International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:4, No:2, 2010

Dichotomous Logistic Regression with Leave-One-Out Validation

Sin Yin Teh, Abdul Rahman Othman, and Michael Boon Chong Khoo

***Abstract*—**In this paper, the concepts of dichotomous logistic regression (DLR) with leave-one-out (L-O-O) were discussed. To illustrate this, the L-O-O was run to determine the importance of the simulation conditions for robust test of spread procedures with good Type I error rates. The resultant model was then evaluated. The discussions included 1) assessment of the accuracy of the model, and 2) parameter estimates. These were presented and illustrated by modeling the relationship between the dichotomous dependent variable (Type I error rates) with a set of independent variables (the simulation conditions). The base SAS software containing PROC LOGISTIC and DATA step functions can be making used to do the DLR analysis.

***Keywords*—**Dichotomous logistic regression, leave-one-out, test of spread.

I. Introduction

EARLY uses of logistic regression were in biomedical studies, but in recent years have also seen much use in business applications, social science research, marketing, and genetics [1-3]. Although logistic regression has gained popularity, there remains considerable confusion about its use and interpretation [4-5]. In short, the literature seems to cover theoretical and mathematical issues related to logistic regression more thoroughly than the practical and applied aspects needed to put this technique to use [6].

Dichotomous logistic regression (DLR) is a common type of generalized linear model that utilizes the logit as its link function [1]. This particular regression enables us to investigate the relationship between a categorical outcome and a set of independent variables. The independent variables can be of any form. DLR does not assume linearity of relationship between the dependent and the independent variables, does not require normally distributed independent variables, and does not assume homoscedasticity. However, it does require that observations to be independent and that the independent variables be linearly related to the *logit* of the dependent. Thus, logistic regression can be used to predict a dependent variable on the basis of continuous and/or categorical independents; to determine the percentage of variance in the dependent variable which has been explained by the independents; to rank the relative importance of independent variables; to assess interaction effects and to understand the impact of covariate control variables [7-8].

The L-O-O classification approach which does not require the assumption of normality was then used to assess the accuracy of the DLR model. This is because the data is made up of categorical independent variables and the normality assumption is violated. Model validation with the L-O-O method produces the highest accuracy estimates for the classification problems due to its capability to process almost all of the available data for training the classifier [9].

In this paper, the concepts of DLR were discussed. And, an illustration to illustrated DLR as one of the data mining techniques were performed to determine the importance of the simulation conditions for robust test of spread procedures on the generating of *p*-values. The discussions included 1) assessment of the accuracy of the model, and 2) parameter estimates. These were presented and illustrated by model the relationship between the dichotomous dependent variable (Type I error rates) with a set of independent variables (the simulation conditions).

II. Probability, DLR Model, Odds, and Logit

The logistic model describes the expected value of Y (i.e., $E(Y)$) in terms of the following "logistic" formula:

$$E(Y|X_i) = \frac{1}{1 + exp\left[-\beta_0 - \sum_{j=1}^{k} \beta_j X_{ij}\right]} \quad (1)$$

where

β_0 = the intercept parameter,

β_j = a vector of t regression parameters, and

X_{ij} = row vector of independent variables corresponding to the j^{th}subpopulation.

For a random variable with values 0 or 1 that

$$E(Y|X_i) = [0 \times P(Y_i = 0) + 1 \times P(Y_i = 1)] = P(Y_i = 1) \quad (2)$$

where

$P(Y_i = 0)$ = **probability** of the event which coded with 0 (failure), and

$P(Y_i = 1)$ = **probability** of the event which coded with 1 (success).

The formula of **DLR model** can be written in a form that describes the variation among probabilities as follows:

$$P(Y_i) = \frac{1}{1 + exp\left[-\beta_0 - \sum_{j=1}^{k} \beta_j X_{ij}\right]} \quad (3)$$

The **odds** of success for the j^{th} group of some event i is defined as the ratio probability of success to the probability of failure i.e.

$$odds_i = \frac{P_i}{1 - P_i} = exp\left[\beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}\right]. \quad (4)$$

In DLR, the dependent variable is a *logit*, which is the natural logarithm of the odds. That is by taking logs on both sides of Equation 4, a linear DLR model for the **logit** were obtain:

$$log(odds_i) = logit(P_i) = ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \quad (5)$$

where

P_i = the predicted probability of the event which coded with 1, and

X_{ij} = independent variables, $i = 1, 2, \ldots, n$.

This is the log odds of success to failure for the j^{th} subpopulation. The logit transformation here is $ln\left(P_i/1 - P_i\right)$. The main reason for using the *logit* form of output is to prevent the predicting probabilities P_i from going out of range, where the required range for P_i is [0,1]. The *logit*(P_i) is assumed to be linear, that means the *log odds* is assumed to be linearly related to X_{ij}.

DLR applies maximum likelihood estimation after transforming the dependent into a *logit* variable. Actually, the maximum likelihood methods are used to estimate β_0 and β_j. In this way, DLR estimates the probability of a certain event occurring. Note that DLR calculates the log odds of the dependent, not changes in the dependent itself. The success of the DLR can be assessed by looking at the classification table which tabulates the correct and incorrect classifications of dichotomous dependent. Also, goodness-of-fit tests such model chi-square is available as an indicator of model appropriateness and statistic, the Wald statistic can be used to test the significance of individual independent variables.

III. Assessment of Model: Fitting

The statistic used to assess the overall fit of the model is based on the likelihood function. The null and the alternative hypotheses for assessing overall model fit are given by

H_0 : The hypothesized model fits the data.

H_A : The hypothesized model does not fits the data.

Obviously, non rejection of the null is desired, as it leads to the conclusion that the model fits the data.

The test statistic for this hypothesis is the likelihood ratio test. The likelihood, *L*, of a model is defined as the probability that estimated hypothesized model represents the input data. To test the null and alternative hypotheses, *L* is transformed to *-2LogL*. The *-2LogL* statistic is referred to as the likelihood ratio. It has a χ^2 distribution with *n-q* degrees of freedom where *q* is the number of parameters in the full model [7-8, 10]. The output of likelihood ratio test provides two *-2LogL* statistics, one for a model that includes only the intercept and another includes intercept and covariates. Deviance is the difference between two log-likelihood values. In comparing a null model (L_{null}) with only the intercept and a model (L_{model}) including intercept and *k* parameters, then the deviance is the difference between $-2LogL_{null} - (-2LogL_{model})$[11]. The smaller the deviance, the better the model fits the data.

