MODIFIED METHOD FOR REMOVING MULTICOLLINEARITY PROBLEM IN MULTIPLE REGRESSION MODEL

UMS
UNIVERSITI MALAYSIA SABAH

YAP SUE JINQ

SCHOOL OF SCIENCE AND TECHNOLOGY

UNIVERSITI MALAYSIA SABAH

2014

MODIFIED METHOD FOR REMOVING MULTICOLLINEARITY PROBLEM IN MULTIPLE REGRESSION MODEL

YAP SUE JINQ

UMS
UNIVERSITI MALAYSIA SABAH

THESIS SUBMITTED IN FULFILLMENT FOR
THE DEGREE OF MASTER OF SCIENCE

SCHOOL OF SCIENCE AND TECHNOLOGY
UNIVERSITI MALAYSIA SABAH
2014

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

JUDUL: MODIFIED METHOD FOR REMOVING MULTICOLLINEARITY PROBLEM IN MULTIPLE REGRESSION MODEL

IJAZAH: SARJANA SAINS

Saya YAP SUE JINQ, Sesi Pengajian 2010-2014, mengaku membenarkan tesis Sarjana ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/)

□ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

□ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/ badan di mana penyelidikan dijalankan)

□ TIDAK TERHAD

UMS
UNIVERSITI MALAYSIA SABAH

Disahkan oleh,

(Tandatangan Penulis) (Tandatangan Pustakawan)

Alamat Tetap:

.

Tarikh: 24 April 2014

(PROF. DR. ZAINODIN HAJI JUBOK)
Penyelia

(PROF. MADYA DR. AINI JANTENG)
Penyelia Bersama

DECLARATION

I hereby declare that the material in this thesis is my own except quotations, excerpts, equations and references, which have been duly acknowledged.

20 May 2014

Yap Sue Jinq
PS2010-8236

UMS
UNIVERSITI MALAYSIA SABAH

CERTIFICATION

NAME	: **YAP SUE JINQ**
MATRIC NO.	: **PS2010-8236**
TITLE	: **MODIFIED METHOD FOR REMOVING MULTICOLLINEARITY PROBLEM IN MULTIPLE REGRESSION MODEL**
DEGREE	: **MASTER OF SCIENCE (MATHEMATICS WITH ECONOMICS)**
VIVA DATE	: **23 DECEMBER 2013**

DECLARED BY

UMS
UNIVERSITI MALAYSIA SABAH

1. SUPERVISOR
Prof. Dr. Zainodin Haji Jubok — Signature

____

2. CO-SUPERVISOR
Assoc. Prof. Dr. Aini Janteng

____

ACKNOWLEDGEMENT

I would like to thank God for his blessings in enabling me to complete my studies smoothly and keeping me on the right path even when I face obstacles in my life. I wish to express my deepest gratitude and appreciation to my supervisor, Professor Dr. Zainodin Haji Jubok and my co-supervisor, Associate Professor Dr. Aini Janteng, the School of Science and Technology, Universiti Malaysia Sabah. Their guidance and encouragement provided me the impetus to complete this thesis.

I am also thankful to Prof Dr Awang Bono (SKTM, UMS) and Prof. Dr. Zuhaimy (FST, UTM) for their painstaking effort in careful reviewing, critical comments and meaningful suggestions in tidying up the final version of this thesis.

Then, gratitude to my beloved parents: Mr. Michael Yap and Mrs. Theresa Chung, my sisters Michelle Yap and Magdalene Yap for their continuous support, love and motivation. My special thank to my friends: especially Choo Ying Ying and Liaw Bing Shen for their guidance and encouragement.

I am also a very grateful and proud recipient of the Biasiswa Kerajaan Negeri Sabah (BKNS) Scholarship (reference number: JPAN(B&L) 600-2007/399). Finally, thanks to all the lecturers in the Mathematics with Economics Programme who have in one way or another helped me.

