

Modeling Medical Data Using MM-Estimation

Applied to Body Mass Index Data

Nor Azlida Aleng¹, Nyi Nyi Naing², Zurkurnai Yusof³, Norizan Mohamed⁴,
Siti Hasliza Ahmad Rusmili⁵ and Nur Farahana Zainudin⁶

^{1,2,3} School of Medical Sciences, Health Campus, Universiti Sains Malaysia 16150
Kota Bharu, Kelantan, Malaysia

^{4,5,6} School of Informatics and Applied Mathematics, University Malaysia
Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia

Copyright © 2015 Nor Azlida Aleng et al. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

In medical statistics research, there are many methodologies used to investigate and to model the relationship between two or more variables. A model is often not useful when it fails to fit the data and the outliers may exist. Outliers play an important role in regression. An outlier (observation) that is quite different from most of the other values or observations in a data set. Robust regression is the most popular method that has been used to detect outliers and to provide resistant results in the presence of outliers in the data set. The purpose of this study is to show that, robust MM-estimation is an alternative approach in dealing with outliers presence in the medical data. This approach is extremely useful in identifying outliers and assessing the adequacy of a fitted model.

Keywords: Robust regression, MM-estimation, Outliers, Body mass index

Introduction

The body mass index (BMI) is a measure of body fat based on height and weight. It is calculated by dividing weight in kilograms (kg) by height in meters squared (m^2). BMI is one method used by healthcare professionals worldwide. BMI is used as a reliable indicator to determine whether a person is overweight or clinically obese [1].

According to the previous research, if someone is overweight, he or she is at risk for many diseases and health condition as such heart disease, stroke, diabetes, cancer, high blood pressure, high cholesterol and blood lipids (LDL) and many more. When overweight or obese people lose their weight, they also lower their blood pressure, total cholesterol, LDL cholesterol, increase their HDL cholesterol, improve their blood sugar levels, and reduce their amount of abdominal fat [2].

In recent years, different ranges of BMI cut-off points for overweight and obesity have been proposed, in particular for the Asia-Pacific region [3]. In USA, over half (53%) of all deaths in women with a BMI > 29 kg/m² could be directly attributed to their obesity [4]. Eating behaviors that have been linked to overweight and obesity include snacking/eating frequency, binge-eating patterns, eating out, and exclusive breastfeeding. Physical activity is an important determinant of body weight. In Australia, for older over the age of 70 years, general health status may be more important than being mildly overweight. Some researchers have suggested that a BMI range of 22-26 is desirable for older Australians.

Materials and Methods

This study is focus on the BMI as a dependent variable and independent variables namely in Table 1. A total of 300 respondents were selected and diagnosed to have BMI problem based on WHO criteria. Material of this study is a hypothetical sample which is composed of ten variables. The explanation of the variables is shown in Table 1 and data were collected from Health Centre in Malaysia.

Table 1: Explanation of the Variables

Code	Variables	Explanation of the variables
Y	BMI	Body Mass Index
X1	SBP	Systolic Blood Pressure
X2	DBP	Diastolic Blood Pressure
X3	CHOL	Cholesterol (Mmol/L)
X4	HEIGHT	Height of a patient in cm
X5	WEIGHT	Weight of a patient in kg
X6	WAIST	The length of waist in cm
X7	AGE	Age (Year)

The linear regression can be expressed in terms of matrices as

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

where y is the $n \times 1$ vector of observed response value, X is the $n \times p$ matrix of p regressors (design matrix), β is the $p \times 1$ regression coefficients and ε is the $n \times 1$ vector error terms. In fitting multiple linear regression model (1.1), the most widely used ordinary least squares (OLS) to find the best estimates of β . Unfortunately, in the presence of outliers, the OLS estimators are still unbiased. According to Barnett and Lewis (1994), an outlier is an observation that is inconsistent with the rest of the data. Outliers in the response variable represent model failure. Outliers in the regressor variable values are extreme in X -space are called leverage points.

In this case, we are interested in using robust MM-estimation to model the BMI data since these methods are highly breakdown point and highly efficient when the errors have a normal distribution. This method is also powerful for analyzing data that are contaminated with outliers. In 1987, Yohai introduced the MM-estimation, combines on both M-estimation and S-estimation to achieve a high breakdown point with high asymptotic efficiency [4].

Yohai [4] defined the MM-estimate in three stages as follows:

S1: Compute an initial regression estimate T^0 .

S2: Compute the residuals $\varepsilon_i(T^0)$ and compute the M-scale S_n using rho function ρ_0 .

The M-scale estimate is defined as the value of S_n which is the solution of

$$\frac{1}{n} \sum \rho_0(\varepsilon_i / S_n) = b \quad (2)$$

where b may be defined by $E_\phi(\rho(\varepsilon)) = b$ where ϕ stands for the standard normal distribution.

