Jurnal
Teknologi

Full Paper

Effects of Different Type of Covariates and Sample Size on Parameter Estimation for Multinomial Logistic Regression Model

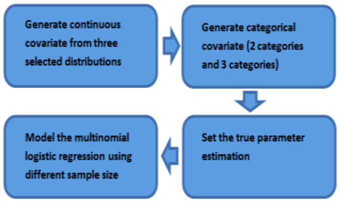
Hamzah Abdul Hamid[a,c*], Yap Bee Wah[a], Xian-Jin Xie[b]

[a]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[b]Department of Clinical Sciences & Simmons Cancer Center, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, Texas, USA
[c]Instute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia

Article history
Received
26 November 2015
Received in revised form
14 January 2016
Accepted
10 October 2016

*Corresponding author
amz_bst@yahoo.com

Graphical abstract


Generate continuous covariate from three selected distributions
Generate categorical covariate (2 categories and 3 categories)
Set the true parameter estimation
Model the multinomial logistic regression using different sample size

Abstract

The sample size and distributions of covariate may affect many statistical modeling techniques. This paper investigates the effects of sample size and data distribution on parameter estimates for multinomial logistic regression. A simulation study was conducted for different distributions (symmetric normal, positively skewed, negatively skewed) for the continuous covariates. In addition, we simulate categorical covariates to investigate their effects on parameter estimation for the multinomial logistic regression model. The simulation results show that the effect of skewed and categorical covariate reduces as sample size increases. The parameter estimates for normal distribution covariate apparently are less affected by sample size. For multinomial logistic regression model with a single covariate study, a sample size of at least 300 is required to obtain unbiased estimates when the covariate is positively skewed or is a categorical covariate. A much larger sample size is required when covariates are negatively skewed.

Keywords: Parameter estimation, simulation, multinomial logistic regression, skewed covariate

1.0 INTRODUCTION

Nominal or unordered response data is commonly found in many research areas [1]. Logistic regression is often used to model a categorical response variable. Binary logistic regression is used when the outcome variable has two categories while multinomial logistic regression is used when the outcome variable has more than two categories. The predictor variable or covariate can be either continuous or categorical [2-3]. The term covariate in this study refers to the independent variable in a statistical model. The covariate can be a continuous or categorical variable.

Logistic regression is widely used because it does not require the assumptions of normality, linearity or homoscedasticity of covariates. An alternative to logistic regression is discriminant function analysis. However, this method is not widely used as it requires the assumptions of multivariate normality. The residuals of a logistic regression model are not normally distributed and variance is not constant as the dependent variable is a binary or polytomous variable. Prabhakar *et al.* [4] in their study on hyperspectral remote sensing of yellow mosaic severity and associated pigment losses in Vignamungo applied multinomial logistic regression technique to build disease prediction models. They used optimal spectral reflectance ratios as an independent variable and found that the model has great potential to be used in prediction of the disease grades. The multinomial logistic regression was used by Venkataraman and Uddameri [5] to model simultaneous exceedance of drinking-water standards of arsenic and nitrate in the Southern Ogallala. They used binary logistic regressions to model separately exceedance and non-exceedance of nitrate and arsenic. Then, they used multinomial logistic regression to model all combinations of exceedance and non-exceedance of nitrate and arsenic. They reported that multinomial logistic regression model has good accuracy and correct prediction compared to separate binary logistic regression model. Varga *et al.* [6] evaluated risk factors for endemic human Salmonella Enteritidis (SE) infections with different phage types (PT) in Ontario, Canada by using multinomial logistic regression and case-case study approach. They considered three types of phage which are SE PT8, PT13a and non-PT8/non-PT13a as dependent variable and set the non-PT8/non-PT13a as a reference category. They found that there is a positive relationship between SE PT8 and contact with dog while negative relationship with pepper consumption and concluded that multinomial logistic regression is a novel method to model relationship between different PTs of SE infections and risk factors.

The method of parameter estimation used in logistic regression is different from ordinary linear regression. This is because the unordered response variable is not a continuous variable and thus the relationship between the categorical outcome and predictor variable will not be linear. Thus, the maximum likelihood parameter estimation method is used for binary and multinomial logistic regression models.