Yap Sue Jinq
PS2010-8236

UMS
UNIVERSITI MALAYSIA SABAH

ABSTRACT

Multicollinearity happens when two or more independent variables in a multiple regression model are highly correlated. This increases the standard errors as the coefficients cannot be estimated accurately. Insignificant variable which does not contribute to a model may also affect the interpretation of data. Therefore, the key objective of this work is to develop a best model that is free from multicollinearity problem and insignificant variables. Originally, there are 25 variables in the data set. Using factor analysis, correlation coefficient values and dummy transformation the following variables are identified: body weight as dependent variable, chest diameter, shoulder girth, chest girth, bicep girth, forearm girth and wrist girth each as single quantitative independent variable and ankle diameter, biacromial diameter, elbow diameter, wrist diameter and gender each as dummy variable. The interaction variables involved here is up to the fifth-order (product of 6 variables). Variables which are lowly correlated with dependent variable are not removed, but are transformed into dummy variables. This work also identifies the significance of interaction variables and variables which are lowly correlated with dependent variables in an analysis. So, applying the concept of backward elimination, multicollinearity and coefficient tests are employed to discard variables systematically from each of all possible models. Multicollinearity source variables are removed using a modified method on the Zainodin-Noraini multicollinearity remedial method. Finally, a best model is obtained, free from multicollinearity problem and insignificant variables. Interaction variables are found to play important role as the best model consists of two single quantitative independent variables (chest diameter, forearm girth), four first-order interaction variables (chest girth and wrist girth, and bicep girth each with biacromial, ankle, gender) and one second-order interaction variable (chest girth, chest diameter and shoulder girth). The highest interaction order found in the best model is up to the second-order. Variables which are lowly correlated with dependent variable (biacromial diameter, ankle diameter and gender) are found to be significant and appear in the best model as interaction variables with bicep girth, respectively. Thus, the results of this work suggest a suitable procedure for researchers when dealing with a large number of independent variables.

Keywords: *Multicollinearity, insignificant variable, multiple regression, interaction variable, dummy variable*

UMS
UNIVERSITI MALAYSIA SABAH

ABSTRAK

TEKNIK TERUBAHSUAI UNTUK MENGATASI MASALAH MULTIKOLINEARITI DALAM MODEL REGRESI BERGANDA

Multikolineariti berlaku apabila dua atau lebih daripada dua pemboleh ubah tak bersandar dalam suatu model regresi berganda adalah berkolerasi tinggi. Hal ini meningkatkan sisihan piawai kerana pekali tidak dapat dianggarkan dengan tepat. Pemboleh ubah tidak signifikan yang tidak menyumbang kepada model juga berkemungkinan mempengaruhi pentafsiran data. Oleh itu, objektif utama kerja ini adalah untuk membentuk satu model terbaik yang bebas daripada masalah multikolineariti dan pemboleh ubah tidak signifikan. Pada asalnya, terdapatnya 25 pemboleh ubah dalam set data. Dengan menggunakan faktor analisis, nilai pekali kolerasi dan penjelmaan patung, pemboleh ubah-pemboleh ubah berikut dikenalpasti: berat badan sebagai pemboleh ubah bersandar, diameter dada, liltan bahu, lilitan dada, lilitan bisep, lilitan lengan dan lilitan pergelangan tangan sebagai pemboleh ubah tidak bersandar kuantitatif tunggal dan diameter pergelangan kaki, diameter biakromial, diameter siku, diameter pergelangan tangan dan jantina sebagai pemboleh ubah patung. Pemboleh ubah-pemboleh ubah interaksi sehingga peringkat kelima terlibat (pendaraban daripada 6 pemboleh-ubah). Pemboleh ubah-pemboleh ubah yang berkolerasi rendah dengan pemboleh ubah bersandar tidak disingkirkan daripada analisis, tetapi dijelmakan menjadi pemboleh ubah patung. Kerja ini juga menerokai kepentingan pemboleh ubah interaksi dan pemboleh ubah yang berkolerasi rendah dengan pemboleh ubah bersandar dalam suatu analisis. Oleh itu, dengan mengaplikasikan konsep kaedah penghapusan ke belakang, ujian multikolineariti dan ujian pekali digunakan untuk menggugurkan pemboleh ubah daripada setiap model berkemungkinan secara sistematik. Satu teknik terubahsuai atas teknik mengatasi multikolineariti Zainodin-Noraini digunakan untuk menggugurkan pemboleh ubah punca multikolineariti. Akhirnya, satu model terbaik yang bebas daripada masalah multikolineariti dan pemboleh ubah tidak signifikan diperolehi. Didapati pemboleh ubah interaksi memainkan peranan penting kerana terdapatnya dua pemboleh ubah tidak bersandar kuantitatif tunggal (diameter dada, lilitan lengan), empat pemboleh ubah interaksi peringkat pertama (lilitan dada dengan lilitan pergelangan tangan, dan lilitan bisep masing-masing dengan biakromial, pergelangan kaki dan jantina) dan satu pemboleh ubah interaksi peringkat kedua (lilitan dada, diameter dada dan lilitan bahu) dalam model terbaik yang diperolehi. Didapati peringkat interaksi tertinggi dalam model terbaik ialah peringkat kedua. Pemboleh ubah-pemboleh ubah yang berkolerasi rendah dengan pemboleh ubah bersandar (diameter biakromial, diameter siku dan jantina) didapati signifikan dan masing masing berinteraksi dengan lilitan bisep. Oleh itu, keputusan kerja ini mencadangkan satu prosedur kepada penyelidik lain apabila berurusan dengan bilangan pemboleh ubah tidak bersandar yang banyak.