The deviance for a large sample given by

$$G_0^2 = -2\ln\left(\frac{L_{null}}{L_{model}}\right) = -2LogL_{null} - (-2LogL_{model}) \quad (6)$$

has a chi-square distribution with *k* degrees of freedom, where L_{null} and L_{model} refer to the likelihood of the null and full models, respectively. This means that the likelihood ratio test was used to compare the likelihood of the full model (i.e. with all the predictors included) with the likelihood of the null model (i.e. a model which contained only the intercept). This is analogous to the overall *F*-test of the model in linear regressions.

IV. Percent of Correct Classification

In any classification method, the percentage of correct classification is the primary indicator of goodness of the method. Classification table (or confusion matrix) is used to show the ability to predict correctly the outcome category (dichotomous dependent variable) for all cases by using 2×2 tables. It shows all correct and incorrect estimates. In fact, the classification table is used to determine the error rate of the model, which is an evaluation measure of the model's predictive performance. Classification of observations is done by first estimating the probabilities, $\hat{P} = P$(each observation belonging to a given group). Table I presented a confusion matrix with a dependent variable with two categories (0 or 1). The columns in the table are the two predicted values of the dependent, while the rows are the two observed values of the dependent. Each cell contains the number of correct/incorrect predictions as the following:

TN = the number of correct predictions that an instance is zero;

FP = the number of incorrect predictions that an instance is one;

FN = the number of incorrect predictions that an instance is zero; and

TP = the number of correct predictions that an instance is one.

The hit ratio or percent of correct classification (PCC) is determined using the equation:

$$\text{Hit ratio} = \frac{TN + TP}{TN + FP + FN + TP} \quad (7)$$

Sensitivity is the ability to predict an event correctly. It is the proportion of observed event responses that were predicted to be events. Specificity is the ability to predict a non-event correctly. It is the proportion of observed non-event responses that were predicted to be non-events. The equations of sensitivity and specificity were:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

TABLE I
CLASSIFICATION TABLE

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

In a perfect model, all cases will be on the diagonal of Table I and the PCC, sensitivity, and specificity will be 100%. Classification of the observations into these groups is based on a cutoff value for $\hat{p}$, which is usually assumed to be 0.5. All observations where $\hat{p}$ is greater than or equal to 0.5 are classified as events and values which are less than 0.5 are classified as non-events. If the observed sample has prior probability of belongs to group 0 is large and the sample has prior probability of belongs to group 1 is small, and vice versa, then 0.5 is not the right cut off value. The cut off were now depends on the sample proportion of group 1. The classification table and the classification rates reported by Statistical Analysis Software (SAS) program are obtained by using the pseudo-jackknife estimation procedure.

V. LEAVE-ONE-OUT CLASSIFICATION

The data is made up of categorical independent variables; hence the normality assumption is violated. Therefore, the L-O-O approach which does not require the assumption of normality is used. The Jackknife-like method also known as the Lachenbruch's holdout is a widely used approach based on estimation with multiple subsets of the sample for validation [12-14]. The L-O-O method represents a special case of the cross-validation technique [15]. Given *n* cases available in a dataset, a classifier is trained on (*n*-1) cases and then is tested on the case that was left out [16-17]. This process is repeated *n* times until every case in the dataset have been included once as a cross-validation instance. The results are averaged across the *n* test cases to estimate the classifier's prediction performance [14]. Therefore, this method produces the highest accuracy estimates for the classification problems [9].

Most researchers suggest that L-O-O approach be used only when the smallest group size is at least five times the number of predictor variables [18]. One of the characteristics of the L-O-O method is that the outside test recognition rate should be able to approach the true recognition rate closely because each classifier uses almost all the data set leaving one entry.

VI. PARAMETER ESTIMATES AND IMPORTANCE OF PARAMETERS

The maximum likelihood estimates of parameters will be used. The coefficient of the independent variable gives the amount by which the dependent variable will increase or decrease if the independent variable changes by one unit. The square of the *t*-values give the Wald χ^2 statistic, which can be used to assess the statistical significance of each independent variable.

A. Wald Test

The test on individual coefficients is based on a *t*-like statistic referred to as the Wald inference [19]. A Wald test is used to test the statistical significance of each coefficient (β_j) in the model. The corresponding null and alternative hypotheses are

$H_0 : \beta_j = 0, \quad j = 0, 1, \ldots, k.$

$H_A : \beta_j \neq 0.$

The Wald test statistic

$$W = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \tag{10}$$

follows the standard normal distribution under the null hypothesis, $\beta_j = 0$. The statistic is essentially the same as the *t*-statistic in the linear model. Under the alternative hypothesis, it is asymptotic to χ^2 distribution and is calculated by

$$Wald\ \chi^2 = \left(\frac{\hat{\beta}}{s.e.(\hat{\beta})} \right)^2. \tag{11}$$

Though the Wald test is used by many, it is less powerful than the likelihood ratio test. This is because the Wald test is biased under certain situations. The Wald test often misleads the user to conclude that the coefficient (consequently the corresponding risk factor) is not significant when reality it indeed is [8]. Certainly, several authors have identified problems with the use of the Wald statistic. Menard [20] warns that for large coefficients, standard error is inflated, lowering the Wald χ^2 statistic value. Agresti [1] stated that the likelihood-ratio test is more reliable for small sample sizes than the Wald test. Therefore, this statistic needs to be interpreted with great caution. In this study, the Wald statistic was considered because it is computationally easy and is provided automatically in the output of most statistical computer packages, i.e. SAS.

B. P-Value

The *p*-value for each parameter estimate of $\hat{\beta}$ is the probability of obtaining a value of the test statistic as extreme as or more extreme (in the appropriate direction) than the one actually computed when the null hypothesis is true. The *p*-value (refer to Fig. 1) is given by:

$$p = P(V > V_0) \tag{12}$$

where

V : χ_v^2,

V_0 = calculated value of test statistic, and

v = degrees of freedom.

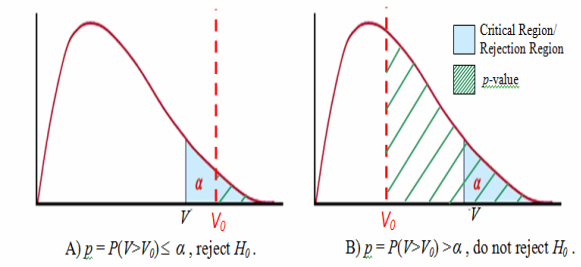

Fig. 1 The *p*-value of χ^2 distribution

The H_0 is rejected when $p \le \alpha$, where α is the level of significance. Thus, the *p*-value for a test can also be defined as the smallest value of α for which the null hypothesis can be rejected. In fact, when controlling the level of significance at $\alpha = 0.05$,

$p < 0.05$ reject H_0 (refer to Figure 1A);

$p \ge 0.05$ accept H_0 (refer to Figure 1B).

