	0.9777
	b_2	-0.0314	0.0529	
	b_3	0.0130	0.0428	18.8106
				2.0328
WLS_F	b_0	47.7726	69.7880	0.0012
	b_1	0.0440	0.0113	0.0170
	b_2	0.0882	0.1310	0.1234
	b_3	0.0036	0.0547	0.0824
WLS_{FMGti}	b_0	113.1523	51.8968	0.0003
	b_1	0.0302	0.0089	0.0168
	b_2	0.0082	0.0307	0.1684
	b_3	0.0402	0.0374	0.0485

Tables 1 and 2 show the results of the education expenditure and modified education expenditure data sets. The results indicate that the new proposed WLS_{FMGt} outperformed the existing methods by providing a small standard errors of the estimates and HC5. It can be concluded that the WLS_{FMGt} is better and more efficient then WLS_F and OLS in the estimation of heteroscedastic model in the presence of HLPs in a data set.

5.0 Conclusion

In this paper, we proposed a new weighting method for robust heteroscedasticity consistent covariance matrix (HCCM) estimator. The robust HCCM estimator based on our weighting method provides an efficient parameter estimates for a heteroscedastic model when there exist high leverage points in a data set. The OLS method becomes inefficient and the Furno's WLS based on leverage weight function also not efficient enough to remedy the problem of heteroscedastic errors with unknown form and high leverage point. The WLS_{FMGt} was found to be the best method as it's provides the lowest standard errors of parameter estimates and HC5.

References

Chatterjee, S. & Hadi, A.S. (2006). *Regression Analysis by Example*, 4th Edition. New York: Wiley.

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

Cribari-Neto, F., Souza, T.C. & Vasconcellos, K.L.P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics-Theory and Methods 36*: 1877-1888
Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Furno, M. (1996). Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator. *Journal of Statistical Computation and Simulation 54*: 115-128.

Furno, M. (1997). A robust heteroskedasticity consistent covariance matrix estimator. *A Journal of Theoretical and Applied Statistics 30*: 201-219.

Habshah M., Norazan MR & Rahmatullah Imon AHM (2009) The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statatistics 36(5):*507–520

Hausman, J. & Palmer, C. (2011) Heteroscedasticity-robust inference in finite samples. *Economics Letters 116*: 232–235

Lim, H. A. & Habshah M. (2016) Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) for the identification of high leverage points in linear models. *Computational statistics*, *31*:859-877

Lima, V.M.C., Souza, T,C., Cribari-Neto, F. & Fernandes, G.B. (2009). Heteroskedasticity- robust inference in linear regressions. *Communications in Statistics-Simulation and Computation 39*: 194-206

Mohammed A. and Habshah M. (2015) A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problem in Engineering* ID 279472

MacKinnon, J.G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*: 305-325.

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

CHAPTER 6

On the Performance of Wild Bootstrap based on MM-GM6 Estimator in the Presence of Heteroscedastic Errors and High Leverage Points

Abstract

The violation of constancy of variance of error terms causes the problem of heteroscedasticity. OLS estimate is no longer efficient in the presence of heteroscedasticity in a data set, because the OLS estimates will be biased and inconsistent. As an alternative, a weighted residuals (wild bootstrap) may be used to remedy this problem. However, the weakness of wild bootstrap is that, in the presence of outliers the estimates of the standard errors become large. Therefore, a robust wild bootstrap is formulated based on MM-GM6 estimator so that the problems of both heteroscedasticty and outliers can be rectified. The results show that the proposed method performs better than the existing ones such as OLS, Wu, and Liu.

Keywords: Heteroscedasticity, outliers, wild bootstrap, high leverage point

Introduction

Bootstrap technique was proposed by Efron and Tibshirani [1].It is a statistical method that can replace theoretical formulation with extensive use of computer .This method does not depend on the distributional assumptions and is able to estimate the standard errors of parameter estimates without theoretical calculations. There are many papers which deal with bootstrap methods (see [2-4]).

Multiple regression analysis is a statistical technique used widely for modelling and analysing the relationship between one dependent variable and two or more independent variables. The standard linear regression model can be defined as:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \qquad (1)$$

Where $Y = (y_1, y_2, \dots, y_n)^T, X = (x_1, x_2, \dots, x_n)^T$,and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$. In equation (1), β is a k×1vector of unknown parameters, Y is an n×1 vector , X is an n×k data matrix of independent variables ,and ε is an n×1 vector of unobservable random errors such that $\varepsilon \sim NID(0, \sigma^2)$. The assumption of stability of variance $Var(\epsilon_i) = (\sigma I)$ is often violated. As a consequence, Wu [5] proposed a wild bootstrap that can be utilized to evaluate the standard errors which are asymptotically and accurate under non stability of variance. Liu [6] proposed another wild bootstrap method which is slightly different from Wu's method. Nevertheless, (Midi ,et.al) [7] stated that there is evidence that the Wild Bootstrap estimator suffer a huge set back in the presence of a few atypical observations that we often call outliers. Therefore, they incorporated the robust MM estimator. However, the MM-estimator does not have a bounded influence property. On the other hand, the GM6 estimator is robust in the X-direction.Thus, In this paper, we want to investigate an

alternative wild bootstrap in the wild bootstrap algorithm based on robust MM-GM6 estimator .

Wild bootstrap technique

Heteroscedasticity is a common problem in linear regression model, occurs when the variance of error terms are not stable. In this case, OLS estimator is no longer efficient. The fixed x bootstrapping the residual method is suggested by Efron and Tibshirani [1].This bootstrapping procedure is based on the ordinary least squares residuals that can be summarized as follows:

Step 1. Fit a model $y_i = f(x_i, \beta_{ols})$by the Ols method to the original sample of observations to get $\hat{\beta}_{ols}$and hence the fitted model is $y_i = f(x_i, \hat{\beta}_{ols})$

Step 2. Compute the OLS residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ and each residual $\hat{\varepsilon}_i$ has equal probability ,$\frac{1}{n}$.

Step 3. Draw a random sample $\varepsilon^*{}_1, \varepsilon^*{}_2, \dots, \varepsilon^*{}_n$ from $\hat{\varepsilon}_i$ with simple random sampling with replacement and attached to $\hat{y}_i$ to obtain Fixed-x bootstrap values $y^{*b}{}_i$ where $y^{*b}{}_i = f(x_i, \hat{\beta}_{ols}) + \varepsilon^{*b}{}_i$.

Step 4. Fit the OLS to the bootstrap value $y^{*b}{}_i$ on the Fixed-x to obtain $\hat{\beta}^{*b}{}_{ols}$

Step 5. Repeat Steps 3 and 4 for B times to get $\hat{\beta}^{*b1}{}_{ols}, \dots, \hat{\beta}^{*bB}{}_{ols}$ where B is the bootstrap replications.

This bootstrap is called $Boot_{ols}$ since it is based on the OLS method

Wu [5] slightly modified Step 3 of the OLS bootstrapping procedure to add the weight in the residual as follows.

$$y^{*b}{}_i = f(x_i, \hat{\beta}_{ols}) + \frac{t_i^* \hat{\varepsilon}_i}{\sqrt{1-h_{ii}}} \qquad (2)$$

Where, t_i's can be selected from a standard normal and h_{ii} is the ith leverage which represents the diagonal of $Hat\ marix = X(X'X)^{-1}X'$

Liu's bootstrap can be conducted by drawing random numbers t_i^* in the following way.

$$t_i^* = H_i D_i - E(H_i)E(D_i), i = 1,2, \dots, n\ and\ H_1, H_2, \dots H_n\ are\ iid\ normally$$

$distributed\ with\ mean\ \frac{1}{2}\left(\sqrt{\frac{17}{6}}+\sqrt{\frac{1}{6}}\right)\ and\ variance\frac{1}{2}$.

$As\ well\ as, \mathrm{D}_1, \mathrm{D}_2, \ldots \mathrm{D}_\mathrm{n} are\ iid\ normally\ distributed\ with\ mean\ \frac{1}{2}\left(\sqrt{\frac{17}{6}}-\sqrt{\frac{1}{6}}\right)$

$and\ variance\frac{1}{2}$.

Proposed Robust Wild Bootstrap Technique

The wild bootstrap is not resistant to outliers because its algorithm is based on the OLS residuals. Thus, in this paper we incorporate the MM- GM6 estimator in the wild bootstrap algorithm. The GM6-estimator aims to down weight outliers both X and Y coordinates to insure that HLPs get lower weights, while the MM estimator is robust to outliers in Y coordinate. The steps of MM-GM6 wild bootstrap can be summarized as follows:

Step 1. Fit the regression model $y_i = x_i\beta + \varepsilon_i$ by using MM estimator to the original data to obtain the robust parameters $\hat{\beta}_{MM}$, then $\hat{y}_i = x_i\hat{\beta}_{MM}$ and hence the fitted model is $\hat{y}_i = \mathrm{x_i}\hat{\beta}_{\mathrm{MM}}$

Step 2. Find the residuals of the MM estimate $\hat{\varepsilon}_i^{MM} = y_i - \hat{y}_i$. Then assign the weight of GM6 to each residual such that $w_i = \min\left(1, \frac{\chi^2{}_{0.95,p}}{MVE}\right)$, where MVE is the minimum-volume ellipsoid.

Step 3. The final weight residual of the MM estimate denoted by $\hat{\varepsilon}_i^{WMM}$ are formulated by multiplying the weight obtained in Step 2 with the residual of the MM estimates.

Step 4. Construct a bootstrap sample $(y^*{}_i, X)$, where

$$\mathrm{y^*{}_i} = \mathrm{x_i}\hat{\beta}_{\mathrm{MM}} + \mathrm{t_i^*} \times \min\left(1, \frac{\chi^2{}_{0.95,\mathrm{p}}}{\mathrm{MVE}}\right) \hat{\varepsilon}_i^{MM} \qquad (3)$$

and t_i^* is a random sample following Liu[6] procedure.

Step 5. The MM procedure is then applied to the bootstrap sample $(y^*{}_i, X)$ and the resultant estimate is denoted by $\hat{\beta}^{*R} = (X^TX)^{-1}X^Ty^*$.

Step 6. Repeat Steps 4 and 5 for B times, where B is the bootstrap replications.

Numerical Example

In this section, a numerical example is presented to assess the performance of the 4 methods, $\mathrm{boot_{ols}}$, $\mathrm{boot_{wu}}$, $\mathrm{boot_{liu}}$, and $\mathrm{boot_{GM6\text{-}MM}}$. A set of real data is used to test the efficiency of the methods. The Concrete Compressive Strength data is taken from Yeh [8]. Concrete compressive strength is the response variable and Age (day), Fine Aggregate, Coarse Aggregate, Super plasticizer, Water , Fly Ash, Blast Furnace Slag, and Cement) are the set of predictor variables. We purposely replaced 2 good

observations with outliers in the y directions, indicated by 1 and 1030 to create outliers in the data set, so that their effect on the parameter estimates can be investigated. The $boot_{ols}$, $boot_{wu}$, $boot_{liu}$, and $boot_{GM6\text{-}MM}$ were then applied to the data set.