UMS
UNIVERSITI MALAYSIA SABAH

TABLE OF CONTENTS

UMS
UNIVERSITI MALAYSIA SABAH

CHAPTER 3: METHODOLOGY

UMS UNIVERSITI MALAYSIA SABAH

CHAPTER 4: DATA PREPARATION

CHAPTER 5: DATA ANALYSIS

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF TABLES

UMS
UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF FIGURES

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF ABBREVIATIONS

8SC	Eight Selection Criteria
AIC	Akaike Information Criterion
APLS	Advanced Paediatric Life Support
BMI	Body Mass Index
cov	Covariance
df	Degrees of Freedom
ED	Emergency Department
FPE	Finite Prediction Error
GCV	Generalized Cross Validation
HQ	Hannan and Quinn criterion
LCL	Lower Control Limit
MAPE	Mean Absolute Percentage Error
MHMR	Moderated Hierarchical Multiple Regression
MSE	Mean Square Error
MSR	Mean Square for Regression
NP	Number of Parameters in the parent model
NPP	Normal Probability Plot
OLS	Ordinary Least Squares
PCA	Principal Components Analysis
PDF	Probability Density Function
re	Error term for Restricted model
RM	Restricted Model in Wald test
se	Standard Error
SPSS	Statistical Package for the Social Sciences
SSE	Error Sum of Squares
$SSE(RM)$	Error Sum of Squares for Restricted model
$SSE(UM)$	Error Sum of Squares for Unrestricted model
SSR	Regression Sum of Squares
SST	Total Sum of Squares
UCL	Upper Control Limit
ue	Error term for Unrestricted model
UM	Unrestricted model in Wald test
var	Variance

UMS UNIVERSITI MALAYSIA SABAH

LIST OF SYMBOLS

a	Number of reserved data for estimation purpose
$\mathbf{a}$	Number of the parent model
A	Ankle diameter
A_i	The ith actual observation value $(i = 1,2,\dots,n)$
$\mathbf{b}$	Number of variables removed in multicollinearity test
B	Elbow diameter
c	Biasing constant
$\mathbf{c}$	Number of variables eliminated in coefficient test
D	Biacromial diameter
E	Mean or expected value
E_i	The ith estimated observation value $(i = 1,2,\dots,a)$
g	Number of single quantitative independent variables
G	Gender
h	Number of single independent dummy variables
H_0	Null hypothesis
H_1	Alternative hypothesis
i	Observation
k	Number of independent variables
$(k+1)$	Number of parameters in the selected model
K	Kurtosis coefficient
m	Number of independent variable in the restricted model
M	Maximum values after rounding off to the nearest integer
Max	Maximum values before rounding off to the nearest integer
Md	Mode
n	Number of observations
N	Total number of all possible models (with interaction variables)
ρ	Correlation coefficient of a population parameter (rho)
R	Wrist diameter
r	Simple correlation coefficient
R^2	Multiple coefficient of determination
S	Skewness coefficient
$\hat{\sigma}$	Standard error of estimate
u	Random error term
u_i	The ith observation value of random error term u
v	Highest order of interaction (quantitative variable) in the model
$w_{j,i}$	The ith observation value of jth independent variable W_j of the general model of Multiple Regression for $j = 1,2,\dots,k$ and $i = 1,2,\dots,n$ (including single quantitative independent variable, dummy variable, interaction variable, generated variable and transformed variable)
W_j	Independent variable of the general model of Multiple Regression for $j = 1, 2, \dots, k$
$x_{j,i}$	The ith observation value of independent variable W_j for $j = 1,2,\dots,k$ and $i = 1,2,\dots,n$
X_j	Independent variable of the related model for $j = 1, 2, \dots, k$
X_i	The ith observation value of independent variable X
Y	Dependent variable
y_i	The ith observation value of dependent variable Y
$\bar{Y}$	Mean of dependent variable Y