Note that in general the sample size must be large in order for the *p*-value to be accurate.

C. Odds Ratio

Quantification of the relationships of the predictors in the logistic model to the dependent variable involves a parameter called the odds ratio [21]. The odds ratio is the ratio of the odds (refer to Equation 4) of having an outcome for one group versus another, that is:

$$\begin{aligned} OR_{X_A \text{ vs } X_B} &= \frac{odds_{X_A}}{odds_{X_B}} \\ &= \frac{P_{X_A}}{1-P_{X_A}} \Big/ \frac{P_{X_B}}{1-P_{X_B}} \\ &= \frac{exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{Aj}\right)}{exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_{Bj}\right)} \\ &= exp\left(\sum_{j=1}^{k} \beta_j \left(X_{Aj} - X_{Bj}\right)\right). \end{aligned} \quad (13)$$

It is normally represented by *Exp(H)* or *Exp(Est)*, where $H = \sum_{j=1}^{k} \beta_j \left(X_{Aj} - X_{Bj}\right)$. Significant Wald values can only be interpreted by transforming the values into odds ratios using the exponential function. The odds ratio can be any non-negative number. When the *Exp(H)* has the value 1, it indicates that the sample is predicted to belong to the event and vice versa.

The odd ratios could also be interpreted by evaluating how the unit changes in *H*, affect *Exp(H)*. Suppose there is an example of lung cancer occurrences, and the purpose is to analyze the predictors of lung cancer, namely smoking status, and some other variable, e.g. age representing a continuous variable. Hence the dependent variable is dichotomous, with having lung cancer considered as an event, while not having lung cancer considered as non-event. The binary independent variable, smoking status has values smoker or non-smoker with non-smoker considered as the reference level.

For the binary independent variable, smoking status, keeping the other variable age constant, the odds ratio could be obtained. For example, X_A and X_B are two specifications of these two independent variables smoking status and age, say, $X_A = (1, 45)$ and $X_B = (-1, 45)$. Here, X_A denotes the group of 45-year-old smokers (smoking status = -1), whereas X_B denotes the group of 45-year-old non-smokers (smoking status = 1). Then, from Equation 13,

$$\begin{aligned} OR_{X_A \text{ vs } X_B} &= exp\left\{\left(X_{A_1} - X_{B_1}\right)\beta_1 + \left(X_{A_2} - X_{B_2}\right)\beta_2\right\} \\ &= exp\left\{\left(1-(-1)\right)\beta_1 + (45-45)\beta_2\right\} \\ &= e^{2\beta_1}. \end{aligned}$$

If the estimate of the $2\beta_1$ coefficient from maximum likelihood estimation turns out to be, say $2\hat{\beta}_1 = 2.303$ then the estimated odd ratio will be $e^{2.303} = 10$. This indicates that a smoker is ten times more likely to get lung cancer compared against a non-smoker.

Similarly, for the continuous independent variable age, keeping smoking status constant, the odds ratio could be obtained. For example, say, $X_A' = (-1, 45)$ and $X_B' = (-1, 21)$. Here, X_A' denotes the group of 45-year-old smokers, whereas X_B' denotes the group of 21-year-old smokers. Then, from Equation 13,

$$\begin{aligned} OR_{X_A \text{ vs } X_B} &= exp\left\{\left(X'_{A_1} - X'_{B_1}\right)\beta_1 + \left(X'_{A_2} - X'_{B_2}\right)\beta_2\right\} \\ &= exp\left\{\left((-1)-(-1)\right)\beta_1 + (45-21)\beta_2\right\} \\ &= e^{24\beta_2}. \end{aligned}$$

If the estimate of the $24\beta_2$ coefficient from maximum likelihood estimation turns out to be, say $24\hat{\beta}_2 = 0.152$ then the estimated odd ratio will be $e^{0.152} = 1.164$. This indicates that the odds of getting lung cancer increases by 16.4% with each increasing age (year) of a smoker.

D. Importance of Parameters

The importance of independent variables is determined by odds ratios and the *p*-values. Independent variables that have influence/importance are those with odds ratio larger than one or odds ratio less than one, with *p*-values significant (<0.05). An odds ratio greater than one means a non-reference level independent variable will be classify into the event group. An odds ratio less than one imply that the reference level independent variable will be classified into the event group.

TABLE II
CATEGORICAL INDEPENDENT VARIABLES AVAILABLE FOR ENTRY

Variable	Variable Label	Level	Level Label
DISTR	Type of distribution	BETA(0.5,0.5)	Symmetric platykurtic
		FLEISHMAN1	Skewed platykurtic
		FLEISHMAN2	Skewed normal-tailed
		G=.225/H=.225	Skewed leptokurtic (severe)
		G=.76/H=-.098	Skewed leptokurtic
		G=0/H=.225	Symmetric leptokurtic
		N(0,1)	Standard normal
SHAPE	Skewness of distribution	SKEW	Skewed
		SYMM	Symmetric
TAIL	Kurtosis of distribution	LEPT	Leptokurtic
		PLAT	Platykurtic
		NORM	Normal
GSIZE	Total group size	120	N=120
		60	N=60
GSCOND	Group size increments	INCR05	Increment of 5
		INCR10	Increment of 10
		EQUAL	Equal sample size

VII. DLR TO DETERMINE THE IMPORTANCE OF SIMULATION CONDITIONS FOR ROBUST TEST OF SPREAD PROCEDURES ON GENERATING OF *P*-VALUES

This paper performed an illustration of DLR as one of the data mining techniques to determine the importance of the simulation conditions for robust test of spread procedures on the generating of *p*-values. That is, DLR was conducted to evaluate the particular simulation conditions that will produce robust Type I error rates, i.e. Type I error rates that fall in [0.045, 0.050]. Essentially, the database consist all *p*-values & attendant information for tests of spread procedures from [22]. In particular, these procedures were compared for their Type I error rates when data were obtained from 7 different distributions within the context of 6 one-way independent groups' designs. The designs differed by total sample size & group sample sizes;

(a) degree of sample size inequality;
(b) shape of the population distribution; and
(c) values of trimming.

For each condition five thousand replications were conducted and the nominal level of significance was 0.05.

The simulation conditions in this study were types of distribution, skewness of distribution, kurtosis of distribution, total group size, and unbalanced group size increments (Refer to Table II). These were also the independent variables in the analysis. The 7 distributions simulated in [22] were used in this study, they were

1) The Fleishman [23] transformation of the standard normal distribution into a skewed platykurtic distribution with skewness, $\gamma_1 = 0.5$ and kurtosis, $\gamma_2 = -0.5$.

2) A second Fleishman transformation of the standard normal distribution into a skewed normal-tailed distribution with $\gamma_1 = 0.75$ and $\gamma_2 = 0$.