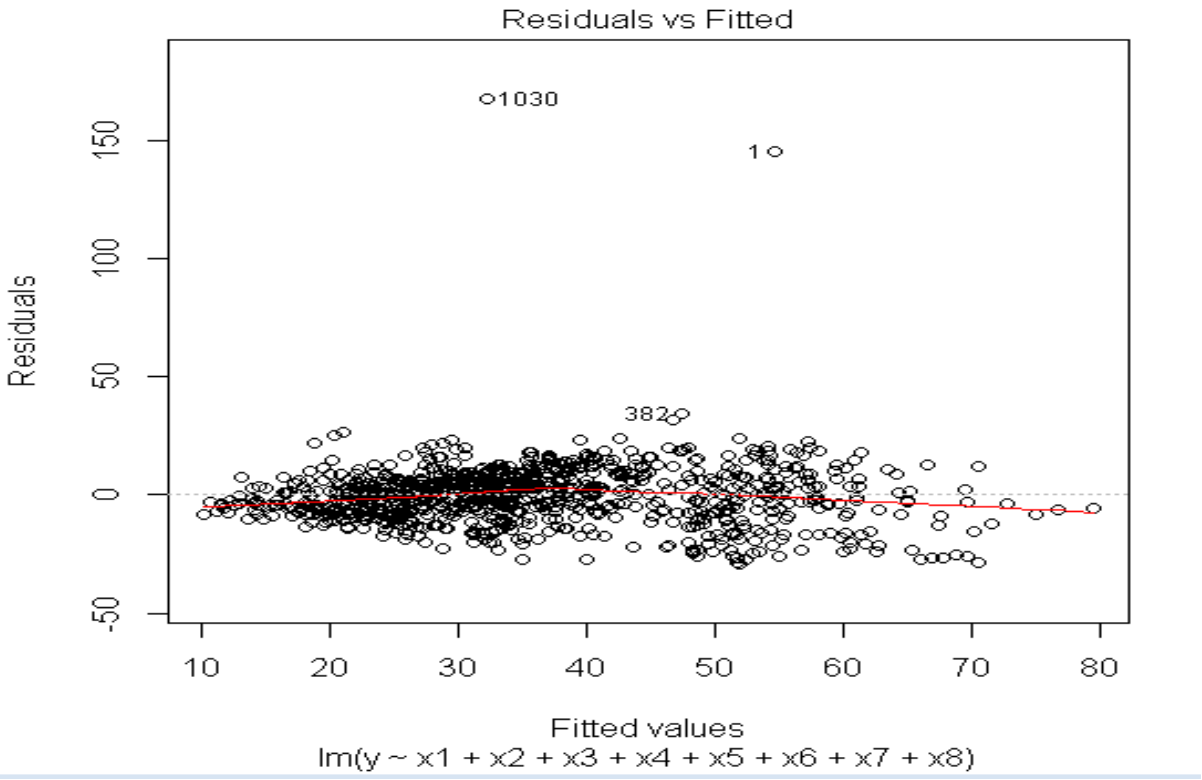


Figure 1: Residuals versus Fitted values plot of Concrete Compressive Strength data.

The residuals versus fitted values are plotted in Figure 1 that show a funnel shape suggesting a heterogeneous error variances for the data.

Table 1: Wild bootstrap standard errors of the parameters for the Concrete Compressive Strength data set.

Standard error (se)	$Boot_{ols}$	$Boot_{wu}$	$Boot_{liu}$	$Boot_{MM\text{-}GM6}$ (liu)
Intercept	31.8681	29.0002	24.5610	2.8731
Cement	0.0104	0.0095	0.0075	0.0009
Blast Furnace Slag	0.0123	0.0115	0.0095	0.0011
Fly Ash	0.0153	0.0144	0.0117	0.0014
Water	0.0487	0.0447	0.0378	0.0043

Super plasticizer	0.1162	0.0986	0.0906	0.0101
Coarse Aggregate	0.0112	0.0101	0.0089	0.0010
Fine Aggregate	0.0127	0.0117	0.0098	0.0012
Age	0.0066	0.0059	0.0052	0.0006

The standard errors of the preceding methods based on 500 bootstrap samples are exhibited in Table 1. It can be observed that our proposed method gives the best result evident by having the smallest standard errors of the parameter estimates, followed by $boot_{liu}$, $boot_{wu}$ and $boot_{ols}$.

Conclusion

This paper examines the performance of classical wild bootstrap techniques which were proposed by Wu [5] and Liu [6] in the presence of heteroscedasticity and outliers. The numerical results show that our proposed $boot_{MM\text{-}GM6}$ outperforms the existing methods when both outliers and heteroscedasticity are present in the data.

References

[1] Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics (pp. 569-593). Springer New York.

[2] B. Efron and R. Tibshirani, “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy,” Statistical Science, vol. 1, no. 1, pp. 54–77, 1986.

[3] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association*, *82*(397), 171-185.

[4] Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[5] .-F. J. Wu, “Jackknife, bootstrap and other resampling methods in regression analysis,” The Annals of Statistics, vol. 14, no. 4, pp. 1261–1350, 1986.

[6] R. Y. Liu, “Bootstrap procedures under some non-i.i.d. models,” The Annals of Statistics, vol. 16, no. 4,pp. 1696–1708, 1988.

[7] Rana, et. al. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. Mathematical Problems in Engineering, 2012.

[8] I.-C. Yeh, “Modeling of strength of high-performance concrete using artificial neural networks,”*Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, 1998.

CHAPTER 7

New Approach to Normalization Technique in K-Means Clustering Algorithm.

Abstract:

The K-Means algorithm first developed in 1960's, is a popular method in cluster algorithms. Data preprocessing methods commonly use raw data to make the data clean, noise free, and consistent. This prevents larger numbers out-weighing features having smaller numbers. Therefore, one of the weaknesses of decimal scaling (DS) approach is that, it has problem of overflow, this makes the approach not robust and prone to outliers. We therefore, introduced a new approach normalization techniques to enhance the K-Means algorithm. This is to remedy the problem of using decimal scaling approach, which has overflow weakness. Hence, the suggested approach is called new approach to decimal scaling (NADS). Furthermore, based on real life datasets, the performance of the suggested method is compared with the existing methods, which evidently indicates that the suggested method outperformed the existing methods with higher average maximum external validity measures, and lower computing time (in minutes). Consequently, the proposed method may be used as data preprocessing methods in distance-based clustering analysis.

Keywords: Normalization, K-Means, Clustering, Validity, Preprocessing.

1. Introduction

Clustering is often applied as the first step in data analysis. It functions as an assessment to discover natural clusters in datasets to identify theoretical patterns that live inside, without having any primary ideas on the features of data (Mohd et al., 2012). It is an unsupervised arrangement method by partitioning data into clusters with main objective of separation, where points in the same cluster are alike, and points belong to different clusters differ significantly, with regard to their attributes (Mohd et al., 2012), and (Suarez-Alvarez et al., 2012). However, it is known that data are taken as unlabeled and clustering is generally understood as the most important unsupervised learning assignment (Patel and Mehta, 2011), (Suarez-Alvarez et al., 2012).

The K-Means is the most widely used cluster algorithm which was first developed by Macqueen 1967, and the algorithm was later highly-developed and expanded by Lloyd, in 1982. The algorithm although is very easy and strong in clustering large data sets, the method suffers from some setbacks (Duwairi and Abu-Rahmeh, 2015). The number of clusters have to be known before hand when applying to most of the real world data sets (Rokach and Maimon, 2014). It has to undergo issues of random selection of initial cluster centers (centroids), which may be sensitive to the algorithm (Barakbah and Kiyoki, 2009). Nonetheless, the algorithm never achieve global optimum results (Rokach and Maimon,

2014). The K-Means algorithm repeatedly converge to a local minimum. The issue of local minimum is being established on the initial cluster centers. Also, the problem of exploratory global minimum is nondeterministic polynomial (NP) time-complete (Oyelade et al., 2010). Usually, K-means algorithm continually updates cluster centers until local minimum is achieved. It is observed that in literature, one of the weaknesses of K-Means clustering algorithm is when unnormalized dataset is used, often that the outcome performance may not reach global optimum (Han et al., 2011).

Data preprocessing methods commonly use raw data to make the data clean, noise free, and consistent (Patel and Mehta, 2011). Data normalization task is to standardize raw data by changing it into classified interval through linear transformation in order to produce good quality clusters and improve the accuracy of clustering algorithms. A normalized dataset is observed to produce better outcomes during the actual clustering process (Patel and Mehta, 2011). This prevents large numbers out weighing features having features with smaller numbers. The main aim is to equalize the magnitude and also prevent the much inconsistency in those features (Mohamad and Usman, 2013).

In this study we were motivated by a problem pointed out in Visalakshi and Thangavel, 2009, that up to date, there is no specific rule for normalizing the datasets, however, the researcher has open options to select whichever approach one wishes to apply. In addition, Wu et al., 2007, stated that the importance of normalizing validation measures has not been completely accepted. Also, it was stated by Aksoy and Haralick, 2001 that it is essential to take into consideration that distance measures like Euclidean distance should not be applied without preprocessing.

This research is organized as follows: Section 2 presents materials and methods; K-Means clustering algorithm, Decimal Scaling, and new approach to decimal scaling. Section 3 gives results and discussion. Section 4, some concluding remarks were given.

2. Materials and Methods

2.1.Conventional Methods

In the literature, there are a number of conventional techniques in normalization and standardization, but the most common methods are min-max, decimal scaling, and Z-score methods. However, in this paper, we are going to limit our study to decimal scaling normalization approach. Furthermore, we also would like to investigate the performance of K-Means clustering algorithm that evaluates dataset without normalization, which is common practice by practitioners.