UMS
UNIVERSITI MALAYSIA SABAH

Z	Z-score
α	Level of significance
μ	Mean
σ	Standard deviation of the mean
σ^2	Variance
Ω_0	Intercept or constant term of the general model of Multiple Regression
Ω_j	Coefficient of the corresponding variable W_j of the general model of Multiple Regression for $j = 1, 2, \dots, k$
$\boldsymbol{\Omega}_j$	The jth parameter value of $\boldsymbol{\Omega}$ for $j = 0, 1, \dots, k$
β_0	Intercept or constant term of the related model
β_j	Coefficient of the corresponding variable X_j for $j = 1, 2, \dots, k$

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF APPENDIX

UMS
UNIVERSITI MALAYSIA SABAH

CHAPTER 1

INTRODUCTION

1.1 Overview

In real life, there are many factors or independent variables affecting one dependent variable. However, researchers may attempt to minimize the number of the possible models and number of parameters rather than estimate coefficients of every possible independent variable initially considered in the data set. Thus, this work focuses on the simplified steps in handling a large number of independent variables. Moreover, by having models in which dependent variable depends on more than one independent variables, this leads this work to the discussion of multiple regression models, which is presented in Subsection 1.2. One of the major challenges in multiple linear regression analysis is to eliminate multicollinearity source variables and insignificant variables from the models. Although some methods have been introduced in encountering this problem, the existing methods are found to have some weaknesses. The method suggested by this work in overcoming this problem is discussed in Chapter 3, while in this chapter, definition on the multicollinearity problem and insignificant variables are described in Subsections 1.3 and 1.4. Another concern in a multiple linear regression model is whether the independent variables interact with each others in affecting the dependent variable. Thus, this chapter also discusses on the interaction variables in Subsection 1.5. Subsection 1.6 presents the problem that are faced in a multiple linear regression model with higher order interaction variables, which lead to the objectives of this work as presented in Subsection 1.8. Subsections 1.9 and 1.10 describe the scope of this work and the importance of implementing this work.

1.2 Multiple Linear Regression

Gujarati and Porter (2009) stated that based on the historical origin of the term “regression” that came from Francis Galton, who observed that although there was a

UMS
UNIVERSITI MALAYSIA SABAH

tendency for short parents to have short children and tall parents to have tall children, the average height of children with a fixed parents' height tended to move or "regress" toward the average height of the population as a whole. In other words, the heights of children of extraordinarily short or extraordinarily tall parents tend to "regress" toward the average height of the population. Karl Pearson, who has collected over a thousand records of heights of members of family groups, verified his friend's, Galton's law of universal regression. He observed that the average height of sons of a group of short fathers was greater than their fathers' height and the average height of sons of a group of tall fathers was less than their fathers' height. In other words, "regressing" short and tall sons alike toward the average height of all men. Galton regarded this as "regression to mediocrity".

However, the modern interpretation of regression is quite different. Regression analysis is about the study of the dependence of one variable (dependent variable) on one or more other variables (independent variables), in order to estimate and or predict the (population) mean of the dependent variable based on the known or fixed (in repeated sampling) values of the independent variables. For instance, multiple regression can be used to predict a student's height (dependent variable) using age, gender and father's and mother's heights (independent variables). Multiple regression is also employed to predict the crop yield in a farm (dependent variable) using the rainfall and amount of fertilizer (independent variables). In the field of studies that are related to body weight, multiple regression is also utilized by Bernal-Orozco *et al.* (2010) in developing a new equation to estimate body weight (dependent variable) in elderly Mexican women by using anthropometric measurements (independent variables). Buckley *et al.* (2012) also employed multiple regression in generating an equation to predict Emergency Department (ED) patients' weights (dependent variable) using the anthropometric measurements, including tibial length and abdominal, neck, chest, arm, and thigh circumferences (independent variables).

UMS
UNIVERSITI MALAYSIA SABAH

1.3 Multicollinearity

Warner (2008) stated that multicollinearity happens when there are high correlations among independent variables (when the correlations among independent variables are more than 0.90). In this case, these independent variables may compete to explain much of the similar variance and it would be difficult to distinguish their contributions to the dependent variable. For better understanding, multicollinearity can be likened to a lovers' triangle. James is in love with two twins who are alike, Kate and Kelly. James feels happy when he is with Kate, and a "meaningful" relationship exists between them. James also feels happy when he is with Kelly, and a "meaningful" relationship also exists between them. However, when James is with both Kate and Kelly, James is confused and cannot separate their individual characteristics as they are so much alike, and so no "meaningful" relationship exists between them. This is just like multicollinearity. There exists a significant ("meaningful") relationship between the dependent variable (James) and either independent variable (Kate or Kelly). But confusion reigns when both independent variables exist at the same time. Or, multicollinearity can also be similar to the case when two people are singing loudly and it is hard to discern who is louder as they offset each other.