3) The Beta (0.5, 0.5) distribution representing symmetric platykurtic distributions with $\gamma_1 = 0$ and $\gamma_2 = -1.5$.

TABLE III
CATEGORICAL INDEPENDENT VARIABLES AVAILABLE FOR ENTRY

Model Information

Data Set	WORK.TRAINING
Response Variable	pval05
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	84
Number of Observations Used	84
Sum of Frequencies Read	25257
Sum of Frequencies Used	25257

Response Profile

Ordered Value	pval05	Total Frequency
1	1	1860
2	0	23397

Probability modeled is pval05=1.

TABLE IV
ASSESSMENT OF MODEL

Part A. Model Fit Statistics

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only [a]	Intercept and Covariates
-2 Log L	13283.252 [b]	12449.159

Part B. Likelihood Ratio Test

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	834.0929 [c]	9	<.0001

[a] The -2 Log L for Intercept Only ($-2\,LogL_{null}$) is defined below, where N_0 and N_1 are observed frequencies for the dichotomous dependent variable. $N_0 + N_1 = N$, total sample size.

$$-2\,LogL_{null} = -2\left[N_0 ln(N_0/N) + N_1 ln(N_1/N)\right]$$

[b] $-2LogL_{null} = -2\left[23397 ln(23397/25257) + 1860 ln(1860/25257)\right] = -2\left[-6641.626\right] = 13283.252$

[c] $G = (-2\,LogL_{null}) - (-2\,LogL_{model}) = 13283.252 - 12449.159 = 834.0929$

4) A g and h distribution [24] where $g = h = 0$. This is the standard normal distribution with $\gamma_1 = \gamma_2 = 0$.

5) A $g = 0$ and $h = 0.225$ long-tailed distribution with $\gamma_1 = 0$ and $\gamma_2 = 154.84$, representing symmetric leptokurtic distributions.

6) A $g = 0.76$ and $h = -0.098$ distribution with $\gamma_1 = 2$ and $\gamma_2 = 6$, representing skewed leptokurtic distribution.

7) A $g = 0.225$ and $h = 0.225$ distribution. This is also a long-tailed skewed leptokurtic distribution ($\gamma_1 = 4.9$, $\gamma_2 = 4673.8$), but more severe than (6).

The skewness of a distribution was either symmetric or skewed, while the kurtosis of distributions ranged from platykurtic to normal-tailed to leptokurtic distributions. The total sample group size was designed as 60 (average sample size of 20) or 120 (average sample size of 40). The unbalanced group size increments followed 3 conditions of sample size equality or inequality. These were equal sample sizes,

TABLE V
BIAS-ADJUSTED CLASSIFICATION TABLE

Classification Table							
Prob Level	Correct		Incorrect		Percentages		
	Event	Non-Event	Event	Non-Event	Correct	Sensitivity	Specificity
0.070	1290	13821	9576	570	59.8 [a]	69.4	59.1

[a] $\text{correct}\% = (\text{correct}_{\text{event}} + \text{correct}_{\text{non-event}})/N \times 100\% = (1290 + 13821)/25257 \times 100\% = 59.8\%$

TABLE VI
ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Analysis of Maximum Likelihood Estimates							
Parameter		DF	Estimate	Standard Error	Wald χ^2	Pr > ChiSq	Exp(Est)
INTERCEPT		1	-2.7557	0.0295	8745.0258	<.0001	0.064
DISTR	**BETA(0.5,0.5)**	1	0.4032	0.0575	49.2344	<.0001	1.497
DISTR	**FLEISHMAN1**	1	-0.5912	0.0817	52.3279	<.0001	0.554
DISTR	**FLEISHMAN2**	1	-0.9563	0.0958	99.6994	<.0001	0.384
DISTR	**G=.225/H=.225**	1	0.1100	0.0630	3.0517	0.0807	1.116
DISTR	**G=.76/H=-.098**	1	-0.5950	0.0823	52.2690	<.0001	0.552
DISTR	**G=0/H=.225**	1	0.5496	0.0548	100.4790	<.0001	1.733
SHAPE	**SKEW**	0	0	.	.	.	.
TAIL	**LEPT**	0	0	.	.	.	.
TAIL	**PLAT**	0	0	.	.	.	.
GSIZE	**120**	1	0.2210	0.0249	78.6714	<.0001	1.247
GSCOND	**INCR05**	1	0.1596	0.0338	22.2760	<.0001	1.173
GSCOND	**INCR10**	1	-0.1810	0.0362	25.0339	<.0001	0.834

increments of 5 (moderately unequal sample sizes), and increment of 10 (extremely unequal sample sizes). On the other hand, the dependent variable had two values, 1 representing *p*-values falling in [0.045, 0.050] and 0 for *p*-values falling outside of this interval after restructure it.

Originally, there were one scale dependent variable and five independent variables with 25,257 records. The five independent variables contain the information of 7 levels of types of distributions, 2 levels of skewness of distributions, 3 levels of kurtosis of distributions, 2 levels of total group size and 3 levels of group size increments. However, the preliminary run on DLR showed that with this particular variables structure, there were zero parameter estimates. This was a sign of presence of multicollinearity. However, this study still believes that the collinear variables are relevant to the model. Thus, the data was restructured by redefining the variables. Since the *SHAPE* and *TAIL* were fixed in the distributions. The combination of the independent variables formed 42 combinations of levels after restructure the original independent variables. For each of the 42 combinations, the number of records in group 0 and group 1 were counted for *PVAL05* (dependent). Hence, there were 42 combination of multiplied by 2 levels of *PVAL05* equaling 84 records. The total number of counts for the 84 records will be 25,257. Basically, this was the number of records before the data restructured. Using these combination, DLR was performed to examine relationship between the dichotomous dependent variable (Type I error rates) with a set of independent variables (the simulation conditions). The restructured variables are given in Appendix A.

A. Dichotomous Logistic Regression (DLR)

The DLR model estimated was

$$PVAL05 = \beta_0 + \beta_1 DISTR + \beta_2 SHAPE + \beta_3 TAIL + \beta_4 SIZE + \beta_5 GSCOND \quad (14)$$

This equation was estimated using the iterative Fisher's scoring method. This is the default method in SAS PROC LOGISTIC as shown in the Table III. The term "Sum of Frequencies" meant the total number of frequencies in the response profile. Since the data in this study did not have missing values, the sum of frequencies read and used were same, i.e. 25,257. The numbers of observations read and used were 84.

Note the level-ordering displayed in the response profile. By default, PROC LOGISTIC in SAS system will attempt to model (i.e. predict the probability of) the lower of the two values of the dependent variable, i.e. PVAL05=0. However, this was not the desired condition. Thus, the DESCENDING option (refer to line 000102 in Appendix B) was included to override this system default. Now, the value 1 became the reference level. Hence, *p* was defined as the probability of being in group 1.