2.1.1. K-Means Clustering Algorithm

The K-means clustering algorithm consist of four steps, which are iterates until convergence are achieved (Mohamad and Usman, 2013). The iteration will stop when the clusters produced are stable, it means that there are no more movement of objects crossing any group. The K-Means algorithms are listed by (Macqueen, 1967), (Lloyd, 1982), and (Shirkhorshidi et al., 2015) as follows:

The K-Means clustering algorithm is broadly used in data mining to group data with similar features together. Assume data points, the algorithm distributes them into ***k*** groups in three stages: (1) evaluate the distances between data points with each of ***k*** clusters and assign the data to the nearest cluster; (2) calculate the center of each cluster; (3) update the clusters repeatedly until the ***k*** clusters stabilize. The aim of the algorithm is to minimize the cost function. The cost function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \left\| x_i - c_j \right\|^2 \qquad (1)$$

Where, $\left\| x_i - c_j \right\|^2$ is an arbitrary distance measure between a data point x_i and the cluster center c_j is a sign of the distance of the *n* data points from their individual centers. The algorithm consists of the following steps (Khan, 2012):

Step 1: Initialize the centers at random;

Step 2: Assign data points to their respective clusters having the nearest mean;

Step 3: Compute new centers as means of the clusters assigned in step 2;

Step 4: Repeat steps 2 and 3 until the centers are stabilized.

This creates a partition of the objects into groups from which the value to be minimized can be calculated, but after data normalization as given in Equation 2, and 3.

2.1.2. Decimal Scaling (DS)

Let *j* be a numeric attribute with *n* observed values $v_1, v_2, ..., v_n$ (Han et al., 2011). It normalizes the dataset by moving the decimal point values of attribute *j*. The number of decimal points moved depends on the maximum absolute value of *j*. The value, v_i of *j* is normalized to v_i' and computed as in (Han etal., 2011):

$$v_i' = \frac{v_i}{10^j} \quad (2)$$

Where *j* is the smallest integer (the integer *j* is equal to the maximum numbers of digits; example, 986, *j* = 3).

2.1.3. New Approach to Decimal Scaling (NADS)

The new approach to decimal scaling is formulated following the ideas of (Zumel and Mount, 2013), but with a slight modifications where normalization is done by replacing the decimal point of values of feature *j* with that of *c + 1.* The number of decimal points moved depends on the maximum absolute value of the attributes (Mohamad and Usman, 2013). The new approach to decimal scaling is calculated using the ideas from Equation 2 with the introduction of *c + 1* to power 10 replacing the maximum absolute integer value with absolute real value using logarithm base 10 as follows:

$$v_i' = \frac{v_i}{10^{(c+1)}} \quad (3)$$

Where, $c = log_{10}\, max(x_i)$; if evaluated without adding 1 to *c,* all the variable values will be slightly greater than 1, which is out of bound for the upper range. Therefore, it is calculated based on the following conditions and rules:

Step 1: We first compute the largest absolute value in each row using logarithm base 10 and plus 1 each.

Step 2: Then, divide the original row value by 10 power of this computed value to obtain the normalized value. It has range of [0, 1].

However, it is important to mention that after the transformation of data by decimal scaling and the proposed method, the following steps are carried out to compare the performance of the proposed method and the existing methods:

Step 1. Perform the K-Means clustering (with unnormalized data).

Step 2. Then, perform the K-Means clustering with the classical and the proposed normalization method.

Step 3. Some external measures such as Purity, Fowlkes-Mallow Index, Rand Index, F-Measure Score, Jaccard Index, Recall, F-Measure (beta varied), Geometric Means, Precision, Specificity, Accuracy, Sensitivity, and the computing time (in minutes) are recorded.

2.2. Real Data Applications

In this section, the Iris, Hayes-Roth and Tae datasets are considered to verify the performance of our proposed method.

Iris dataset: The dataset contains 3 classes of 150 sample size each, where each class refers to a type of iris plant. It comprises the following attributes information: (1) sepal length in cm, (2) sepal width in cm, (3) petal length in cm, and (4) petal width in cm. The classes are listed as follows: (1) iris Setosa, (2) iris Verisiclor, and (3) iris Virginica (Bache and Lichman, 2013).

Hayes-Roth dataset: The contains 3 classes of 160 sample size each, with 4 attributes namely: (1) hobby, (2) age, (3) educational, and (4) marital status (Bache and Lichman, 2013).

Tae (Teaching Assistant Evaluation) dataset: The dataset contains 3 classes of 151 sample size each, with 5 attributes namely: (1) native, (2) instructor, (3) course, (4) semester, and (5) size (Bache and Lichman, 2013).

After data is transformed by DS, and NADS; the K-Means clustering algorithm is applied to the transformed data as well as the Conventional (not transformed data). The performance of the methods are evaluated based on the external validity measures such as: Purity, Fowlkes-Mallow Index, Rand Index, F-Measure Score, Jaccard Index, Recall, F-Measure (beta varied), Geometric Means, Precision, Specificity, Accuracy, Sensitivity, and the computing time (minutes) are recorded. Then, the average external validity measures and computing time (minutes) are computed under each distance functions, in order to ascertain a good normalization method that has average external validity measure closer to 1 or (1) at maximum, and minimum computational time.

3. Results and Discussion

Table 1: *Average External Validity Measures and Computing Time for Iris Dataset*

Distance Functions	Conventional	Decimal Scaling	NADS
Purity	0.9072	0.9072	0.9319
Fow. Mallow I.	0.9188	0.9174	0.9305
Rand Index	0.9206	0.9233	0.9413
F-Measure(Score)	0.9003	0.9052	0.9219
Jaccard Index	0.8933	0.8933	0.9308
Recall	0.9067	0.9069	0.9393
F-Measure(varied)	0.9016	0.9022	0.9145
Geometric Means	0.9045	0.9211	0.9291
Precision	0.9038	0.9182	0.9252

Specificity	0.9219	0.9347	0.9472
Accuracy	0.9137	0.9161	0.9291
Sensitivity	0.9067	0.9069	0.9393
Average	0.9083	0.9127	0.9317
Com. Time (minutes)	43	42	38

Table 2: *Average External Validity Measures and Computing Time for Hayes-Roth Dataset*

Distance Functions	Conventional	Decimal Scaling	NADS
Purity	0.4250	0.4309	0.4497
Fow. Mallow I.	0.4129	0.4256	0.4384
Rand Index	0.4375	0.4450	0.4603
F-Measure(Score)	0.4261	0.4280	0.4391
Jaccard Index	0.4132	0.4259	0.4395
Recall	0.4355	0.4393	0.4453
F-Measure(varied)	0.4152	0.4247	0.4300
Geometric Means	0.5363	0.5483	0.5561
Precision	0.4133	0.4141	0.4219
Specificity	0.6171	0.6188	0.6229
Accuracy	0.5037	0.5133	0.5290
Sensitivity	0.4355	0.4393	0.4453
Average	0.4559	0.4627	0.4731
Com. Time (minutes)	44	43	41

Table 3: *Average External Validity Measures and Computing Time for Tae Dataset*

Distance Functions	Conventional	Decimal Scaling	NADS
Purity	0.4845	0.4859	0.5203
Fow. Mallow I.	0.5036	0.5148	0.5282
Rand Index	0.4941	0.5032	0.5172
F-Measure(Score)	0.5060	0.5131	0.5210
Jaccard Index	0.4504	0.4613	0.4883
Recall	0.5167	0.5233	0.5310
F-Measure(varied)	0.5122	0.5244	0.5371
Geometric Means	0.5833	0.5867	0.5987
Precision	0.5172	0.5194	0.5264
Specificity	0.6836	0.6866	0.6965
Accuracy	0.6171	0.6271	0.6293

Sensitivity	0.5167	0.5233	0.5310
Average	0.5321	0.5390	0.5520
Com. Time (minutes)	43	42	37

Tables 1,2, and 3 present the average performance of external validity measures and computing time under each distance functions. However, it is important to mention some effect of our suggested method on clustering before detail discussion on the tables mentioned above. The importance of suggested normalization method like NADS is to eliminate redundant data and ensures that good quality clusters are evaluated which can enhance the efficiency of clustering algorithm. The suggested method is evaluated before clustering is specifically needed for distance metric like the Euclidean distance that is sensitive to variations within the magnitudes from the attributes. It also prevents outweighing features having a large number over features with smaller numbers. Therefore, clustering with this suggested method normalized matrices yield tighter (compact) and hence better clusters.

It can be clearly observed from the three tables above the proposed NADS is the best method as it has external validity measures closest to 1 most especially in Table 1 and also has the lowest computational time. In Table 1, the average external validity measures for NADS (0.9317), Decimal Scaling (0.9127), Conventional method (0.9083), while computational time in minutes for NADS (38), Decimal Scaling (42), and Conventional method (43), respectively. Table 2, the average external validity measures for NADS (0.4731), Decimal Scaling (0.4627), and Conventional method (0.4559), while computational time in minutes for NADS (41), Decimal Scaling (43), and Conventional method (44), respectively. Table 3, the average external validity measures for NADS (0.5520), Decimal Scaling (0.5390), and Conventional method (0.5321), while computational time in minutes for NADS (37), Decimal Scaling (42), and Conventional method (43), respectively.

This indicates that the performance of NADS is more accurate and efficient compared to the existing methods. It is evidently shown and based on this experiment that the conventional method without transformation give the poor results. Therefore, based on this real life data results, the proposed method may be used especially in distance-based data preprocessing clustering analysis methods in many sectors of real life situations.

4. Conclusion

In this research, we proposed normalization approach to overcome attributes with initially large range from overweighting attributes with initially smaller ranges. The new normalization approach is called new approach to decimal-scaling (NADS).

To investigate the performance of our proposed approach, real life datasets are considered. The results indicate that the conventional K-Means without normalization has the least performance. This is due to the fact that distance measures like Euclidean distance, should not be applied without normalization of datasets. Although, the proposed method has good performance; evidently, by achieving nearly maximum points in the average external validity measures and clustering the object points to almost all their cluster centers and recorded lower computing time; but, it has failed to perform very well in the integer-based datasets. From the results, it can be concluded that the NADS approach is better in the data preprocessing methods; which down weight the magnitudes of larger values.

References

Aksoy, S. and Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters,* 22(5), 563-582.

Bache, K. and Lichman, M. (2013). UCI machine learning repository. *University of California, School of Information and Computer Science, Irvine, ca. Retrieved from the World Wide Web October, 27: 2014*

Barakbah, A. R. and Kiyoki, Y. (2009). A pillar algorithm for K-means optimization by distance maximization for initial centroid designation. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on,* pages 61-68.

Duwairi, R. and Abu-Rahmeh, M. (2015). A novel approach for initializing the spherical K-means clustering algorithm. *Simulation Modeling Practice and Theory, 54,* 49-63.

Han, J., Pei, J. and Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Khan, F. (2012). An initial seed selection algorithm for k-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application. *Applied Soft Computing, 12(11):* 3698-3700.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory, 28(2),* 129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* 1(14), 281-297.

Mohamad, I.and Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology,* 6(17), 3299-3303.