It is well known that most statistical procedures are affected by the distribution of covariates. The normality assumption is often required in many statistical techniques [7–10]. Hamid *et al.* [11] investigated the effects of covariate distribution and sample size on parameter estimation for binary logistic regression via simulation study. Three types of distribution: N (0,1), Beta (4,2) and U (-3,3) were simulated. They found that the parameter estimates for logistic regression model are affected by covariate distribution and sample size. This paper extends the simulation study by Hamid *et al.* [11] by considering the multinomial logistic regression model. The multinomial logistic regression model is a more complex model as the outcome variables can have more than two categories.

2.0 METHODOLOGY

2.1 The Multinomial Logistic Regression Model

To develop a multinomial logistic regression model, assume that *Y* is an outcome variable with c possible value (0, 1, … ,c-1) and let *Y*=0 be the reference category. Let x=$\left(x_1, x_2, ..., x_p\right)$ be the independent predictor variables. Thus, the conditional probabilities of each outcome category can be expressed as [12]:

$$P\left(Y = 0 \mid \mathrm{x}\right) = \frac{1}{1+\mathrm{e}^{\mathrm{g}_1(x)+...+\mathrm{g}_{\mathrm{c}\text{-}1}(x)}} \qquad (1)$$

$$P\left(Y = 1 \mid \mathrm{x}\right) = \frac{\mathrm{e}^{\mathrm{g}_1(x)}}{1+\mathrm{e}^{\mathrm{g}_1(x)+...+\mathrm{g}_{\mathrm{c}\text{-}1}(x)}} \qquad (2)$$

$$P\left(Y = \mathrm{c}\text{-}1 \mid \mathrm{x}\right) = \frac{\mathrm{e}^{\mathrm{g}_{\mathrm{c}\text{-}1}(x)}}{1+\mathrm{e}^{\mathrm{g}_1(x)+...+\mathrm{g}_{\mathrm{c}\text{-}1}(x)}} \qquad (3)$$

It follows that the logit function of category *j* is

$$g_j(x) = \ln\left[\frac{\mathrm{P}(\mathrm{Y} = \mathrm{j} \mid \mathrm{x})}{\mathrm{P}(\mathrm{Y} = 0 \mid \mathrm{x})}\right] = \beta_{\mathrm{j}0} + \beta_{\mathrm{j}1}\mathrm{x}_1 + ... + \beta_{\mathrm{jp}} x_p \qquad (4)$$

for*j*=1, 2,…,*c*-1.

2.2 The Maximum Likelihood Parameter Estimation

Let the outcome variable *Y* has three possible outcomes, *j*=0,1,2. To construct the likelihood function, three binary variables are created and coded as 0 and 1 to represent the group of membership of an observation. The variable Y_j is coded as 1 if *Y*=*j* while other categories are coded as 0. Therefore, the sum of these variables is $\Sigma_{j=0}^{2} Y_j = 1$. The conditional likelihood function for a sample of *n* independent observation can be written as [3]:

$$l(\beta) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \left[\pi_0(x_i)^{y_{0i}}\, \pi_1(x_i)^{y_{1i}}\, \pi_2(x_i)^{y_{2i}}\right] \qquad (5)$$

By using the fact that $\Sigma_{j=1}^{2}\Sigma_{i=1}^{n} y_{ji} = 1$ for each i of (5), the log-likelihood function is

$$L(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} y_{1i} g_1(x_i) + y_{2i} g_2(x_i) \text{-} \ln\left(1 + \mathrm{e}^{\mathrm{g}_1(x_i)} + \mathrm{e}^{\mathrm{g}_2(x_i)}\right) \qquad (6)$$

The first partial derivatives of $L(\beta)$ with respect to each of the 2(*p*+1) unknown parameters are used to obtain the likelihood equations. Let $\pi_{ji} = \pi_j(x_i)$, then the general form of likelihood equation is

$$\frac{\partial L(\beta)}{\partial \beta_{jk}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} x_{ki}\left(y_{ji} \text{ - } \pi_{\mathrm{ji}}\right) \qquad (7)$$

for j=1,2 and *k*=0,1,2,…,p, with $x_{0i} = 1$ for each subject. To obtain the maximum likelihood estimator $\hat{\beta}$, (7) is set equal to zero to solve for β [3]. The solution to obtain the MLE estimates requires Newton-Raphson iterative method. This efficient method is based on the idea of linear approximation.