Let ρ_0 in (2) be a real function satisfying the following assumptions. (i) $\rho(0) = 0$; (ii) $\rho(-\varepsilon) = \rho(\varepsilon)$; (iii) $0 \leq u \leq v$ implies $\rho(u) \leq \rho(v)$; (iv) ρ is continuous; (v) let $a = \sup \rho(\varepsilon)$, then $0 < a < \infty$; (vi) if $\rho(u) < a$ and $0 \leq u < v$, then $\rho(u) < \rho(v)$. (A1)

S3: Compute the estimate T^1 .

Let ρ_1 be another function which satisfies assumption (A1) such that

$$\rho_1(\varepsilon) \leq \rho_0(\varepsilon) \quad (3)$$

$$\sup \rho_1(\varepsilon) = \sup \rho_0(\varepsilon) \quad (4)$$

Then the MM-estimate T^1 is defined as any solution to the equation

$$\sum \rho_1(\varepsilon_i(T^1) / S_n) x_i = 0 \quad (5)$$

$$\text{which verifies} \quad \sum \rho_1(\varepsilon_i(T^1) / S_n) x_i \leq \sum \rho_1(\varepsilon_i(T^0) / S_n) x_i \quad (6)$$

Yohai [4] revealed that $\rho_0(\varepsilon)$ and $\rho_1(\varepsilon)$ can be taken be $\rho_H(\varepsilon/k_0)$ and $\rho_H(\varepsilon/k_1)$, respectively. Stromberg [5] stated that selecting $k_0 = 0.212$ and $k_1 = 0.9014$ will guarantee a high breakdown estimate and the result in 95% efficiency under normal errors, respectively.

Results and Discussions

Robust regression analysis provides an alternative model to improve the fit of the model. If the data are contaminated in the x -space, M-estimation does not do well. Therefore, we used MM-estimation to solve the problems when outliers presence in both the y -direction and the x -space. Table 2 provided the comparisons estimates of β associated with the 95% confidence interval and p -value for MM-estimation and OLS.

Table 2: Comparison of parameter estimates among MM-estimation and OLS for BMI data

Variables	MM-estimation			OLS		
	S.E	β (95% CI)	p -value	S.E	β (95% CI)	p -value
SBP	0.0008	-0.0012 (-0.0027, 0.0004)	0.1407	0.0011	-0.00203 (-0.004, 0.000)	0.0742
DBP	0.0013	0.0004 (-0.0022, 0.0031)	0.7587	0.0019	0.00034318 (-0.003, 0.004)	0.8594
CHOL	0.0003	-0.0003 (-0.0009, 0.0004)	0.3867	0.0004	-0.00056231 (-0.001, 0.000)	0.2201
HEIGHT	0.0024	-0.3060 (-0.3106, -0.3014)	<.0001**	0.0030	-0.31233 (-0.318, -0306)	<.0001**
WEIGHT	0.0021	0.3233 (0.3192, 0.3274)	<.0001**	0.0029	0.32275 (0.317, 0.328)	<.0001**
WAIST	0.0024	0.0014 (-0.0032, 0.0060)	0.5507	0.0033	0.00678 (0.001, 0.013)	0.0434*
AGE	0.0087	0.0042 (-0.0129, 0.0214)	0.6286	0.0126	-0.01289 (-0.038, 0.012)	0.3054

Note: Significant levels: ** $p < 0.01$, * $p < 0.05$

The results of Table 2 show that the standard error for MM-estimation is less than OLS estimation. Results strongly support that when outliers and leverage points

presence in the data set, MM-estimates are highly stable (resistant) compared with the OLS estimation are affected by the outliers.

The model of BMI is, $\hat{y} = 63.2399 - 0.306x_4 + 0.3233x_5 + \varepsilon$. R squared value is 0.882 (88.2%), this indicated the greater the ability of that model to predict a trend. Based on Figure 1, proved that the existing of outliers observations in the BMI data. 13 observations are considered as outliers. Although, the BMI data presence of outliers, results remain robust. The residual analysis plays an important role for determining the adequacy of the model to ensure that the interpretation of the model is valid. Figure 1 presents the Q-Q plot straight lines, which clearly indicate that the error is normally distributed and the histogram also demonstrate normality.