Mohd, W. M. B. W., Beg, A. H., Herawan, T. and Rabbi, K. F. (2012). An improved parameter less data clustering technique based on maximum distance of data and Lioyd k-means algorithm. *Procedia Technology, 1,* 367-371.

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. *arXivpreprint arXiv: 1002.2425.*

Patel, V. R. and Mehta, R. G. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. *IJCSI International Journal of Computer Science Issues,* 8(5), 331-336.

Rokach, L. and Maimon, O. (2014). *Data mining with decision trees: theory and applications. World scientific.*

Shirkhorshidi, A. S., Aghabozorgi, S. and Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one, 10(12), e0144059.*

Suarez-Alvarez, M. M., Pham, D. T., Prostov, M. Y., and Prostov, Y. I. (2012l). Statistical approach to normalization of feature vectors and clustering of mixed datasets. *In Proc. R. Soc. A (p. rspa20110704). The Royal Society.*

Visalakshi, N. K. and Thangavel, K. (2009). Impact of normalization in distributed k-meansclustering. *International Journal of Soft Computing,* 4(4), 168-172.

Wu, J., Xiong, H., Chen, J. and Zhou, W. (2007). A generalization of proximity functions for k-means. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on,* pages 361-370.

Zumel, N. and Mount, J. (2013). Log Transformations for Skewed and Wide Distributions. *Retrieved March, 18,* 2014.

CHAPTER 8

The Effect of High Leverage Points on Collinearity Diagnostic in Logistic Regression Model

Abstract

The problem of collinearity among regressors and weighted regressors in the observed Fisher information matrix of maximum likelihood is examined. Both ill-conditioned situations inflate the estimated variances and regression coefficients. There is evident that the pattern of collinearity can change substantially in the presence of high leverage points. The existence of high leverage points creates the ambiguous understanding of collinear regressors and it is affecting the interpretation of model parameter estimation. In this article, we investigate the behavioral of high leverage points in collinear data. While the previous studies focusing on a common result that collinearity increase variance estimates, we found that under a certain condition, collinearity can reduce variance estimates in the presence of high leverage points. Simulation plots and real example illustrate the methodology.

Keywords: Logistic regression, maximum likelihood, collinearity, high leverage point, eigenvalue, condition number

1.0 Introduction

The logistic regression using maximum likelihood (ML) estimator has been in the epidemiological study to model the probability of survival and the assessment of risk factors in diseases growth. Although the methodology of ML estimator is well developed, there is still lack of investigation on the model assessment in the situation when both collinearity and high leverage points occur together.

Collinearity seriously affects the ML estimator in that the variance estimates is inflated in much the same way the collinearity inflates the variance estimates of Least Square (LS) estimator in linear regression (Månsson and Shukur, 2011). Another effect of collinearity is having a non-significant of Wald statistic for a single regressor while the overall model may be strongly significant (Lesaffre and Marx, 1993). Furthermore, collinearity may also result in changing signs and increase the magnitudes of estimated regression coefficients (Kibria et al., 2012). Moreover, collinearity introduces a near singularity in the cross product of regressors $X^T X$ (Marx and Smith, 1990; Lesaffre and Marx, 1993). However, collinearity is not the only problem that involves the regressors for the logistic regression. Marx and Smith (1990) and Lesaffre and Marx (1993) investigated the collinearity among

weighted regressor in the Fisher information matrix, X^TWX and both X^TX and X^TWX produce imprecision of ML estimates. The collinearity diagnostic procedure of Lesaffre and Marx (1993) can be applied to the final solution of X^TWX and X^TX. When applied to the matrix X^TX, the collinearity diagnostic will detect near dependencies in the regressors. Meanwhile, the collinearity diagnostic to the matrix X^TWX will detect near singularities in the weighted regressors which affect the stability of the estimated regression coefficients. Dependencies in the matrix X^TX are often associated with near singularities in X^TWX. However, this will not always be the case in logistic regression since singularities in X^TX and X^TWX may be dissimilar depending on how the magnitude of weights, w_i vary. The effect of weights, w_i where W is the $n \times n$ diagonal matrix is found to be a factor in reducing or inducing ill-conditioning in X^TWX.

High leverage point (HLP) which is an outlying observation in covariates space causes more difficulties to collinear data. The HLPs biased the estimated regression coefficients and obscure other observations (Ariffin and Midi, 2010). According to Bagheri et al. (2012) and Bagheri and Midi (2012), the presence of HLPs in collinear data are capable of enhancing or decreasing the effect of collinearity. Therefore, they have developed high leverage collinearity influential observation (HLCIO) diagnostic procedure in the linear regression model.

The focus of this paper is to investigate the effect of HLPs on collinear data. In Section 2, we provide collinearity diagnostic procedure by Lesaffre and Marx (1993). Section 3 contains simulation experiment and numerical example is exhibits in Section 4. Section 5 offers a conclusion.

2.0 Materials and Methods

The logistic regression model with binary response can be formulated in link linear logit function

$$\boldsymbol{logit}(\boldsymbol{\pi_i}) = \boldsymbol{ln}\{\boldsymbol{\pi_i}/(\mathbf{1}-\boldsymbol{\pi_i})\} = \boldsymbol{x_i^T\beta} \qquad (\mathbf{1})$$

or in probability of occurrence of an event (success)

$$\boldsymbol{\pi_i} = \boldsymbol{exp}\left(\boldsymbol{x_i^T\beta}\right)/\left\{\mathbf{1}+\mathbf{exp}\left(\boldsymbol{x_i^T\beta}\right)\right\} \qquad (\mathbf{2})$$

where $\boldsymbol{x_i}$ is the *i*-th row of an $\boldsymbol{n} \times (\boldsymbol{p}+\mathbf{1})$ matrix $\boldsymbol{X}$ with $\boldsymbol{p}$ explanatory variables and $\boldsymbol{\beta}$ is

a $(\boldsymbol{p}+\mathbf{1})\times\mathbf{1}$ vector of regression coefficients. The iterative maximum likelihood scheme for logistic regression can be expressed as:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{ML}} = (\boldsymbol{X^T W X})^{-1}(\boldsymbol{X^T W z}) \qquad (\mathbf{3})$$

where $\widehat{W} = diag\big(\pi_i(1-\pi_i)\big)$ and $z_i = x_i^T\beta + \{y_i - \hat{\pi}_i/\hat{\pi}_i(1-\hat{\pi}_i)\}$.

Similar to linear regression model, collinearity diagnostic in logistic regression model is detected using condition indices (CI) and condition number (CN) by computing eigenvalues from X^TX and X^TWX (Lesaffre and Marx,1993). We summarize their algorithm as below:

Step 1: Scaling the columns of X (including the intercept term) to unit length $x_{ij}^* = x_{ij}/\|X_j\|, i = 1,\dots,n\,;\ j = 1,\dots,p.$

Step 2: Compute eigenvalues $\hat{\lambda}_0,\dots,\hat{\lambda}_p$ from the information matrix, $\widehat{W} = (X^TWX)$ and $\hat{X} = (X^TX)$ and arranged in decreasing order.

Step 3: Define the condition indexes of matrix $\widehat{W}$ and $\hat{X}$ $\kappa_{W_j} = \big(\hat{\lambda}_0/\hat{\lambda}_j\big)^{1/2}$ and $\kappa_{X_j} = \big(\hat{\lambda}_0/\hat{\lambda}_j\big)^{1/2}$.

Step 4: Define the condition numbers of $\kappa_W = \hat{\lambda}_0/\hat{\lambda}_p$ and $\kappa_X = \hat{\lambda}_0/\hat{\lambda}_p$ and ratio $r_{WX} = \kappa_W/\kappa_X$.

Step 5: Determine whether there is an ill-conditioned in X and ML. According to the threshold given by Lesaffre and Marx (1993), if $\kappa_X \geq 30$, there is collinearity in X , if $\kappa_W \geq 30$ and κ_X is not high, there is ML-collinearity. If both (≥ 30) and the ratio is $r_{WX} \geq 1$, there are collinearity exist in both X and ML.

Step 6: Calculate the variance decomposition proportion table of $\hat{X}$ and $\widehat{W}$ to determine which regressor that highly correlated

3.0 Simulation Experiment

We generate the first explanatory variable $x_1 \sim N(0,1)$ from the Normal distribution. In order to create a collinear data, the second explanatory variables, x_2 is generated in the form of $x_2 = x_1 + U(0,0.1)$. Thus, both x_1 and x_2 are now highly correlated. The outcome variable y_i are generated by comparing $u \sim U(0,1)$ with the true probability $\pi(x_i) = exp(\beta_0+\beta_1x_1+\beta_2x_1)/\big(1+exp(\beta_0+\beta_1x_1+\beta_2x_1)\big)$ given true $\beta = (0,1,2)$. If $u < \pi(x_i)$, then $y = 1$, otherwise $y = 0$. The sample size is fixed for $n = 100$. The data is contaminated by 5% of HLPs $h \sim U(10,15)$ where HLP are generates using Uniform distribution and they are allocated at the last five rows of the x_2 variable. The estimation

of parameters using the ML estimator is repeated for 2500 times. Then, the $\hat{\beta}_2$ and the $se(\hat{\beta}_2)$ for 2500 estimates for each are plotted using the histogram to show how they vary. Figure 1 and Figure 2 illustrate the behavioral of two estimates for three types of data (i) without collinearity (ii) with collinearity and (iii) with collinearity and HLPs.