These probabilities were used to group the simulation conditions combinations (the independent variables). The classification depended upon a cutoff point. Generically, SAS set the cutoff probability as 0.5. In order to define a cutoff probability, the option PPROB was invoked (refer to line 000107 in Appendix B). Simulation conditions combinations with predicted values that exceeded the classification cutoff were classified as group 1, while those with predicted values smaller than the cutoff were classified as group 0. In this case, the value of the cutoff point for classifying cases was 0.07 (1860/25257).

Since the data were in count form, it is indicated to PROC LOGISTIC by writing the FREQ statement (refer to line 000105 in Appendix B). The main effects model was specified in the MODEL statement, which also included the options SCALE=NONE and AGGREGATE (refer both to line 000106 in Appendix B). The SCALE option enabled the PROC LOGISTIC to treat each unique combination of the

independent variable values as a distinct group in computing the goodness-of-fit statistics. The SCALE=NONE statement specifies that no correction was needed for the dispersion parameter. The AGGREGATE option grouping the observations into subpopulations and compute the goodness-of-fit test statistics for them.

The results for simulation conditions are discussed in this manner:

1) Assessment of model
2) Percent of correct classification
3) Parameter estimates

B. Assessment of Model

In order to assess the model fit, the likelihood ratio test was used. The test statistic for the null hypothesis that model fits the data, was the likelihood ratio test involving log likelihoods. The model fit statistics from Table IV, Part A showed the model convergence status and statistics for testing the overall model significance.

The output of the likelihood ratio test provided two *-2LogL* statistics. The result of testing this hypothesis and the *p*-value for this decision was presented in Table IV, Part B. The χ^2 was 834.0929. The *p*-value was less than 0.0001 implying the rejection of H_0. This indicated that the overall DLR model was highly significant and at least one and perhaps all of the parameter estimates were significantly different from zero. The model with the independent variables was significantly better then the model with just the intercept. In other words, the inclusion of the independent variables significantly improved model fit and contributed to predicting the likelihood of being classified as group 1. Other than the testing of model fit, L-O-O classification table is used to shows the accuracy of the model to assign records into correct group.

C. Percent of Correct Classification

In any classification method, the hit ratio is still the primary indicator of the goodness of the method. Usually, the class display of this assessment was in the form of contingency table of observations versus predicted grouping. In SAS PROC LOGISTIC, this was given as bias-adjusted classification table (refer to Table V).

In computing the bias-adjusted classification table, SAS used an approximate pseudo jack-knife method known as the L-O-O technique. Essentially, for a given observation, a model was fitted by excluding an observation from the data and then classifies the observation using the resulting model. The CTABLE option (refer to line 000107 in Appendix B) allowed one to use L-O-O technique which gave us the unbiased estimate of the correct classification. Table V was the bias-adjusted classification table produced by CTABLE option. This particular model constructed from the training data set has 59.8% hit ratio caused by moderate high sensitivity (69.4%) and specificity (59.1%).

D. Parameter Estimates

The maximum likelihood method was used to estimate the parameters. Then, the Wald chi-square was used to test the statistical significance of each of the coefficient (β_j). Next, in order to interpret the DLR model, the logits were changed into odds ratio. The odd ratios can use to determine independent variables that were included in the DLR model to obtain robust Type I error rates.

Noticed that there were some peculiar values for the estimate (refer to Table VI). These were zero values that obtained for $SHAPE_{SKEW}$, $TAIL_{LEPT}$ and $TAIL_{PLAT}$. This was due to the presence of multicollinearity in the data. The SAS output gave three equations (refer to Equations 15-17) with regard to $SHAPE_{SKEW}$, $TAIL_{LEPT}$ and $TAIL_{PLAT}$. They were

$$\begin{aligned}\text{SHAPE}_{\text{SKEW}} =& 0.14286 * \text{INTERCEPT} - 1.14286 * \text{DISTR}_{\text{BETA(0.5,0.5)}} + 0.85714 * \text{DISTR}_{\text{FLEISHMAN1}} + 0.85714 * \text{DISTR}_{\text{FLEISHMAN2}} + 0.85714 * \text{DISTR}_{\text{G=.225/H=.225}} + 0.85714 * \text{DISTR}_{\text{G=.76/H=.098}} - 1.14286 * \text{DISTR}_{\text{G=0/H=.225}} \quad (15)\end{aligned}$$

$$\begin{aligned}\text{TAIL}_{\text{LEPT}} =& 0.14286 * \text{INTERCEPT} - 0.14286 * \text{DISTR}_{\text{BETA(0.5,0.5)}} - 0.14286 * \text{DISTR}_{\text{FLEISHMAN1}} - 1.14286 * \text{DISTR}_{\text{FLEISHMAN2}} + 0.85714 * \text{DISTR}_{\text{G=.225/H=.225}} + 0.85714 * \text{DISTR}_{\text{G=.76/H=.098}} + 0.85714 * \text{DISTR}_{\text{G=0/H=.225}} \quad (16)\end{aligned}$$

$$\begin{aligned}\text{TAIL}_{\text{PLAT}} =& \text{DISTR}_{\text{BETA(0.5,0.5)}} + \text{DISTR}_{\text{FLEISHMAN1}} - \text{DISTR}_{\text{FLEISHMAN2}} \quad (17)\end{aligned}$$

Notice that these three equations were linear combinations of the variables that were non-zero estimates. These variables were not included in the model because of their linear relationships. However, this did not imply that they were not important variables in the model. This actually implied that these variables were characteristics of other variables that existed in the model.

The *SHAPE* and *TAIL* variables are the skewness and kurtosis of the distributions, respectively. Each distribution comes with known values of skewness and kurtosis indices. Therefore, the *SHAPE* and *TAIL* were inherent in the distribution. Technically, a linear relationship can be formulated for each distribution, i.e. $DISTR_i = \lambda_0 + \lambda_1 SHAPE_i + \lambda_2 TAIL_i$, where i = type of distribution. Then, there were seven of these equations for all the seven distributions in the data set. Base upon these seven equations, they can be reformulated into Equations 15, 16 and 17.

From Table VI, the parameter column showed the simulation conditions and the second column was the levels of the conditions. Each level of the parameter consisted of dichotomous dummy variables. Originally, *DISTR* has seven levels. Reformulating these as dummy variables, there were six dummy variables. Each of the six distributions was compared against the standard normal distribution. The rest of the variables undergo the same process, where a level was considered as a reference level and every other level was compared against this reference level.