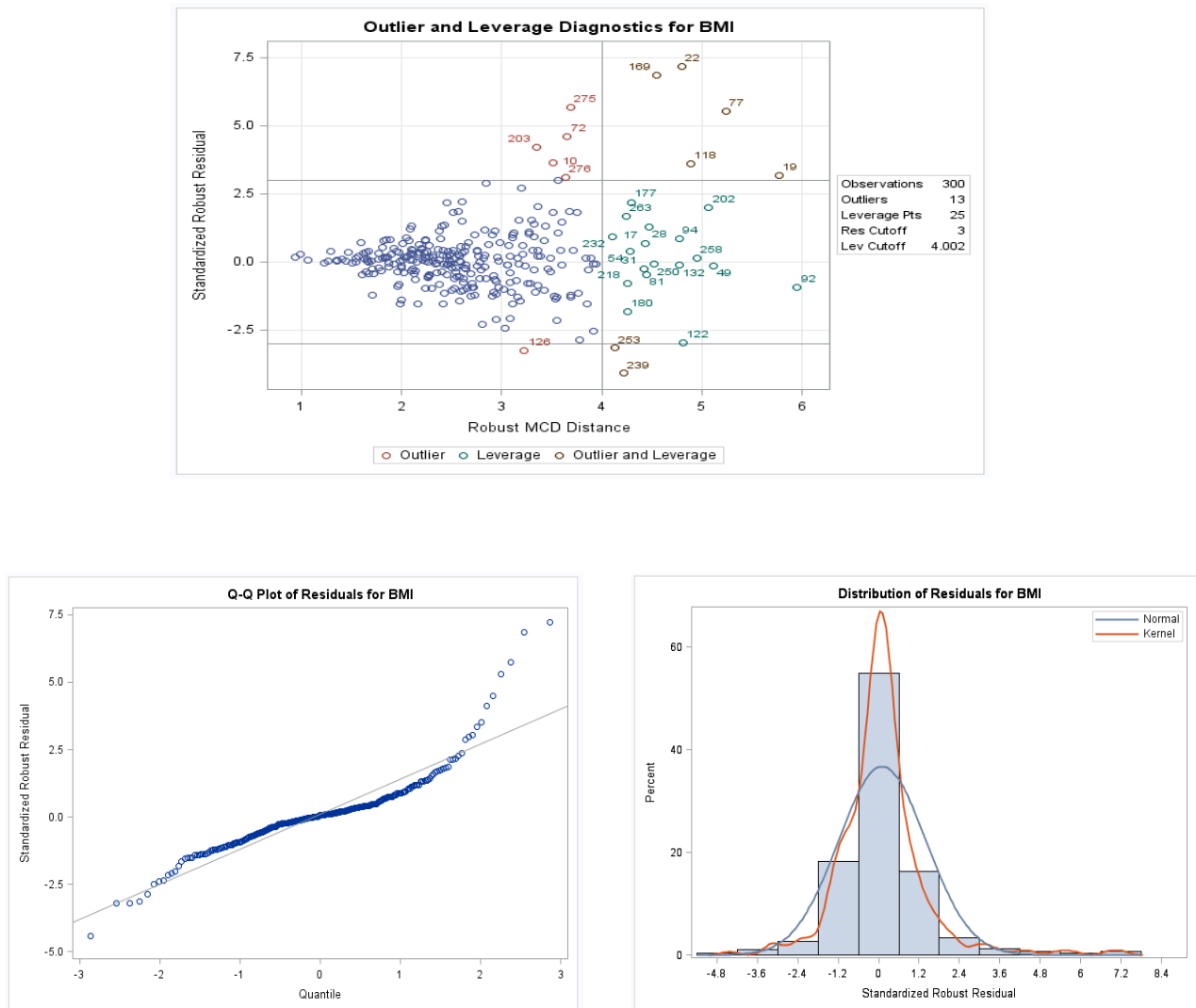


Figure 1: The leverage diagnostics, Q-Q plot and histogram for MM-estimate

Conclusion

Based on these results, it can be concluded that in the presence of outliers in the data, MM-estimation is more sensitive than OLS estimation.

Acknowledgements. The authors would like to thank the Research Management Centre, Universiti Malaysia Terengganu for the financial and moral support in the form of FRGS Grant no. 59266 for the delivery of this paper.

References

- [1] Cosmetic Surgery Consultants.
<http://www.cosmeticsurgeryconsultants.co.uk/obesity-what-is-bmi.htm> [16 April 2014].
- [2] S.P. Walker, E.B. Rimm, A. Ascherio, I. Kawachi, M.J. Stampfer and W.C. Willett, Body size and fat distribution as predictors of stroke among US men, *American Journal Epidemiol.*, **144** (1969), 1143-1150.
<http://dx.doi.org/10.1093/oxfordjournals.aje.a008892>
- [3] WHO, *Diet, Nutrition, and the Prevention of Chronic Disease*, Technical Report Series no 797. Geneva: World Health Organization, 1990.
- [4] V.J. Yohai, High Breakdown-Point and High Efficiency Robust Estimates for Regression, *Ann. Statist.*, **15** (1987), no. 20, 642-656.
<http://dx.doi.org/10.1214/aos/1176350366>
- [5] A.J. Stromberg, Computational of High Breakdown Non-linear Regression Parameters, *J. Am. Stat. Assoc.*, **88** (1993), no. 421, 237-244.
<http://dx.doi.org/10.1080/01621459.1993.10594315>

Received: September 14 2015; Published: October 28, 2015