2.3 Simulation Procedures

The effects of covariate distribution and sample size in estimating the multinomial logistic regression parameters were evaluated using simulation study. The sample sizes of 50, 100, 150, 300, 500, 1000, 1500, 3000 and 5000 were considered to represent small to large sample. The simulation procedure was carried out using R, open source programming software. The simulation involves 10,000 replications. The data were generated by using the same technique introduced by Fagerland *et al.* [12].

- The value of x is generated from the stated distribution.
- Evaluate the multinomial logistic probabilities for each category (π_0, π_1, π_2).
- The value of *u* is generated by using an independent U(0,1) distribution.
- Assign outcome variable (*Y*) based on the rule (i) y=2 if $u > \pi_0 + \pi_1$, (ii) y=1 if $u < \pi_0 + \pi_1$ and $u > \pi_0$ and (iii) y=0 otherwise.

The distribution of covariate and true parameters value for the multinomial logistic regression model are presented in Table 1. The distribution of N (0,1) is selected to represent symmetric distribution while Beta(12,1) represent negative skewed and χ^2 (4) represent positive skewed distributions. The categorical data were generated by using Binomial (1/2) and Binomial (1/3) distribution. Binomial (1/2) represent a binary dependent variable while Binomial (1/3) represent dependent variable with three categories.

Table 1 Distributions of Covariate and True Logistic Regression Coefficient

Setting	Covariate distribution	Skewness	Kurtosis	Coefficient
1	N(0,1)	0.000	2.996	β_{10} = -2.10 β_{11} = -0.35 β_{20} = -1.90 β_{21} = -0.21
2	Beta(12,1)	-1.577	6.108	β_{10} = -2.10 β_{11} = -0.35 β_{20} = -1.90 β_{21} = -0.21
3	χ^2 (4)	1.405	5.931	β_{10} = -2.10 β_{11} = -0.35 β_{20} = -1.90 β_{21} = -0.21
4	Binomial(1/2) (2 categories)	-	-	β_{10} = -2.10 β_{11} = -0.35 β_{20} = -1.90 β_{21} = -0.21
5	Binomial(1/3) (3 categories)	-	-	β_{10} = -2.10 β_{11} = -0.35 β_{12} = 1.08 β_{20} = -1.90 β_{21} = -0.21 β_{22} = 2.00

3.0 RESULTS AND DISCUSSION

This section presents the simulation results. The performance of maximum likelihood parameter estimation method is evaluated by considering different types of covariate and sample size.

3.1 Continuous Covariate

In this study, we considered three continuous distributions as tested by Hamid *et al.* [11]. The distributions N (0,1), Beta (12,1) and χ^2 (4) were chosen to represent symmetric, negatively skewed and positively skewed distribution. Table 2 summarizes the results of parameter estimates for different distribution for different sample size. The parameter estimates $\hat{\beta}$ does not deviate far from the true parameter value for symmetric normal covariate for all small sample sizes. Interestingly, the parameter estimates were more severely affected the model with negatively skewed covariate. The estimation of parameter improves and can be considered close to the true parameter value at sample size of 300 and above for the model with positively skewed covariate while the model with negatively skewed covariate needs larger sample size.