Then, the Wald χ^2 was used to test the statistical significance of each of the coefficient (β_j). From Table VI, the DLR equation for the model could be express as

$$PVAL05_{predicted} = -2.7557 + 0.4032DISTR_{BETA(0.5,0.5)} - 0.5912DISTR_{FLEISHMAN1} - 0.9563DISTR_{FLEISHMAN2} - 0.5950DISTR_{G=.76/H=-.098} - 0.5496DISTR_{G=0/H=-.225} + 0.2210GSIZE_{120} + 0.1596GSCOND_{INCR05} - 0.1810GSCOND_{INCR10} \quad (18)$$

These estimates described the relationship between the dependent variables and the independent variables, where the dependent variable was on the *logit* scale. From the same table, all parameters were significant under Wald test, except for $DISTR_{G=.225/H=-.225}$.

The coefficients (β_j) in the model Equation 18 were *logits*. To interpret the model, the *logits* was changed into odds ratio. This was represented in Exp(Est) column. From Table VI, the independent variables that have influence/important are those with Exp(Est)>1 or Exp(Est)<1, with *p*-values significant (<0.05). The variables that influenced classification into group 1 were $DISTR_{BETA(0.5,0.5)}$, $DISTR_{G=0/H=.225}$, $DISTR_{N(0,1)}$, $GSIZE_{120}$, $GSCOND_{INCR05}$ and $GSCOND_{EQUAL}$.

The odds ratio for $DISTR_{BETA(0.5,0.5)}$ favored the *BETA(0.5,0.5)* distribution over *N(0,1)* distribution. This meant that the likelihood of getting good rates of Type I error using the *BETA(0.5,0.5)* distribution was about twice that of the *N(0,1)* distribution, when other variables were controlled. The same result was observed for the $DISTR_{G=0/H=-.225}$ distribution. Noticed that the *BETA(0.5,0.5)* was a symmetric playkurtic distribution with γ_1 = 0 and γ_2 = -1.5 and the *G=0/H=.225* was a symmetric leptokurtic distribution.

On the contrary, the odds ratios for $DISTR_{FLEISHMAN1}$, $DISTR_{FLEISHMAN2}$, $DISTR_{G=.76/H=-.098}$ favored the *N(0,1)* distribution. This meant that when type of distribution was standard normal, it was more likely to result in good rates of Type I error compared with the skewed platykurtic distribution ($DISTR_{FLEISHMAN1}$), the skewed normal-tailed distribution ($DISTR_{FLEISHMAN2}$) and the skewed leptokurtic distribution ($DISTR_{G=.76/H=-.098}$).

From the same table, noticed that the odds ratio for $GSIZE_{120}$ favored $GSIZE_{120}$ over $GSIZE_{60}$. This meant that the likelihood of getting good rates of Type I error using large total sample size (*N*=120) was about twice that of the small total sample size (*N*=60).

The odds ratio for $GSCOND_{INCR05}$ favored the $GSCOND_{INCR05}$ over $GSCOND_{EQUAL}$. However, the odds ratio for $GSCOND_{INCR10}$ favored the $GSCOND_{EQUAL}$ over $GSCOND_{INCR10}$. This meant that unbalanced group size increments by five units obtained from a (15, 20, 25) design or (35, 40, 45) design (representing by $GSCOND_{INCR05}$) was more likely to give good rates of Type I error compared with balanced group size such as (20, 20, 20) design or (40, 40, 40) design. While, balanced group size ($GSCOND_{EQUAL}$) was more likely to give good rates of Type I error compared with unbalanced group size increments by ten units obtained from a (10, 20, 30) design or (30, 40, 50) design (representing by $GSCOND_{INCR10}$). The latter design represented extremely unequal sample size.

VIII. Conclusion

The most common method to use for analyzing data with binary response variables is DLR. In DLR model, the response variable is Bernoulli distributed mean value related to the independent variables through the logit transformation. The SAS system facilitates the building of a program to conduct DLR analysis by using PROC LOGISTIC and DATA step. In this study, the response variables are binary random variables, taking values 1 and 0, where 1 representing *p*-values falling in [0.045, 0.050] and 0 for *p*-values falling outside of this interval. In order to test hypotheses in DLR, the likelihood ratio test was have used. Wald test and *p*-values, and odds ratios were used to analyze maximum likelihood estimates. In this study, independent variables that were included in the DLR model to obtained robust Type I error rates falling in [0.045, 0.050] were successfully determined. That is, the model should include either symmetric platykurtic distributions ($DISTR_{BETA(0.5,0.5)}$) or symmetric leptokurtic distributions ($DISTR_{G=0/H=-.225}$), with a (35, 40, 45) design. The (35, 40, 45) design indicated conditions of large total sample size ($GSIZE_{120}$) and moderately unequal sample size ($GSCOND_{INCR05}$).

Usually, if one is interested to do prediction of model, the hit ratio of 80% is necessary. However, it is not required in this study because the hit ratio is used for the purpose of model accuracy assessment. Hence, the model constructed could not be used for prediction purpose.

Appendices

Appendix A
Restructure Variables for Training Data Set

No.	DISTR	SHAPE	TAIL	GSIZE	GSCOND	PVAL05	COUNT
1	BETA(0.5,0.5)	SYMM	PLAT	60	EQUAL	0	551
2	BETA(0.5,0.5)	SYMM	PLAT	60	EQUAL	1	59
3	BETA(0.5,0.5)	SYMM	PLAT	60	INCR05	0	533
4	BETA(0.5,0.5)	SYMM	PLAT	60	INCR05	1	59
5	BETA(0.5,0.5)	SYMM	PLAT	60	INCR10	0	565
6	BETA(0.5,0.5)	SYMM	PLAT	60	INCR10	1	37
7	BETA(0.5,0.5)	SYMM	PLAT	120	EQUAL	0	550
8	BETA(0.5,0.5)	SYMM	PLAT	120	EQUAL	1	51
9	BETA(0.5,0.5)	SYMM	PLAT	120	INCR05	0	515
10	BETA(0.5,0.5)	SYMM	PLAT	120	INCR05	1	82
11	BETA(0.5,0.5)	SYMM	PLAT	120	INCR10	0	560
12	BETA(0.5,0.5)	SYMM	PLAT	120	INCR10	1	32
13	FLEISHMAN1	SKEW	PLAT	60	EQUAL	0	567
14	FLEISHMAN1	SKEW	PLAT	60	EQUAL	1	36
15	FLEISHMAN1	SKEW	PLAT	60	INCR05	0	594
16	FLEISHMAN1	SKEW	PLAT	60	INCR05	1	13
17	FLEISHMAN1	SKEW	PLAT	60	INCR10	0	594
18	FLEISHMAN1	SKEW	PLAT	60	INCR10	1	11