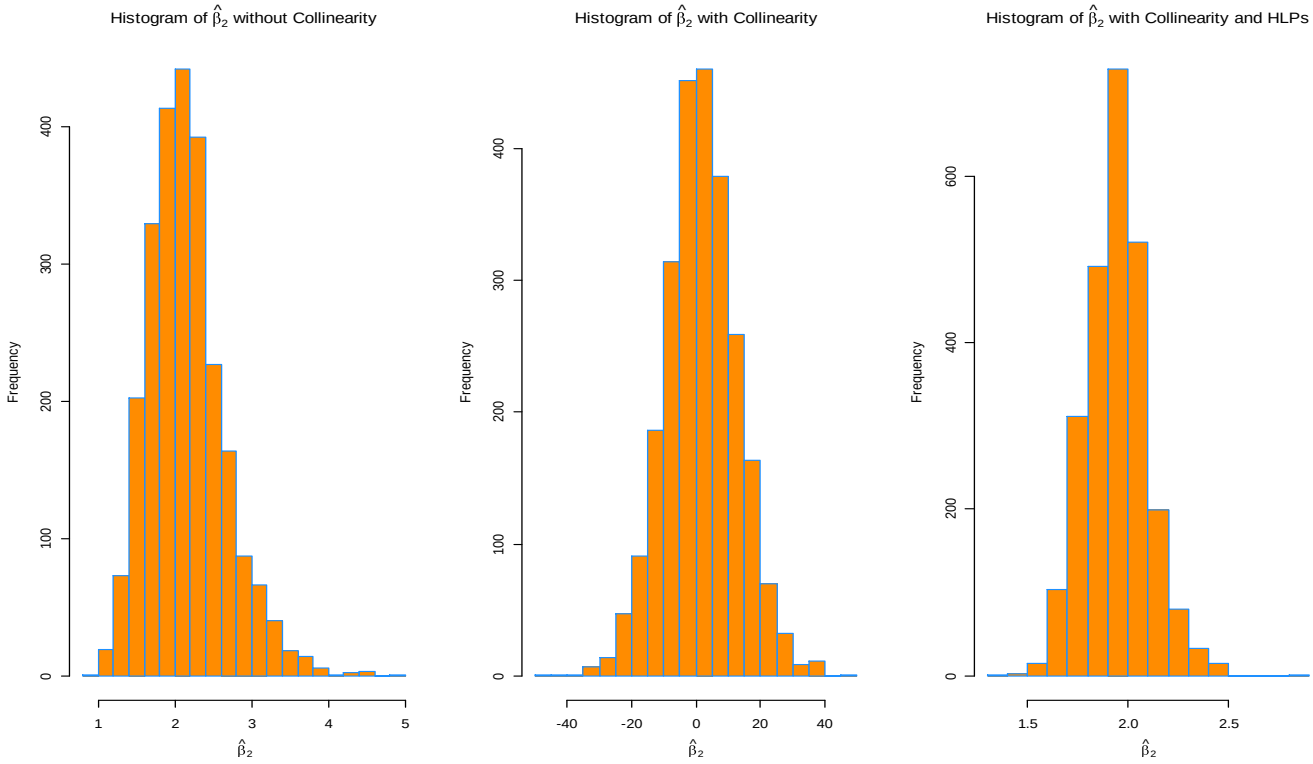


Figure 1. Histogram of $\hat{\beta}_2$ for 2500 replications

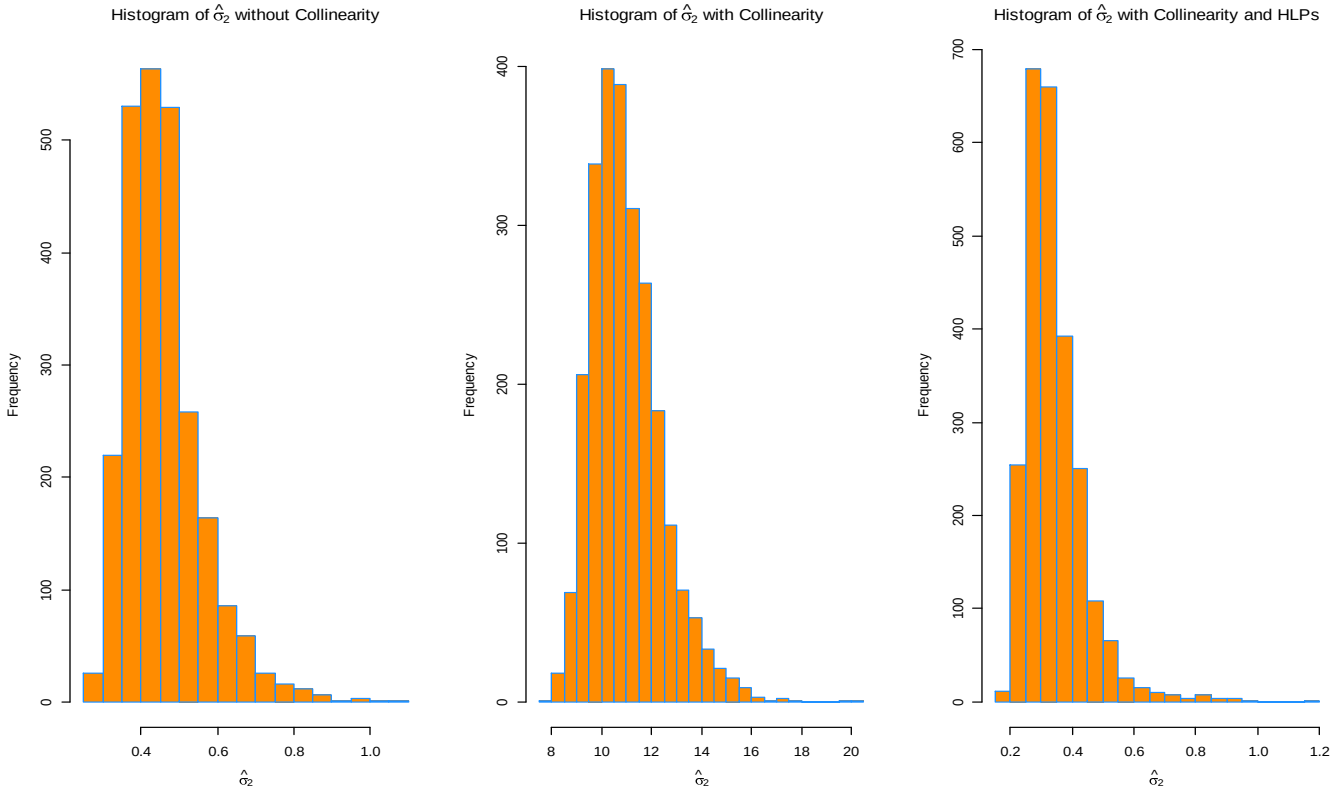


Figure 2. Histogram of $se(\hat{\beta}_2)$ for 2500 replications

Figure 1 shows the histogram plot for 2500 $\hat{\beta}_2$ estimates for three simulated data i.e. without collinearity (left side), with collinearity (middle) and colinearity in the presence of 5% of HLPs (right side). $\hat{\beta}_2$ has a true value of $\beta_2 = 2$ and the histogram plot for three simulated data show that the $\hat{\beta}_2$ values are centered near to β_2. However, the magnitude of $\hat{\beta}_2$ is much larger for collinear data with the interval [-40,40] compared to without collinearity with the interval [1,5] and some of $\hat{\beta}_2$ also change the sign to negative. Surprisingly, the distribution of $\hat{\beta}_2$ estimates in collinear is drastically change in the presence of HLPs where the magnitude $\hat{\beta}_2$ estimates dropped from the interval [-40,40] to [1.5,2.5] which is not too different from the interval of $\hat{\beta}_2$ [1,5] for good data (without collinearity and no HLPs). Segerstedt and Nyquist (1992) mentioned that the change of collinear pattern is strongly related to the matrix weight $\widehat{W} = diag(\hat{\pi}_i(1-\hat{\pi}_i))$ in Eq. (3). They found in some cases, strong collinearity can be weakened by the weight. Meanwhile, Hosmer and Lemeshow (2000) showed that the observations with HLPs correspond to fitted probability with either large $\hat{\pi}_i \approx 1$ or small $\hat{\pi}_i \approx 0$. The effect of collinearity is seemed to be eliminated as the weights $\widehat{W}$ appear to be small when the $\hat{\pi}_i$ values are large or small due to the presence of HLPs. Thus, this explains why the $\hat{\beta}_2$ estimates in both collinearity and HLPs are smaller compared to $\hat{\beta}_2$ estimates in collinearity. The variance estimates are larger in the simulations performed with collinearity with the interval [8,20] compared to without collinearity with the interval

[0.4,1.0] as shown in Figure 2. However, the variance estimates in both collinearity and HLPs reduce to the interval [0.2,1.2].

4.0 Numerical Example

Our example is on cancer remission data by Lesaffre and Marx (1993) which is taken to illustrate severe collinearity in logistic regression. The continuous risk factors associated with cancer remission are cell index (CELL), temperature (TEMP), and li index (LI). The binary response is 1 if the patient experiences a complete cancer remission and 0 otherwise. There were 27 patients involved and 9 of which experienced a complete cancer remission. The modified cancer remission data contains two extreme HLPs on a temperature variable at rows 25 and 26 where the two original observations are replaced with values 10 and 11. Table 1 and Table 2 show the results on collinearity diagnostic for original and modified cancer remission data followed by the parameter estimates, as displayed in Table 3.

Table 1. Collinearity Diagnostic for

High Correlated Cancer Remission Data

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	LI	TEMP	CELL
		X'X			
3.843	1	0	0.010	0	0.003
0.129	5.448	0	0.979	0	0.020
0.028	11.799	0.001	0.003	0.001	0.969
1.06E-04	190.776	0.999	0.008	0.999	0.008
		X'WX			
0.576	1	0	0.005	0	0
0.015	6.165	0	0.454	0	0.007
5.52E-04	32.287	0.005	0.097	0.003	0.816
5.29E-06	329.954	0.995	0.444	0.997	0.176

Table 2. Collinearity Diagnostic for

High Correlated Cancer Remission Data with HLPs

Eigenvalue	Condition Index	Variance Decomposition Proportion			
		Intercept	LI	TEMP	CELL
		X'X			
3.293	1	0.003	0.014	0.031	0.003
0.572	2.400	0.004	0.012	0.952	0.003
0.115	5.346	0.049	0.970	0.002	0.057
2.06E-02	12.652	0.943	0.004	0.016	0.936
		X'WX			
0.419	1	0.001	0.015	0	0.001
0.013	5.644	0.009	0.897	0.001	0.022
1.07E-03	19.785	0.005	0.006	0.861	0.158
5.89E-04	26.668	0.985	0.082	0.137	0.819

The condition number of $\boldsymbol{\kappa_x = 190.776}$ and $\boldsymbol{\kappa_W = 329.954}$ with ratio $\boldsymbol{r_{wx} = 1.73}$ determined the ill-conditioning in matrix $\boldsymbol{X}$ and information matrix of ML (see Table 1). The variance decomposition proportion table shows the high correlation between temperature variable and the intercept term with correlation values 0.99 as also being pointed the same by Lesaffre and Marx (1993). In the presence of HLPs (see Table 2), the condition numbers reduce to $\boldsymbol{\kappa_x = 12.652}$ and $\boldsymbol{\kappa_W = 26.668}$. We also observe that the correlation between Intercept term and temperature variable is now change to cell variable.