Table 2 Parameter Estimates for Different Distribution

Sample size	Model	β_{10} = **-2.10**	β_{11} = **-0.35**	β_{20} = **-1.90**	β_{21} = **-0.21**
50	A	-1.903	-0.405	-1.789	-0.240
	B	-3.423	1.551	-3.313	1.530
	C	-0.688	-0.480	-1.069	-0.297
100	A	-2.182	-0.375	-1.968	-0.223
	B	-3.626	1.270	-3.040	0.936
	C	-1.455	-0.452	-1.687	-0.272
150	A	-2.167	-0.367	-1.946	-0.221
	B	-3.226	0.775	-2.626	0.509
	C	-1.797	-0.429	-1.851	-0.254
300	A	-2.133	-0.358	-1.924	-0.215
	B	-2.660	0.208	-2.245	0.131
	C	-2.062	-0.400	-1.892	-0.230
500	A	-2.118	-0.355	-1.915	-0.214
	B	-2.354	-0.102	-2.163	0.056
	C	-2.093	-0.375	-1.895	-0.222
1000	A	-2.110	-0.354	-1.907	-0.210
	B	-2.254	-0.197	-2.007	-0.102
	C	-2.097	-0.362	-1.900	-0.216
1500	A	-2.105	-0.349	-1.905	-0.210
	B	-2.188	-0.263	-1.941	-0.171
	C	-2.093	-0.359	-1.898	-0.214
3000	A	-2.103	-0.351	-1.902	-0.210
	B	-2.143	-0.307	-1.934	-0.178

Sample size	Model	β_{10} = **-2.10**	β_{11} = **-0.35**	β_{20} = **-1.90**	β_{21} = **-0.21**
	C	-2.100	-0.354	-1.900	-0.212
5000	A	-2.102	-0.350	-1.901	-0.210
	B	-2.126	-0.324	-1.905	-0.207
	C	-2.101	-0.352	-1.900	-0.211

[a]Model A-N(0,1); Model B-Beta(12,1); Model C- χ^2 (4)

Figure 1 and Figure 2 show the box-plots of the parameter estimates for symmetric normal covariate at different sample size. The dispersion (standard deviation) of parameter estimates decreases as sample size increases. The value of parameter estimates get closer to the true parameter value at large sample size.

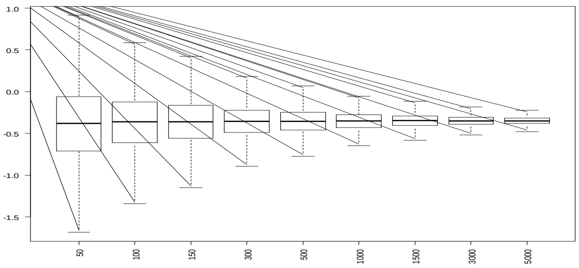


Figure 1 Box-plots of parameter estimates $\left(\hat{\beta}_{11}\right)$ for symmetric normal covariate

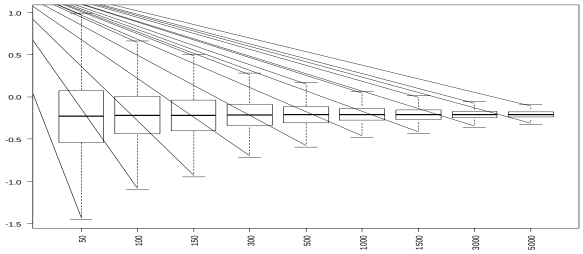


Figure 2 Box-plots of parameter estimates $\left(\hat{\beta}_{21}\right)$ for symmetric normal covariate

Figure 3 and Figure 4 show the box-plots of parameter estimates for negatively skewed covariate at different sample size. The parameter estimation improves when the sample size increases. Figure 5 and Figure 6 show the box-plot of the parameter estimates for positively skewed distribution covariate at different sample size. The estimates improve as sample size increases.

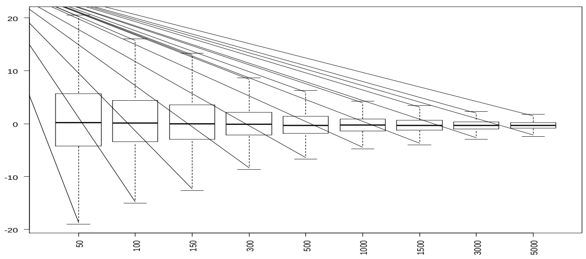


Figure 3 Box-plots of parameter estimates $\left(\hat{\beta}_{11}\right)$ for negatively skewed covariate