19	FLEISHMAN1	SKEW	PLAT	120	EQUAL	0	572
20	FLEISHMAN1	SKEW	PLAT	120	EQUAL	1	30
21	FLEISHMAN1	SKEW	PLAT	120	INCR05	0	584
22	FLEISHMAN1	SKEW	PLAT	120	INCR05	1	24
23	FLEISHMAN1	SKEW	PLAT	120	INCR10	0	585
24	FLEISHMAN1	SKEW	PLAT	120	INCR10	1	13
25	FLEISHMAN2	SKEW	NORM	60	EQUAL	0	586
26	FLEISHMAN2	SKEW	NORM	60	EQUAL	1	13
27	FLEISHMAN2	SKEW	NORM	60	INCR05	0	571
28	FLEISHMAN2	SKEW	NORM	60	INCR05	1	23
29	FLEISHMAN2	SKEW	NORM	60	INCR10	0	596
30	FLEISHMAN2	SKEW	NORM	60	INCR10	1	12
31	FLEISHMAN2	SKEW	NORM	120	EQUAL	0	568
32	FLEISHMAN2	SKEW	NORM	120	EQUAL	1	14
33	FLEISHMAN2	SKEW	NORM	120	INCR05	0	582
34	FLEISHMAN2	SKEW	NORM	120	INCR05	1	16
35	FLEISHMAN2	SKEW	NORM	120	INCR10	0	592
36	FLEISHMAN2	SKEW	NORM	120	INCR10	1	10
37	G=.225/H=.225	SKEW	LEPT	60	EQUAL	0	564
38	G=.225/H=.225	SKEW	LEPT	60	EQUAL	1	33
39	G=.225/H=.225	SKEW	LEPT	60	INCR05	0	552
40	G=.225/H=.225	SKEW	LEPT	60	INCR05	1	45
41	G=.225/H=.225	SKEW	LEPT	60	INCR10	0	581
42	G=.225/H=.225	SKEW	LEPT	60	INCR10	1	18
43	G=.225/H=.225	SKEW	LEPT	120	EQUAL	0	573
44	G=.225/H=.225	SKEW	LEPT	120	EQUAL	1	32
45	G=.225/H=.225	SKEW	LEPT	120	INCR05	0	552
46	G=.225/H=.225	SKEW	LEPT	120	INCR05	1	54
47	G=.225/H=.225	SKEW	LEPT	120	INCR10	0	540
48	G=.225/H=.225	SKEW	LEPT	120	INCR10	1	64
49	G=.76/H=-.098	SKEW	LEPT	60	EQUAL	0	553
50	G=.76/H=-.098	SKEW	LEPT	60	EQUAL	1	41
51	G=.76/H=-.098	SKEW	LEPT	60	INCR05	0	584
52	G=.76/H=-.098	SKEW	LEPT	60	INCR05	1	16
53	G=.76/H=-.098	SKEW	LEPT	60	INCR10	0	585
54	G=.76/H=-.098	SKEW	LEPT	60	INCR10	1	15
55	G=.76/H=-.098	SKEW	LEPT	120	EQUAL	0	570
56	G=.76/H=-.098	SKEW	LEPT	120	EQUAL	1	18
57	G=.76/H=-.098	SKEW	LEPT	120	INCR05	0	581
58	G=.76/H=-.098	SKEW	LEPT	120	INCR05	1	19
59	G=.76/H=-.098	SKEW	LEPT	120	INCR10	0	583
60	G=.76/H=-.098	SKEW	LEPT	120	INCR10	1	16
61	G=0/H=.225	SYMM	LEPT	60	EQUAL	0	570
62	G=0/H=.225	SYMM	LEPT	60	EQUAL	1	27
63	G=0/H=.225	SYMM	LEPT	60	INCR05	0	530
64	G=0/H=.225	SYMM	LEPT	60	INCR05	1	79
65	G=0/H=.225	SYMM	LEPT	60	INCR10	0	572
66	G=0/H=.225	SYMM	LEPT	60	INCR10	1	33
67	G=0/H=.225	SYMM	LEPT	120	EQUAL	0	542
68	G=0/H=.225	SYMM	LEPT	120	EQUAL	1	69
69	G=0/H=.225	SYMM	LEPT	120	INCR05	0	532
70	G=0/H=.225	SYMM	LEPT	120	INCR05	1	69
71	G=0/H=.225	SYMM	LEPT	120	INCR10	0	514
72	G=0/H=.225	SYMM	LEPT	120	INCR10	1	92
73	N(0,1)	SYMM	NORM	60	EQUAL	0	530
74	N(0,1)	SYMM	NORM	60	EQUAL	1	77
75	N(0,1)	SYMM	NORM	60	INCR05	0	538
76	N(0,1)	SYMM	NORM	60	INCR05	1	65
77	N(0,1)	SYMM	NORM	60	INCR10	0	571
78	N(0,1)	SYMM	NORM	60	INCR10	1	35
79	N(0,1)	SYMM	NORM	120	EQUAL	0	479
80	N(0,1)	SYMM	NORM	120	EQUAL	1	128
81	N(0,1)	SYMM	NORM	120	INCR05	0	463
82	N(0,1)	SYMM	NORM	120	INCR05	1	145
83	N(0,1)	SYMM	NORM	120	INCR10	0	473
84	N(0,1)	SYMM	NORM	120	INCR10	1	135
					Total	42	25,257

APPENDIX B
PARTIAL PROGRAM FOR LOGISTIC REGRESSION OF SIMULATION CONDITIONS

```
000001 /*********************************************************/
000002 /*          DICHOTOMOUS LOGISTIC REGRESSION              */
000003/*********************************************************/
000004
000005 Data trainingsimcond42;
000006 LABEL    DISTR  = 'Type of distribution'
000007      SHAPE  = 'Skewness of distribution'
000008      TAIL   = 'Kurtosis of distribution'
000009      GSIZE  = 'Total group size'
000010      GSCOND = 'Unbalanced group size increments';
000011 LENGTH DISTR $ 13;
000012 INPUT DISTR $ SHAPE $ TAIL $ GSIZE $ GSCOND $PVAL05
000013      COUNT @@;
000014 CARDS;
000015 BETA(0.5,0.5)  SYMM  PLAT60 EQUAL    0  551
000016 BETA(0.5,0.5)  SYMM  PLAT60 EQUAL    1  59
  :
000098 N(0,1)     SYMM  NORM  120   INCR10 1  135
000099 ;
000100 RUN;
000101
000102 PROC LOGISTIC DATA=TRAININGSIMCOND42
         DESCENDING;
000103  CLASS DISTR(REF='N(0,1)') SHAPE(REF='SYMM')
         TAIL(REF='NORM')
000104  GSIZE(REF='60') GSCOND(REF='EQUAL');
000105  FREQ COUNT;
000106  MODEL PVAL05 = DISTR SHAPE TAIL GSIZE GSCOND /
000107  SCALE=NONE AGGREGATE EXPB RSQUARE CTABLE
         PPROB=0.07;
000108  /*PPROB is the prior probabilities to the sample size*/
000109  OUTPUT OUT=PROBSLR REDPROBS=(CROSSVALIDATE);
000110  /*PROBSLR saves posterior probabilities for classification*/
000111 RUN;
```

ACKNOWLEDGEMENT

The authors would like to acknowledge the work that led to this paper publication funded by the School of Mathematical Sciences, and supported by the Universiti Sains Malaysia Fellowship.