Table 3. Parameter Estimation of Cancer Remission Data

	ESTIMATOR			
	ML	LR	WBY	RLR
	Collinearity			
Intercept	67.634 (56.888)	-2.99E-03 (6.76E-03)	66.745 (73.569)	-3.06E-03 (6.87E-03)
LI	3.867 (1.778)	1.34E-03 (8.43E-03)	3.818 (2.238)	1.34E-03 (8.56E-03)
TEMP	-82.074 (61.712)	-2.94E-03 (6.72E-03)	-80.967 (81.163)	-3.02E-03 (6.82E-03)
CELL	9.652 (7.751)	-2.71E-03 (6.44E-03)	9.494 (6.497)	-2.77E-03 (6.54E-03)
	Collinearity and High Leverage Points			
Intercept	-8.752 (6.054)	-7.07E-02 (4.90E-02)	51.425 (82.316)	-6.593E-03 (1.00E-02)
LI	2.862 (1.298)	-1.07E-02 (6.08E-02)	3.631 (2.331)	4.800E-04 (1.253E-02)
TEMP	0.716 (1.963)	-6.22E-02 (5.22E-02)	-62.963 (89.662)	-6.501E-03 (9.986E-03)
CELL	4.408 (5.724)	-6.26E-02 (4.62E-02)	6.977 (6.705)	-6.039-03 (9.483E-03)

On the parameter estimates, we compare the ML estimator with several estimators i.e. logistic ridge (LR) by Månsson and Shukur (2011), the robust weighted Bianco and Yohai (WBY) by Croux and Haesbroeck (2003) and robust logistic ridge (RLR) by Midi and Ariffin (2017). The LR estimator is expected to be the best estimator in collinear and uncontaminated data. Meanwhile, the WBY estimator performs the best estimates for non-collinear contaminated data. However, the LR and WBY estimates are no longer reliable for both collinear and contaminated data. Therefore, RLR estimator is proposed to remedy this problem. Detailed explanation of the methodology for mentioned estimators are not given due to the space constraint.

The LR estimator is always expected to give the best estimates in collinear and uncontaminated data. Refer to Table 3, the RLR estimates are fairly close to the LR estimates. On the other hand, the ML and the WBY estimators fail to provide good estimates as they have larger values for both estimated regression coefficients and standard errors, while the estimated regression coefficient of Intercept and cell variable change sign.

A good estimator for both collinear and contaminated data is the one that has smallest standard errors and estimated regression coefficients which are closest to estimates for the LR in collinear and uncontaminated data. As to be expected, the RLR outperforms other estimators in collinear and contaminated data. Even though the RLR standard errors are slightly larger compared to the LR standard errors in collinear and uncontaminated data, the RLR estimated regression coefficients are not strayed too far or not changing the sign. The WBY estimated regression coefficients and standard errors do not change much from its previous estimated regression coefficients and standard errors in collinear and uncontaminated data and, but they give faulty inference. Meanwhile, there are reductions in the standard errors using ML estimator. Although the ML estimator gives smaller standard errors in collinear and contaminated data compared to the standard errors in collinear and uncontaminated data, these estimates are misleading and not reliable. We also observed the sign different for estimated regression coefficients of temperature and cell variables. The LR also affected by the presence of high leverage points in correlated data, evident by having larger standard errors and different sign for the estimated coefficient of li variable compared to the LR standard errors and estimated regression coefficients in collinearity and uncontaminated data.

5.0 Conclusion

Many circumstances in logistic regression model encountered a problem of having a severe collinearity and high leverage points. The simulation plot and empirical result indicate that collinearity seriously affects the ML estimator by producing large and unstable estimates. The ML estimator gives a misleading conclusion on parameter

estimation in the presence both colinearity and HLPs, whereby the standard error of the ML estimates are reduced but they are not reliable. The RLR estimator offers substantial improvement over the ML, the WBY, and the LR estimators for the collinearity and high leverage points. The findings obtained from real example indicate that the RLR method is the best estimator.

References

Ariffin, S.B. and Midi, H. (2010). Robust logistic diagnostic for the identification of high leverage points in logistic regression model. *Journal of Applied Sciences*, 10(23): 3042-3050.

Bagheri, A., and Midi, H. (2012). On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations, *Mathematical Problems in Engineering*, vol. 2012, Article ID 531607, 16 pages. doi:10.1155/2012/531607.

Bagheri, A., Midi, H. Midi and Imon, A.H.M.R. (2012). *Communications in Statistics – Simulation and Computation,* 41**:** 1379-1396.

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, 44(1-2): 273-295.

Hosmer, D.W., and Lemeshow, S. (2002). Applied logistic regression. New York. John Wiley & Sons.

Kibria, B.M.G., Månsson, K. and Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics*, 40(4): 401-414.

Lesaffre, E. and Marx, B.D. (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods*, 22(7): 1933-1952.

Månsson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. *Communications in Statistics – Theory and Methods*, 40(18): 3366-3381.

Marx, B.D. and Smith, E.P. (1990). Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(6): 1128-1135.

Midi, H., Ariffin, S. B. (2017). Weighted High Leverage Collinear Robust Ridge Estimator in Logistic Regression Model. *Pakistan Journal of Statistics.* 34(1): 55-75.

Segerstedt, B. and Nyquist, H. (1992). On the conditioning problem in generalized linear models. *Journal of Applied Statistics*, 19(4): 513-526.

CHAPTER 9

Analysis of genetic diversity in closely related plant species using multivariate analyses in comparison with molecular marker evidence

Abstract

Knowledge on genetic diversity among the closely related plant species is important for species separation, breeding works, conservation and management of crop germplasm. A number of approaches are currently available for species identification and germplasm characteristics. These methods have relied on morphological studies and or in combination with molecular approaches. The present study focused on nine accessions of closely related 5 *Passiflora* species; i.e, *Passiflora quadrangularis*, *Passiflora maliformis*, *Passiflora incarnata,* 2 varieties of *Passiflora foetida* and 4 varieties of *Passiflora edulis* as an example aimed to study the purposes of multivariate analyses for species separation. Previous studies showed the effectiveness of floral characteristics to identify *Passiflora* species, and distinction between several closely related species was difficult. In order to resolve the taxonomic uncertainty, we tested a total of 43 quantitative characteristics (12 vegetative and 31 floral) of 715 specimens corresponding to the different species/varieties. Multivariate analysis of principal component (PCA) was employed to reduce the data sets from 43 to 26 quantitative characteristics as a selection criterion for species separation. The selected 26 characteristics were subjected further to discriminant analysis (DA) and those traits were discriminate best among the nine *Passiflora* accessions and to obtain reliable discriminant functions for provision of maximum separation among the species. As a result, the nine *Passiflora* accessions were clustered into five distinct groups with no overlapping between species. *Hierarchical* cluster analysis further confirmed the species separation. The classification of *Passiflora* species were further elucidated using molecular methods (ITS) and the genetic diversity was consistent with morphological classification. Combination of morphological traits using appropriate set of multivariate analyses and molecular approaches are useful for distinguishing the closely related *Passiflora* species.

Keywords: Discriminant analysis (DA), *hierarchical* cluster analysis, multivariate analyses, Passiflora species, principle component analysis (PCA)

Introduction

Passiflora plants generally known as passion fruit may well be the most fascinating plant of the tropics. The *Passiflora* has unique flowers, usually complicated in forms, variable in shape, structure and colour (Ulmer & MacDougal, 2004). *Passiflora* plants belongs to the family of Passifloracea consists of 18 genera including genus *Passiflora*. The species

of the *Passiflora* genus have a wide range of morphological characteristics and anatomical differences (Krosnick & Freudenstein, 2005).

Morphological characteristic have traditionally been the most important criteria in making taxonomic decisions and remain so despite of the wide uses of different molecular markers. The phenotypic expressions of morphological quantitative characters usually are determined by several genes, and different populations may remain similar polymorphism at many characters for a long time (Oja & Paal, 2007). According to Sanchez et al. (1999), *Passiflora* species are difficult to classify as some species vary widely in morphology while other species closely resemble each other. Additionally, existing inter- and intra-species dissimilarity among the *Passiflora* species makes understanding the link between morphological plasticity, genotypic diversity, and speciation challenging. A more critical situation is the fact that the external coloration (purple, pink red, red, yellow, orange yellow, red purple, dark purple) of the fruits are a character of complex inheritance and is not dominant, thus displaying a number of intermediate colors, making it not possible to identify (Bernacci, Soares-Scott, Junqueira, Passos, & Meletti, 2008). The variations in morphology were attributed to their adaptation to various habitats or conditions that could produce plants phenotypically different from their native environment.

Multivariate methods are useful for characterization, evaluations and classifications of plant genetic resources when a large number of accessions are to be assessed for several characters of agronomic and physiological importance (Ayana & Bekele, 1999). The usefulness of multivariate methods for handling morphological variation in germplasm collections have been demonstrated in many crop plants, e.g., cereals include barley, maize, oat, rice and wheat. The information generated is useful for identifying groups of accessions that have desirable characters for crossing, for planning efficient germplasm collections, for revealing the pattern of variation in germplasm collection, for establishing core collection and investigating some aspects of crop evaluation (Oja & Paal, 2007; Ayana & Bekele, 1999).

With regards to morphological variation of the *Passiflora* germplasm, some studies have been conducted in the past from various geographical locations and very little work has been done in Malaysia. Taxonomic studies on *Passiflora* are based on the morphology especially their flowers and fruits (Crochemore, Molinari, & Stenzel, 2003; Souza, Pereira, Viana, Pereira, & Madureira, 2004; Viana, Souza, Araujo, Correa, & Ahnert, 2010; Santos et al., 2011) leading to a classification of this genus. However, the above mentioned findings were based on only univariate analysis. Since the morphological features are the most important tools for identifying the plants and breeding works, the present study was conducted to determine the taxonomic rank of these species and to resolve the existing taxonomic confusion in this species.

Materials and Methods

Sample collection

Plant parts from 4 *Passiflora* species; *Passiflora maliformis, P. quadrangularis, P. incarnata* and *Passiflora edulis* with 2 varieties; *P. edulis* (Purple) and *P. edulis* (Frederick) were collected from the cultivated passion fruit farm in UPM Campus Bintulu (03° 12.45' N and 113° 4.68' E), Sarawak . In addition, two other varieties of *P. edulis*; *P.*

edulis (Pink) and *P. edulis* (Yellow) were sampled from small-scale *Passiflora* farms at Kota Kinabalu (05° 58.28' N and 116° 5.72' E), Sabah and Ba'kelalan (03° 58.44' N and 115° 37.08' E), Sarawak, respectively. Two wild cultivars of *P. foetida; P. foetida* (Yellow) and *P. foetida* (Orange) were collected from the bush area (03° 10.25' N and 113° 2.39' E) at Bintulu, Sarawak. Plant materials (Vegetative and reproductive parts) were collected from randomly selected plants.