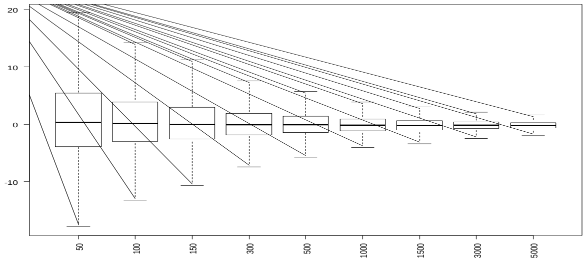


Figure 4 Box-plots of parameter estimates $\left(\hat{\beta}_{21}\right)$for negatively skewed covariate

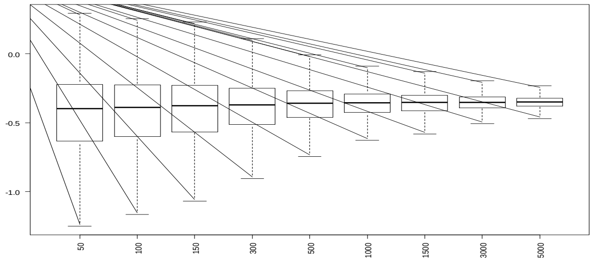


Figure 5 Box-plots of parameter estimates $\left(\hat{\beta}_{11}\right)$for positively skewed covariate

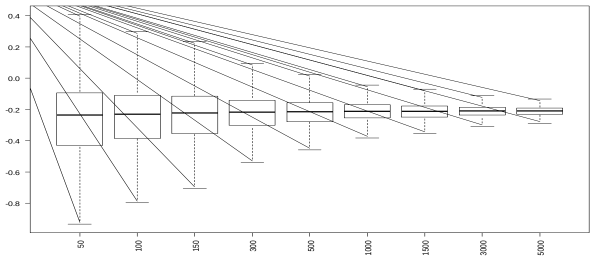


Figure 6 Box-plots of parameter estimates $\left(\hat{\beta}_{21}\right)$for positively skewed covariate

In addition, we present the combined box-plots of parameter estimates for all three distributions for sample size 50 to 500. These sample sizes were selected because the parameter estimation is highly affected in this range of sample size. Figure 7 and Figure 8 show the box-plots of parameter estimates β_{11} and β_{21}. It is clearly shown that the dispersion (standard deviation) of parameter estimates is very much higher for negatively skewed compared to symmetric normal and positively skewed covariate.

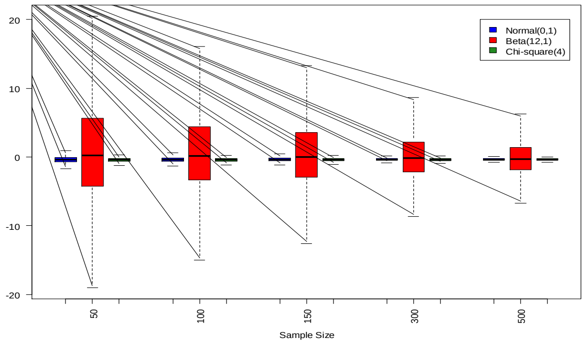


Figure 7 Box-plots of parameter estimates $(\hat{\beta}_{11})$ for continuous covariates

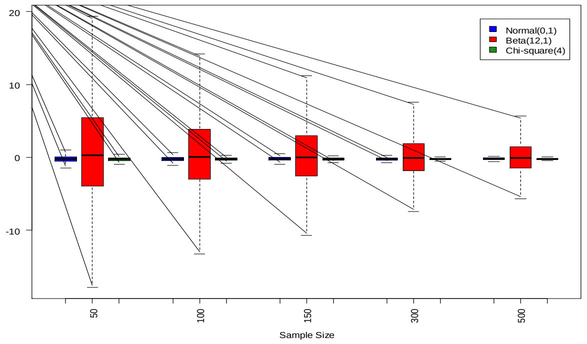


Figure 8 Box-plots of parameter estimates $(\hat{\beta}_{21})$ for continuous covariates

3.2 Categorical Covariate

To investigate the effect of categorical covariate on parameter estimation, we then generate Binomial distribution with *p*=1/2 for binary covariate and Binomial distribution with *p*=1/3 for a covariate with three categories. Table 3 summarizes the results for binary covariate. The parameter estimates, $\hat{\beta}$ are far from the true parameter value at small sample size (*n*=50). The parameter estimates started to get closer to the true parameter value at sample size of 300 and above.

