

ON GOODNESS-OF-FIT OF  
LOGISTIC REGRESSION MODEL

by

YING LIU

M.S., University of Rhode Island, 2003  
M.S., Henan Normal University, 2000

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics  
College of Arts and Sciences

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

2007

## Abstract

Logistic regression model is a branch of the generalized linear models and is widely used in many areas of scientific research. The logit link function and the binary dependent variable of interest make the logistic regression model distinct from linear regression model.

The conclusion drawn from a fitted logistic regression model could be incorrect or misleading when the covariates can not explain and /or predict the response variable accurately based on the fitted model---that is, lack-of-fit is present in the fitted logistic regression model.

The current goodness-of-fit tests can be roughly categorized into four types. (1) The tests are based on covariate patterns, e.g., Pearson's Chi-square test, Deviance D test, and Osius and Rojek's normal approximation test. (2) Hosmer-Lemeshow's  $\hat{C}$  and Hosmer-Lemeshow's  $\hat{H}$  tests are based on the estimated probabilities. (3) Score tests are based on the comparison of two models, where the assumed logistic regression model is embedded into a more general parametric family of models, e.g., Stukel's Score test and Tsiatis's test. (4) Smoothed residual tests include le Cessie and van Howelingen's test and Hosmer and Lemeshow's test. All of them have advantages and disadvantages.

In this dissertation, we proposed a partition logistic regression model which can be viewed as a generalized logistic regression model, since it includes the logistic regression model as a special case. This partition model is used to construct goodness-of-fit test for a logistic regression model which can also identify the nature of lack-of-fit is due to the tail or middle part of the probabilities of success. Several simulation results showed that the proposed test performs as well as or better than many of the known tests.

ON GOODNESS-OF-FIT OF  
LOGISTIC REGRESSION MODEL

by

YING LIU

M.S., University of Rhode Island, 2003  
M.S., Henan Normal University, 2000

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics  
College of Arts and Sciences

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

2007

Approved by:

Major Professor  
Shie-shien Yang

## Abstract

Logistic regression model is a branch of the generalized linear models and is widely used in many areas of scientific research. The logit link function and the binary dependent variable of interest make the logistic regression model distinct from the linear regression model.

The conclusion drawn from a fitted logistic regression model could be incorrect or misleading when the covariates can not explain and /or predict the response variable accurately based on the fitted model---that is, lack-of-fit is present in the fitted logistic regression model.

The current goodness-of-fit tests can be roughly categorized into four types. (1) The tests are based on covariate patterns, e.g., Pearson's Chi-square test, Deviance D test, and Osius and Rojek's normal approximation test. (2) Hosmer-Lemeshow's  $\hat{C}$  and Hosmer-Lemeshow's  $\hat{H}$  tests are based on the estimated probabilities. (3) Score tests are based on the comparison of two models, where the assumed logistic regression model is embedded into a more general parametric family of models, e.g., Stukel's Score test and Tsiatis's test. (4) Smoothed residual tests include le Cessie and van Howelingen's test and Hosmer and Lemeshow's test. All of them have advantages and disadvantages.

In this dissertation, we proposed a partition logistic regression model which can be viewed as a generalized logistic regression model, since it includes the logistic regression model as a special case. This partition model is used to construct goodness-of-fit test for a logistic regression model which can also identify the nature of lack-of-fit is due to the tail or middle part of the probabilities of success. Several simulation results showed that the proposed test performs as well as or better than many of the known tests.

# Table of Contents

|  |      |
|--|------|
| List of Figures .....  | vii  |
| List of Tables.....  | viii |
| Acknowledgements.....  | x    |
| CHAPTER 1 - Introduction.....                                    | 1    |
| 1.1 Basic Concepts on Logistic Regression Model .....            | 1    |
| 1.2 The Goodness-of Fit Test for Logistic Regression Model ..... | 8    |
| CHAPTER 2 - Literature Review.....                               | 16   |
| 2.1 Pearson Chi-square $\chi^2$ Test and Deviance D Test.....    | 18   |
| 2.2 The Hosmer and Lemeshow Tests.....                           | 24   |
| 2.2.1 Hosmer and Lemeshow's $\hat{C}$ .....                      | 24   |
| 2.2.2 Hosmer and Lemeshow $\hat{H}$ .....                        | 26   |
| 2.3 Osius and Rojek Normal Approximation Test.....               | 28   |
| 2.4 Stukel's Score Test .....                                    | 30   |
| 2.5. An Illustrative Example .....                               | 33   |
| CHAPTER 3 - The Proposed Method .....                            | 36   |
| 3.1 Description of the Proposed Test.....                        | 38   |
| 3.2 The First Example .....                                      | 43   |
| 3.3 The Second Example.....                                      | 46   |
| 3.4 The Third Example .....                                      | 53   |
| 3.5 The Fourth Example.....                                      | 55   |
| CHAPTER 4 - Simulation Study.....                                | 61   |
| 4.1 The First Simulation Study.....                              | 62   |
| 4.2 The Second Simulation Study.....                             | 68   |
| 4.3 The Third Simulation Study.....                              | 75   |
| 4.4 The Fourth Simulation Study.....                             | 80   |
| 4.5 Summary on Simulation Studies.....                           | 85   |
| 4.6 Convergence Problem in the Simulation Studies.....           | 86   |