UMS UNIVERSITI MALAYSIA SABAH

According to Gujarati and Porter (2009), the term multicollinearity is first used by Ragnar Frisch. Originally, multicollinearity means the existence of an exact or perfect linear relationship among some or all independent variables of a regression model. However, the term multicollinearity is utilized in a broader sense nowadays to include the case of perfect multicollinearity, as well as the less than perfect multicollinearity. The multicollinearity problem may exist due to several factors: the data collection method employed, constraints on the model or in the population that being sampled, model specification and an overdetermined model. Firstly, multicollinearity may happen due to the method employed in collecting data; where the sampling is carried over on a limited range of the values taken by the independent variables in the population and thus insufficient variability in the values of the independent variables. Secondly, multicollinearity also exists when there are constraints on the model; for instance, in the regression of blood pressure (dependent variable) with BMI (weight/height^2) and weight (independent variables), there is a constraint in the model as weight is involved in the formulation of BMI. Thirdly, model specification such as adding polynomial terms to a regression model also causes multicollinearity, especially when the range of independent variable is small. Next, an overdetermined model, or which is defined as a model with more independent variables than the number of observations in the sample is also one of the factors that causes multicollinearity.

1.4 Insignificant Variables

Ramanathan *et al.* (1997) mentioned that insignificant variables can affect the precision of the coefficients that are left in the model and the powers of hypothesis tests. Therefore, in their work on developing a number of models to produce very short run forecasts of hourly system loads, they omitted variables which were insignificant to improve the efficiency of their work. In the recent year, Jackson (2012) also discussed on the inclusion of variables that are highly insignificant in an equation. He mentioned that these variables tends to raise the standard error of estimate and may cause variables that are significant in reality to be described as insignificant. In the following year, Haines and Fiori (2013) also supported the idea of Jackson (2012). They stated that the elimination of insignificant variables from model can aid in reducing the model dimensionality and thus, gives a more accurate standard error of estimate.

UMS
UNIVERSITI MALAYSIA SABAH

1.5 Interaction Variables

Allen (1997) mentioned that in multiple regression analysis, people always make the initial assumption that the dependent variable can be predicted most accurately by a linear function of the independent variables. Nevertheless, the effects of independent variables on a dependent variable are not always additive, nonadditive effects or interaction effects may also present in some cases. According to Allison (1999), interaction involves two or more than two independent variables. Black and Eldredge (2002) stated that an interaction variable can be created by multiplying the data values of one variable by the values of another variable. They also supported the idea of Allen (1997) that the effects of two variables are not additive in some cases. They believed that there are interacting effects between the two variables. For instance, interaction effects may present in the independent variables age and education level in estimating monthly income. Interaction effects may also present between temperature and humidity in estimating the annual crop. The later work carried out by Ge and Frick (2007) supported the idea of Black and Eldredge (2002), namely the interaction variables are useful in cases where the dependent variable does not exhibit a linear relationship with the independent variables. They have included 3 first-order interaction variables in their work on modeling of beach bacteria concentrations using multiple regression. They found that omitting interaction variables may result in biased models. In this sense, researchers should include interaction variables in the model and identify the significance of interaction variables by using multiple regression analysis.

1.6 Problem Statement

Researchers may always attempt to include all the possible independent variables in their analyses so that they would not lose any important information. However, in real life, there may happen that a dependent variable is affected by too many independent variables. Thus, when facing a large number of independent variables, the doubt of whether to use all the independent variables in the analysis or to discard some of the independent variable from analysis remains a problem. Besides, researchers also attempt to get a best model which is free from multicollinearity problem and insignificant variables that would give the best estimation on the dependent variable. However, the procedures in getting a best model which is free from multicollinearity problem and insignificant variables remain unclear because the existing methods are found to have their own weaknesses.

UMS UNIVERSITI MALAYSIA SABAH

1.7 Motivation of Work

The studies carried out by Tay *et al.* (2012), Zainodin and Yap (2010) and Zainodin *et al.* (2011) have motivated this work. This is because in the work carried out by Tay *et al.* (2012), SPSS did not run the regression on all of the predictors in one of their models. This can be seen from the excluded variables table shown in their work, where three predictor variables are excluded from the regression model. It is possible that that there exists multicollinearity among the predictors and this may degrade the