Table 3 Parameter Estimates (Binary Covariate)

Sample size	Binomial(1/2)			
	β_{10} = **-2.10**	β_{11} = **-0.35**	β_{20} = **-1.9**	β_{21} = **-0.21**
50	-1.815	-0.621	-1.811	-0.340
100	-2.148	-0.509	-1.973	-0.265
150	-2.167	-0.416	-1.946	-0.228
300	-2.138	-0.359	-1.927	-0.214
500	-2.121	-0.355	-1.916	-0.211
1000	-2.109	-0.356	-1.908	-0.209
1500	-2.106	-0.354	-1.905	-0.211
3000	-2.102	-0.352	-1.902	-0.212
5000	-2.102	-0.350	-1.901	-0.211

Figure 9 and Figure 10 show the box-plot of the parameter estimates for binary covariate for different sample sizes. Similar patterns were observed, whereby the dispersion (standard deviation) decreases and parameter estimation improves when sample size increases.

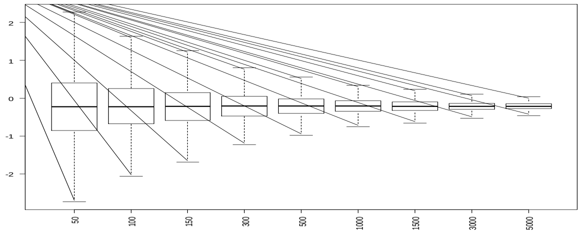


Figure 9 Box-plots of parameter estimates $(\hat{\beta}_{11})$ for binary covariate

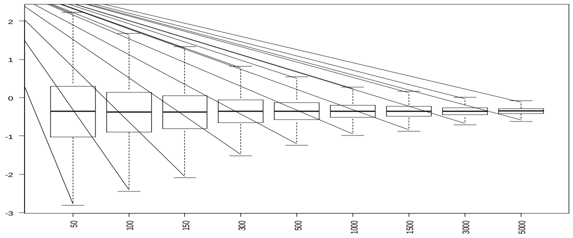


Figure 10 Box-plots of parameter estimates $(\hat{\beta}_{21})$ for binary covariate

The simulation results for a categorical covariate with three categories are summarized in Table 4. The results were consistent with the results achieved for the model with binary covariate. The parameter estimates $\hat{\beta}$ are also far from the true parameter value at small sample size (n=50) and started to get close to the true parameter value at sample size of 300 and above.

Table 4 Parameter Estimates (Covariate with 3 Categories)

Sample size	Binomial(1/3)					
	β_{10} = **-2.10**	β_{11} = **-0.35**	β_{12} = **1.08**	β_{20} = **-1.90**	β_{21} = **-0.21**	β_{22} = **2.00**
50	-2.039	-0.866	-0.296	-2.298	-0.563	3.001

Sample size	Binomial(1/3)					
	β_{10} = **-2.10**	β_{11} = **-0.35**	β_{12} = **1.08**	β_{20} = **-1.90**	β_{21} = **-0.21**	β_{22} = **2.00**
100	-2.231	-0.648	-0.210	-2.030	-0.287	2.183
150	-2.185	-0.456	0.376	-1.954	-0.237	2.068
300	-2.144	-0.360	0.987	-1.930	-0.213	2.035
500	-2.122	-0.357	1.052	-1.917	-0.213	2.019
1000	-2.109	-0.358	1.067	-1.909	-0.209	2.010
1500	-2.108	-0.353	1.071	-1.905	-0.211	2.007
3000	-2.103	-0.352	1.077	-1.902	-0.212	2.001
5000	-2.103	-0.350	1.081	-1.901	-0.211	2.003

Figure 11 to Figure 14 shows the box-plots of the parameter estimates. The precision of the estimates increases as sample size increases.