Morphological observation on vegetative and reproductive variables

Fifty-one (51) quantitative and 50 qualitative data on the arils parts; leaves, stems, flowers, fruits and seeds (Figure 1). The variables were recorded and measured using ruler and Mitutoya Digimatic Vernier Caliper. The detail morphological structures were observed under 3D microscope (Keyence VH-S30K) and digital images of plants parts were captured using Olympus FE-320 digital camera. Specimen identification and botanical nomenclature were based on the taxonomic keys of Ulmer & MacDougal (2004).

Statistical analysis and data processing

Data on morphological variables were statistically analyzed using the SAS 9.1 for Windows. Single-factor analysis of variance (ANOVA) with post hoc Tukey's test ($p<0.05$) was used to compare the mean values. Principal component analyses (PCA) based on Spearman correlation coefficient were performed with total of 43 characteristics *Passiflora* species (12 for vegetative and 31 for floral) is aim to reduce a large sets of variables (Abdi & Williams, 2010) and to check whether data reduction obtained through the new set of variables (PCs) revealed a pattern of variation that is consistant with grouping when largest component of the overall variance were contributed by differences among groups. Variables excluded were sepal length, peduncle length, filament width, petiole length, number of petals, sepals, stigma, style, anther and filament, corona length, ovary length and width, stigma length and width and pollen polar and equatorial diameter. Discriminant analysis (DA) generalization by Fisher (1936) is based on linear combinations of the predictor variables was used to find the maximum separation between the species. The predictive model of group member based on 26 quantitative characteristics of the variables remained after data reduction by PCA mentioned above. Clustering was carried out using hierarchical cluster analysis to specify the distance or similarity measure to be used in grouping (Jacquez, 2009) with Spearmen correlation coefficient method. The analyses were performed using XLSTAT 2013 for Windows (Figure 2).

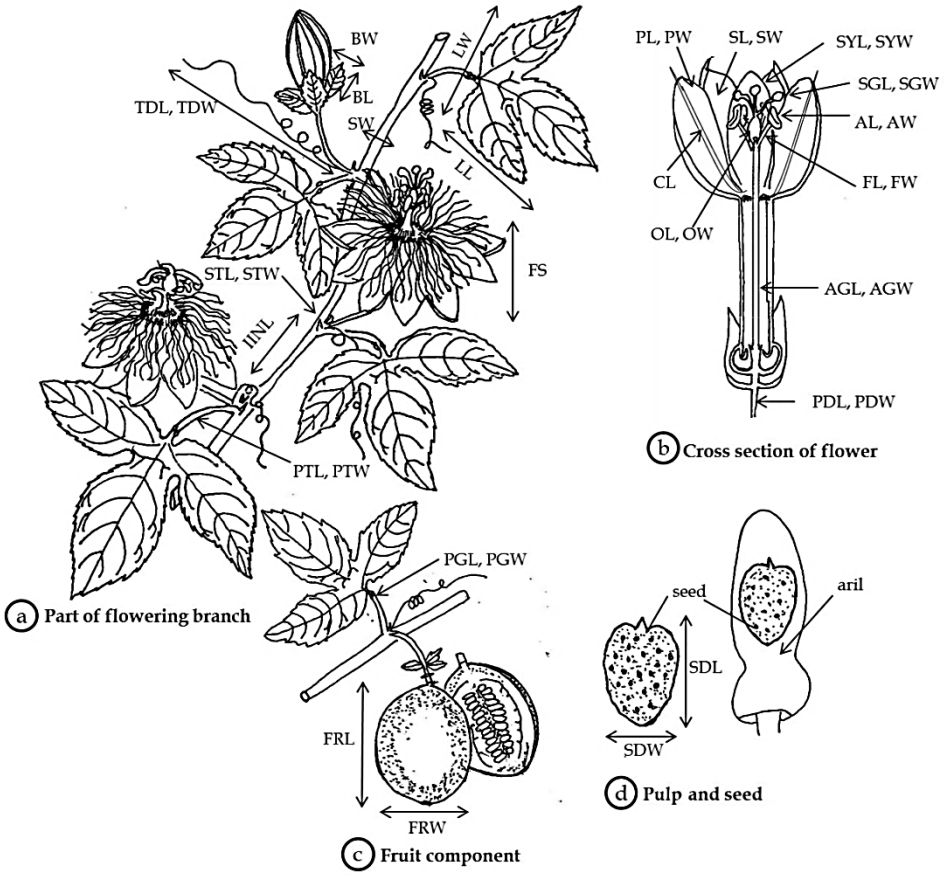


Figure 1: The illustration of quantitative measurement recorded for various parts of *Passiflora* plants. Leaf length (LL) and width (LW); Petiole length (PTL) and width (PTW); Petiole gland length (PGL) and width (PGW); Stem width (SW); internode length (INL); Tendril length (TDL) and width (TDW); Stipule length (STL) and width (STW); Bract length (BL) and width (BW); Peducle length (PDL) and width (PDW); Flower size (FS); Series of corona and coronoa length (CL); Petals number, length (PL) and width (PW); Sepals number, length (SL) and width (SW); Stigma number, length (SGL) and width (SGW); Style number, length (SYL) and width (SYW); Anther number, length (AL) and width (AW); Filament number, length (FL) and width (FW); Ovary length (OL) and width (OW); Androgynophore length (AGL) and width (AGW); Pollen polar and equatorial diameter; Fruit length (FRL), width (FRW), fruit mass, pulp weight and rind weigth; Number of seeds per fruit, length (SDL) and width (SDW).

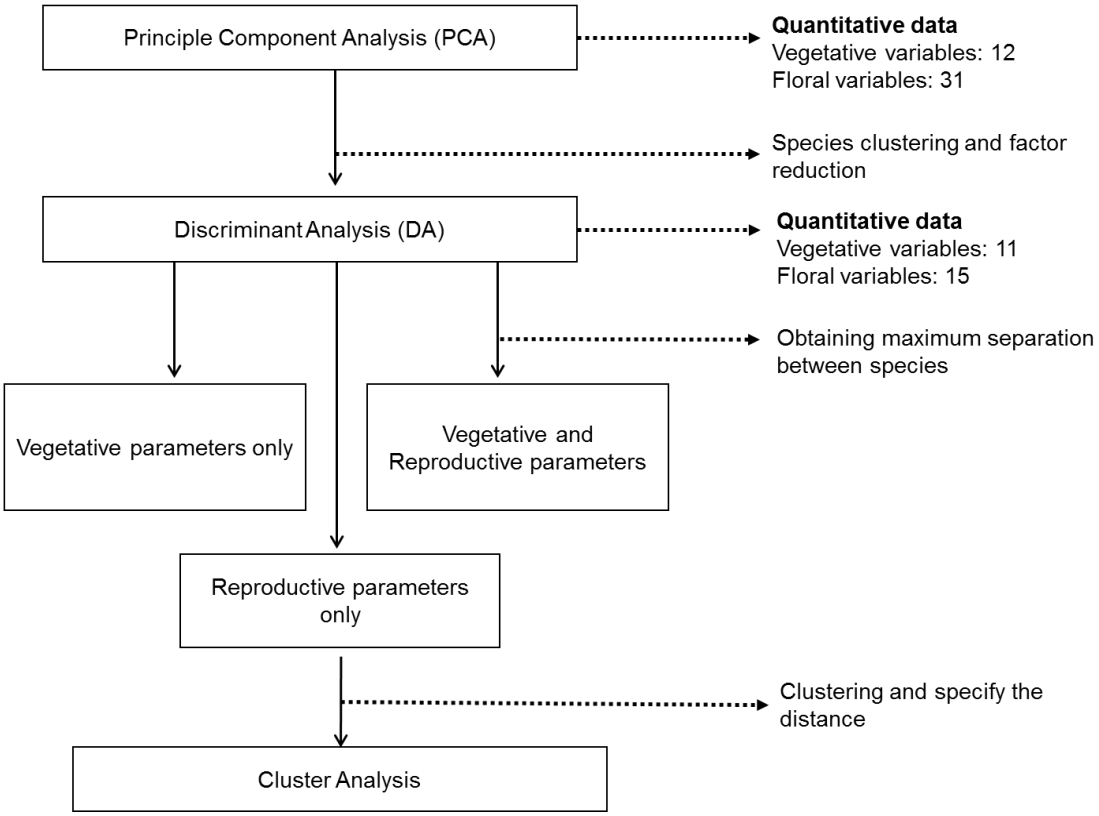


Figure 2: Steps of clustering and confirming the identity of the *Passiflora* species studied.

Results and Discussion

Principal component analysis (PCA)

Principal component analysis (PCA) based on Spearman correlation coefficient was performed in order to evaluate morphological differentiation between species. Data reduction obtained through the new set of variables (PCs) revealed a pattern of variation that is consistent with grouping when largest components of overall variance are contributed by differences among species. The first three PCs of the 43 quantitative characteristics accounted 56.82% of the varience (31.69%, 14.11% and 11.02%, respectively). Although the grouping was consistent, but the total variance of first three principal components after factors reduction was higher (72.50%) and the 9 accessions studied were clustered into four groups (Figure 3a) with overlapping characters observed with *P. edulis* and *P. quadrangularis*. The variables with higher loading factor (>0.60) were chosen for the subsequences analyses. Total of 17 variables were excluded including sepal length, peduncle length, filament diameter, petiole length, number of petals, sepals, stigma, style, anther and filament, corona length, ovary length and width, stigma length and width and pollen polar and equatorial diameter. After the factor reduction, the first two PCs of the remaining 26 quantitative characteristics explained 72.50% of the total varience. Most of the variables were heavily loaded on the postitive axis of PC1 (Figure 3b).

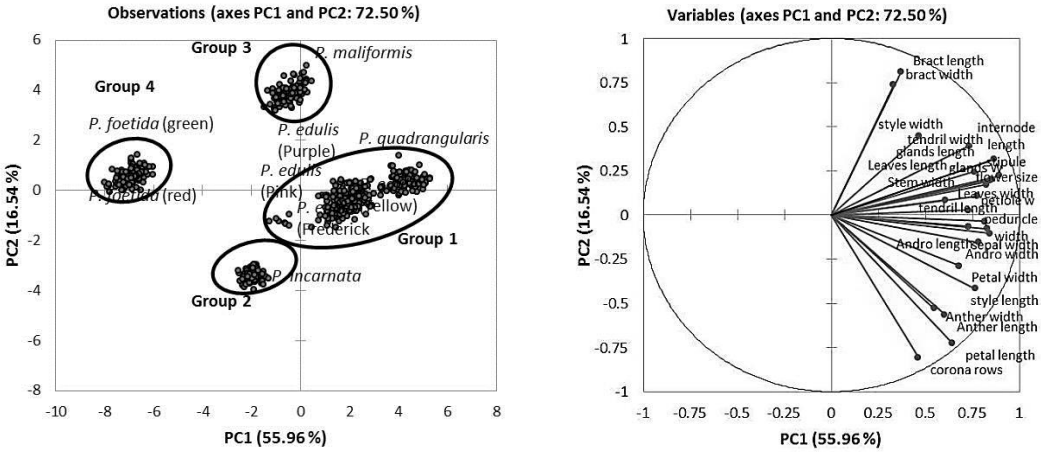


Figure 3: Plot of morphological parameters of the *Passiflora* accessions after factor reduction. (a) bi-plot generated by variables clustered into four groups with overlapping characters and (b) position of the PC score of the variables.