REFERENCES

[1] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. New York: Wiley, 2002.

[2] J. M. Henshall, and M. E. Goddard, "Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression," *Genetics*, vol.151, pp. 885-894, 1999.

[3] D. F. Levinson, P. Holmans, R. E. Straub, M. J. Owen, D. B. Wildenauer, P. V. Gejman, A. E. Pulver, C. Laurent, K. S. Kendler, D. Walsh, N. Norton, N. M. Williams, S. G. Schwab, B. Lerer, B. J. Mowry, A. R. Sanders, S. E. Antonarakis, J. L. Blouin, , J. F. DeLeuze, and J. Mallet, "Multicenter linkage study of schizophrenia candidate regions on chromosomes 5q, 6q, 10p and 13q: Schizophrenia linkage collaborative group III," *American Journal of Human Genetics*, vol. 67, pp. 652-663.

[4] A. DeMaris, "Feedback: Interpreting logistic regression results: A critical commentary," *Journal of Marriage and the Family*, vol. 52, pp. 271-277, 1990.

[5] P. S. Morgan, and J. D. Teachman, "Logistic regression: Description, examples, and comparisons," *Journal of Marriage and the Family*, vol.*50*, pp. 928-936, 1988.

[6] I. L. Lottes, M. A. Adler, and A. DeMaris, "Using and interpreting logistic regression: A guide for teaching and students," *Journal of Teaching Sociology*, vol. 24, pp. 284-298, 1996.

[7] D. R. Cox, and E. J. Snell, *The Analysis of Binary Data*, 2nd ed. London: Chapman & Hall, 1989.

[8] D. W. Hosmer, and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 1989.

[9] F. Azuaje, "Genomic data sampling and its effect on classification performance assessment," *BMC Bioinformatics*, vol.4, pp. 1-14, 2003.
[10] D. Pregibon, "Logistic regression diagnostics," *The Annals of Statistics*, vol.9, pp. 705-724, 1981.
[11] P. McCullagh, and J. A. Nelder, *Generalized linear models*, 2nd ed. London: Chapman and Hall, 1989.
[12] M. Crask, and Perreault, "Validation of discriminant analysis in marketing research," *Journal of Marketing Research*, vol.14, pp.60-68, 1977.
[13] W. R. Dillon, and M. Goldstein, *Multivariate Analysis: Methods and Applications*. New York: Wiley, 1984.
[14] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.
[15] G. Gong, "Cross-validation, the jackknife and the bootstrap excess error estimation in forward regression logistic regression," *Journal of the American Statistical Association*, vol.81, no.393, pp.108-113, 1986.
[16] C. J. Huberty, *Applied Discriminant Analysis*. New York: Wiley, 1994.
[17] R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.
[18] C. J. Huberty, J. M. Wisenbaker, J. D. Smith, and J. C. Smith, "Using categorical variables in discriminant analysis," *Multivariate Behavioral Research*, vol.21, pp. 479-496, 1986.
[19] D. C. Montgomery, E. A. Peck, and G. Vinning, *An Introduction to Linear Regression Analysis*, 3rd ed. New York: Wiley, 2001.
[20] S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. Thousand Oaks: Sage Publications, 2002.
[21] D. G. Kleinbaum, L. L. Kupper, K. E. Muller, and A. Nizam, *Applied Regression Analysis and Other Multivariate Methods*. New York: Duxbury Press, 1998.
[22] H. J. Keselman, R. R. Wilcox, J. Algina, A. R. Othman, and K. A. Fradette, "Comparative study of robust tests for spread: asymmetric trimming strategies," *British Journal of Mathematical and Statistical Psychology*, vol.61, pp.235-253, 2008.
[23] A. I. Fleishman, "A method for simulating non-normal distributions," *Psychometrika*, vol.43, pp. 521-532, 1978.
[24] D. C. Hoaglin, Summarizing Shape Numerically: The *g*- and *h*-Distributions. In D. C. Hoaglin, F. Mosteller & J. Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes* (pp. 461-513). New York: Wiley, 1985.

Teh, S. Y. currently serves as a graduate teaching assistant and Universiti Sains Malaysia (USM) fellow at the School of Mathematical Sciences, USM. She obtained her B.Sc. (2006) and M.Sc. (2008) in the field of applied statistics both from USM, Penang, Malaysia.

Her first degree research project is 'Process Improvement (Quality Control) in Motorola Penang'. Her master degree thesis is 'Data Mining for Robust Tests of Spread' and ph.D. thesis is 'Monitoring Process Mean and Variability With one DEWMA Chart'. She was the research officer for a survey of 'Computer and Internet Usage and Users' an Information, Communication and Technologies project for the Penang state government. She was also a USM fellow at the School of Distance Education, USM. Her research interests are data mining, robust tests of spread, quality control, and process control.

Othman, A. R. graduated with a PhD from University of California, Santa Barbara in 1995 in the field of educational research methodology.

He is a deputy dean of Institute of Graduate Studies and an associate professor of mathematics at the School of Distance Education, Universiti Sains Malaysia (USM). At present he is involved in applied statistical research and also consultant in taste and smell sensor research at the School of Pharmacy, USM. His research interests are in robust statistics and psychometrics.

Michael, B. C. Khoo is an Associate Professor in the School of Mathematical Sciences, Universiti Sains Malaysia (USM). He specializes in statistical process control. He publishes extensively in International journals. His papers are either published or accepted for publications in renown International journals, such as International Journal of Production Economics, Quality and Reliability Engineering International, Computers and Industrial Engineering, Communications in Statistics (Series A & B), Computational Statistics and Data Analysis, Computational Statistics, European Journal of Operational Research, International Journal of Production Research, Quality Engineering and International Journal of Reliability, Quality and Safety Engineering. He has vetted numerous papers in International journals, a large number of which are indexed in the ISI database. He is a member of the American Society for Quality and a life member of the Malaysian Mathematical Society. He also serves as a member of the editorial boards of Quality Engineering and a few other international journals.