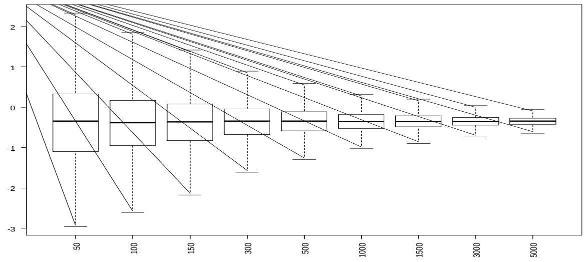


Figure 11 Box-plots of parameter estimates $\left(\hat{\beta}_{11}\right)$

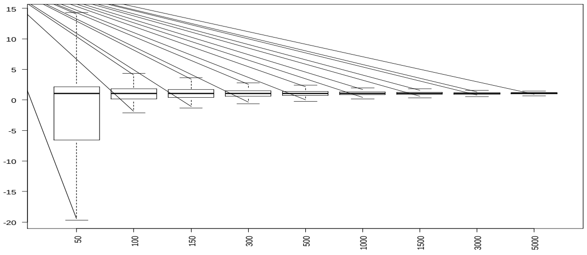


Figure 12 Box-plots of parameter estimates $\left(\hat{\beta}_{12}\right)$

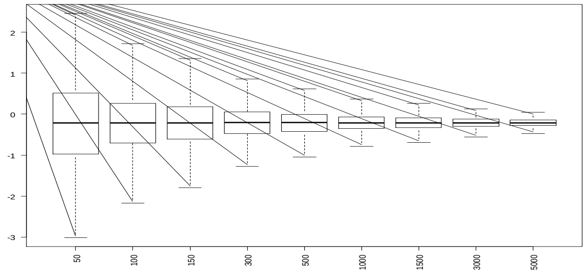


Figure 13 Box-plots of parameter estimates $\left(\hat{\beta}_{21}\right)$

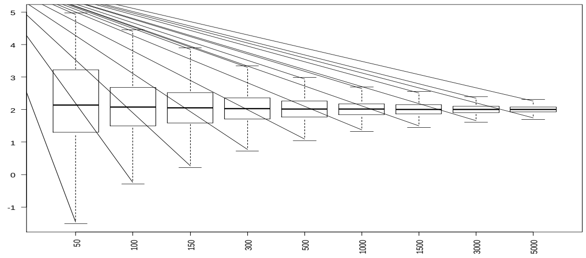


Figure 14 Box-plots of parameter estimates $\left(\hat{\beta}_{22}\right)$

4.0 CONCLUSIONS

Although the multivariate logistic regression model does not require the assumption of normality, this simulation study results show that the parameter estimates of a multinomial logistic regression model is affected when data is not normal. The parameter estimates are more severely biased when distribution of the covariate is negatively skewed compared to covariate which is positively skewed. In addition, we also provide the results for categorical covariate. We found that small sample size below 300 produced biased parameter estimates. The parameter estimates for model with positively skewed and categorical covariates approach the true parameter value at sample size of 300 and above. However, a larger sample size is required for negatively skewed covariates. These results are consistent with simulation study by Hamid *et al.* [11] for binary logistic regression. Further simulation is in progress to determine the sample size cut-off when the logistic regression model has more covariates. Results of this simulation study confirm that distribution of covariates and sample size obviously plays a very important role in obtaining reliable parameter estimates.

Acknowledgement

The authors would like to acknowledge the financial support received from the Ministry of Higher Education of Malaysia and Tabung Amanah Pembangunan Pelajar (TAPA) Universiti Teknologi MARA Malaysia.

References

[1] Hedeker, D. 2003. A Mixed-Effects Multinomial Logistic Regression Model. *Stat. Med*. 22(9):1433-1446.

[2] Kutner, M. H., Nachtsheim, C. J., and Neter, J. 2004. *Applied Linear Regression Models.* 4th Edition. McGraw-Hill/Irwin.