Discriminant analysis (DA)

Discriminant function analysis based on linear combinations of the variables produced a better discrimination of the *Passiflora* species than the PCA. We produced a scatter plot of 715 specimens for the first two discriminant functions based on three different sets of variables; a) vegetative variables alone, b) reproductive variables alone and c) combined both vegetative and reproductive descriptors. The DA results accounted 89.61% of the total variance in vegetative (Figure 4a), 85.64% in reproductive (Figure 4b) and 84.04% in combination of both vegetative and reproductive descriptors (Figure 4c). The biplot generated from combination of vegetative and reproductive descriptors after factor reduction from PCA were chosen with no overlapping characteristics observed showing five well distinct groups discriminated by morphological variables. Overlapping characters were observed when using vegetative or reproductive descriptors alone.

The discriminant factors grouped the *Passiflora* species into five main clusters (Figure 4c). The specimens belonging to Group 1 comprised cultivars of *P. edulis* (Purple, Frederick, Pink and Yellow) were highly discriminated based on most of the vegetative variables; i.e., leaf, stem, petiole, glands on petiole, internode length, stipule, floral characteristics; i.e., flower size, filament and ovary length. Accordingly, with the exception of fruit colour and sizes, there were no significant differences (p>0.05) in morphological variables among *P. edulis* accessions*.* Group 2, consisting of *P. quadrangularis,* located at the positive sites of DF1 and DF2 axis, and the member of this group was highly discriminated by flower features. *Passiflora quadrangularis* produced the largest flowers and fruits of all analyzed species. *Passiflora maliformis* was clearly separated into Group 3 and the members of this group were highly correlated with respect to bract length and width and style width. The bract structure of *P. maliformis* differed from that of other

species; in particular, the three bracts of this species fused together and formed large cups around the bud or flowers. Group 4, which was composed of *P. incarnata* accessions. The species of this group formed an independent cluster with no discrimination on any of the variables. The last group (Group 5) consisted of *P. foetida* (Yellow) and *P. foetida* (Orange). These species clearly diverged from the other *Passiflora* accessions that were located at the negative ends of DF1 and positive end of DF2 axes.

Cluster analysis

Assessment of the morphological traits clustered the species based on their morphological similarities is reflected in the hierarchical cluster analysis dendrogram (Figure 5) and confirmed with grouping using DA (Figure 4c). The dendrogram on Spearman correlation coefficient similarities separates all the species studied into five distinct groups with no overlaps. Group 1 consisted of *P. edulis* species, Group 2 comprised of *P. maliformis*, Group 3 consisted of *P. quadrangularis* and Group 4 composed of *P. incarata* and *P. foetida* clustered in Group 5. The variance composition for clustering was 0.26% for within class and 99.74% for between classes. The dendrogram shower a relative distance value varying from 1.000 to 0.583. The visual evaluations of the dendrogram allow the identifications of homogenous group formed by genotypes showing low variability.

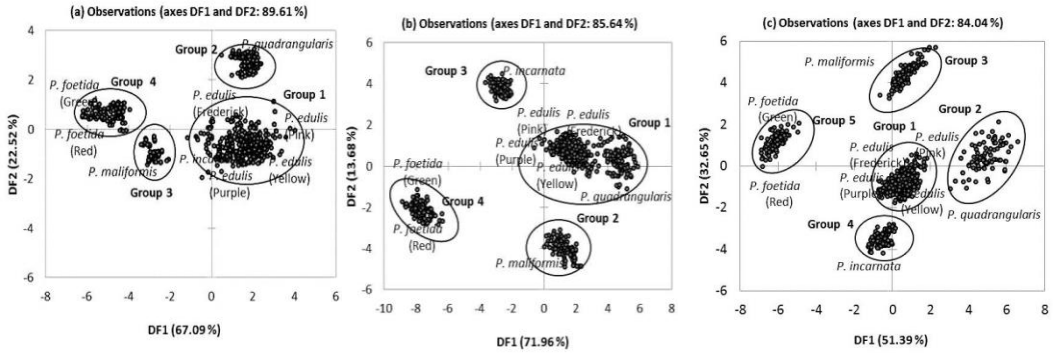


Figure 4: Plot of the morphological parameters of the *Passiflora* accessions. a) Bi-plot generated by vegetative variables, b) bi-plot generated floral variables and c) bi-plot generated by vegetative and reproductive variables.

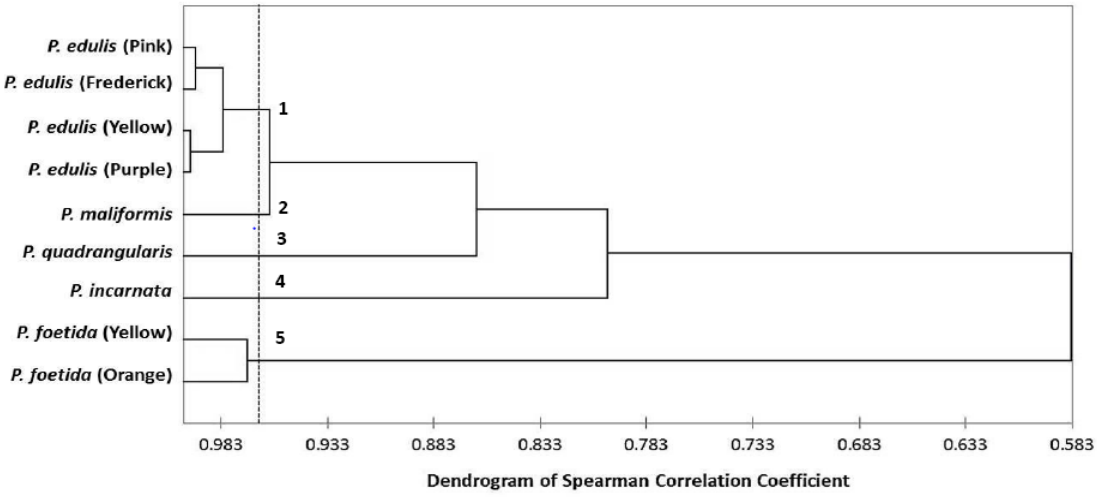


Figure 5: Similarity cluster dendogram of Spearman Correlation Coefficient based on vegetative and reproductive morphological characteristics.

Conclusion

The morphological study provided a useful tool for identification of *Passiflora* species. This study complemented previous classification by Feuillet & MacDougal (2004) and contributed new information regarding the differences in traits such as firm and size. Multivariate analyses using PCA and DA the morphological traits could distinguished the *Passiflora* species. The classification analysis after factor reduction using PCA was very important for infrageneric discrimination in *Passiflora* species. Using DA the species description and separation has become more precise and the species have been clearly distingushed by the combination of vegetative and reproductive features. Discriminant analysis clustered the 9 *Passiflora* accessions into 5 distict groups based on their morphological similarities with no overlapping between the species. All the *P. edulis* sampled from various locations were clustered together in a single group based on their morphological similarities and with the exception of fruit color and fruit size. In addition, cluster analysis also further supported the species separation. The classification of *Passiflora* species were further elucidated using molecular methods (ITS) and the genetic diversity was consistent with morphological classification. Combination of morphological traits using appropriate set of multivariate analyses and molecular approaches are useful for distinguishing the *Passiflora* closely related species.

Acknowledgments

This study was funded by the Ministry of Higher Education Malaysia and UPM under the RUGS-01-01-12-1592RU entitled ‘Comparative studies on passion fruit species and their potential uses’.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459.

Ayana, A., & Bekele, E. (1999). Multivariate analysis of morphological variation in sorghum (*Sorghum bicolor* (L.) Moench) germplasm from Ethiopia. *Genetic Resources and Crop Evolution*, *46*(3), 273-284.

Bernacci, L. C., Soares-Scott, M. D., Junqueira, N. T. V., Passos, I. R. D. S., & Meletti, L. M. M. (2008). *Passiflora edulis* Sims: the correct taxonomic way to cite the yellow passion fruit (and of others colors). *Revista Brasileira de fruticulture*, *30*(2), 566-576.

Crochemore, M. L., Molinari, H. B., & Stenzel, N. M. (2003). Agromorphological characterization of passion fruit (*Passiflora* spp.) germplasm. *Revista Brasileira de Fruticultura*, *25*(1), 5-10.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179-188.

Jacquez, G. M. (2009). Cluster morphology analysis. *Spatial and Spatio-Temporal Epidemiology*, *1*(1), 19-29.

Krosnick, S. E., & Freudenstein, J. V. (2005). Monophyly and floral character homology of old world *Passiflora*. *Systematic Botany, 30*, 139-152.

Oja, T., & Paal, J. (2004). Multivariate analysis of morphological variation among closely related species *Bromus japonicus*, *B. squarrosus* and *B. arvensis* (Poaceae) in comparison with isozyme evidences. *Nordic Journal of Botany*, *24*(6), 691-702.

Sanchez, I., Angel, F., Grum, M., Duque, M., Lobo, M. C., Tohme, J., & Roca, W. (1999). Variability of chloroplast DNA in the genus *Passiflora* L. *Euphytica*, *106*, 15-26.

Santos, E. A., Souza, M. M., Viana, A. P., Almeida, A. A., Freitas, J. C., & Lawinscky, P. R. (2011). Multivariate analysis of morphological characteristics of two species of passion flower with ornamental potential and of hybrids. *Genetics and Molecular Research*, *10*(4), 2457-2471.

Souza, M. M., Pereira, S. T. N., Viana, A. P., Pereira, G. M., & Madureira, H. C. (2004). Flower receptivity and fruit characteristic associated to time of pollination in the yellow passion fruit *Passiflora edulis* Sims f. *flavicarpa* Degener (Passifloracea). *Scientia Horticulturae*, *101*, 373-385.

Ulmer, T., & MacDougal, J. M. (2004). *Passiflora*: *Passionflowers of the World,* USA: Timber Press.

Viana, A. J. C., Souza, M. M., Araujo, I. S., Correa, R. X., & Ahnert, D. (2010). Genetic diversity in *Passiflora* species determined by morphological and molecular characteristics. *Biologia Plantarum*, *54*(3), 535-538.