[3] Hosmer, D. Jr., and Lemeshow, S. 2004. *Applied Logistic Regression*. John Wiley & Sons; 2nd edition.

[4] Prabhakar, M., Prasad, Y. G., Desai S., Thirupathi, M., Gopika, K., Rao, G. R., and Venkateswarlu, B. 2013. Hyperspectral Remote Sensing Of Yellow Mosaic Severity And Associated Pigment Losses In Vignamungo Using

Multinomial Logistic Regression Models. *Crop Prot.,* 45(2013): 132-140.

[5] Venkataraman, K. and Uddameri, V. 2012. Modeling Simultaneous Exceedance Of Drinking-Water Standards Of Arsenic And Nitrate In The Southern Ogallala Aquifer Using Multinomial Logistic Regression. *J. Hydrol*. 458(2012): 16-27.

[6] Varga, C., Middleton, D., Walton, R., Savage, R., Tighe, M.-K., Allen, V., Ahmed, R. and Rosella, L. 2012. Evaluating Risk Factors For Endemic Human Salmonella Enteritidis Infections With Different Phage Types In Ontario, Canada Using Multinomial Logistic Regression And A Case-Case Study Approach. *BMC Public Health*. 12(1): 866.

[7] Erceg-Hurn, D. M. and Mirosevich, V. M. 2008. Modern Robust Statistical Methods: An Easy Way To Maximize The Accuracy And Power Of Your Research. *Am. Psychol*. 63(7): 591-601.

[8] Jahan S. and Khan, A. 2012. Power Of T-Test For Simple Linear Regression Model With Non-Normal Error Distribution: A Quantile Function Distribution Approach. *J. Sci. Res*. 4(3): 609-622.

[9] Khan A. and Rayner G. 2003. Robustness To Non-Normality Of Common Tests For The Many-Sample Location Problem. *Journal of Applied Mathematics and Decision Sciences*. 7(4): 187-206.

[10] Curran, P. J., West, S. G. and Finch, J. F. 1996. The Robustness Of Test Statistics To Nonnormality And Specification Error In Confirmatory Factor Analysis. Psychological Methods. *American Psychological Association, Inc*. 1(1): 16-29.

[11] Hamid, H. A., Wah, Y. B., Xie, X.-J. and Rahman, H. A. A. 2015. Assessing The Effects Of Different Types Of Covariates For Binary Logistic Regression. *The 2nd ISM International Statistical Conference 2014 (ISM-II): Empowering the Applications of Statistical and Mathematical Sciences.* AIP Publishing. 425(2015): 425-430.

[12] Fagerland, M., Hosmer, D. and Bofin, A. 2008. Multinomial Goodness-Of-Fit Tests For Logistic Regression Models. *Statist. Med*. 27(21): 4238-4253.

[13] Stokes, M. E., Davis, C. S. and Koch, G. G. 2009. *Categorical Data Analysis Using the SAS System*. 2nd edition. SAS Institute.

[14] Hosmer, D. W. and Lemesbow, S. 1980. Goodness Of Fit Tests For The Multiple Logistic Regression Model. *Commun. Stat.-Theory Methods.* 9(10): 1043-1069.

[15] Xie, X.-J., Pendergast, J. and Clarke, W. 2008. Increasing The Power: A Practical Approach To Goodness-Of-Fit Test For Logistic Regression Models With Continuous Predictors. *Comput. Stat. Data Anal.* 52(5): 2703-2713.

[16] Motrenko, A., Strijov, V. and Weber, G.-W. 2014. Sample Size Determination For Logistic Regression. *J. Comput. Appl. Math*. 255(2014): 743-752.

[17] Fishman, G. 1971. Estimating Sample Size In Computing Simulation Experiments. *Manage. Sci.* 18(1): 21-38.

[18] Ancel, P. Y. 1999. Value Of Multinomial Model In Epidemiology: Application To The Comparison Of Risk Factors For Severely And Moderately Preterm Births. *Rev. Epidemiol. Sante Publique*. 47(6): 563-9.

[19] Cramer, J. S. 2002. The Origins of Logistic Regression. Tinbergen Inst. Work. Pap. 2002-119/4.