|  |     |
|--|-----|
| CHAPTER 5 - Summary.....                                 | 94  |
| CHAPTER 6 - Extension and Future Study .....             | 96  |
| CHAPTER 7 - References .....                             | 97  |
| Appendix A - Data on O-Ring Failure .....                | 104 |
| Appendix B - Data on Vasoconstriction .....              | 105 |
| Appendix C - Data on Leukemia Survival .....             | 107 |
| Appendix D - Beetle Data Set .....                       | 109 |
| Appendix E - Warsaw Girl Data .....                      | 110 |
| Appendix F - Data set For Illustrative Example 3.....    | 111 |
| Appendix G - Data set for Illustrative Example 4 .....   | 112 |
| Appendix H - SAS Program Code for Simulation Study ..... | 113 |

## List of Figures

|  |    |
|--|----|
| Figure 1-1 Estimated Failure Probability from Logistic Regression Model and Probability Linear Model ..... | 7  |
| Figure 1-2 Three Dimensional Plot on Transformed Vasoconstriction Data .....                               | 11 |
| Figure 2-1 the Plot of $\pi(x)$ versus $\eta(x)$ for Stukel's Model .....                                  | 31 |
| Figure 3-1 Plot of Estimated probabilities from the Assumed and .....                                      | 45 |
| Figure 3-2 the Observed Probability of Menstruating on Warsaw Girls .....                                  | 46 |
| Figure 3-3 the Plot of Residual versus Age .....   | 47 |
| Figure 3-4 Probability versus Age on Warsaw Girl Data .....  | 51 |
| Figure 4-1 Plots of Rejection Rate for Study One (n=100).....  | 65 |
| Figure 4-2 Plots of Rejection Rate for Study One (n=200).....  | 66 |
| Figure 4-3 Plots of Rejection Rate for Study One (n=500).....  | 66 |
| Figure 4-4 Plots of Rejection Rate for Study Two (n=100).....  | 72 |
| Figure 4-5 Plots of Rejection Rate for Study Two (n=200).....  | 72 |
| Figure 4-6 Plots of Rejection Rate for Study Two (n=500).....  | 73 |
| Figure 4-7 Plots of Rejection Rate for Study Three (n=100).....  | 78 |
| Figure 4-8 Plots of Rejection Rate for Study Three (n=200).....  | 78 |
| Figure 4-9 Plots of Rejection Rate for Study Three (n=500).....  | 79 |
| Figure 4-10 Plots of Rejection Rate for Study Four (n=100).....  | 82 |
| Figure 4-11 Plots of Rejection Rate for Study Four (n=200).....  | 83 |
| Figure 4-12 Plots of Rejection Rate for Study Four (n=500).....  | 83 |
| Figure 4-13 Scatter Plot Sample Points with Complete Separation Data.....                                  | 90 |

## List of Tables

|  |    |
|--|----|
| Table 1-1 Link Functions for the Generalized Linear Models .....   | 3  |
| Table 1-2 A Comparison of Linear & Logistic Regression Models.....   | 5  |
| Table 1-3 Information on Covariates from Logistic Regression.....  | 9  |
| Table 1-4 Goodness-of-Fit Test.....  | 9  |
| Table 1-5 Model Fit Statistics from Logistic Regression Model.....   | 10 |
| Table 1-6 Information on Covariates from Logistic Regression .....   | 12 |
| Table 1-7 Model Fit Statistics from Logistic Regression with.....  | 12 |
| Table 1-8 Goodness-of-Fit Test with Log-transformation.....  | 13 |
| Table 2-1 the Results of Hosmer-Lemeshow $\hat{C}$ Test .....  | 26 |
| Table 2-2 Parameter Estimates of Leukemia Data.....  | 33 |
| Table 2-3 Result from Six Goodness-of-Fit Tests on Leukemia Data .....   | 34 |
| Table 2-4 Group Scheme of Hosmer-Lemeshow ' s $\hat{C}$ test with SAS Software .....                                   | 35 |
| Table 3-1 Summary Table for Known and Proposed Tests .....   | 42 |
| Table 3-2 Results from the Known and Proposed Methods for Beetle Data.....   | 43 |
| Table 3-3 Prediction interval of Number of Beetles Killed by $CS_2$ .....  | 44 |
| Table 3-4 Results from the Known for Warsaw Girls Data.....  | 47 |
| Table 3-5 Results from Different Grouping Schemes of the Proposed Method.....  | 49 |
| Table 3-6 the Parameter Estimates for Model with Grouping Scheme I.....  | 50 |
| Table 3-7 Model Fit Statistics of Model 3.6.....   | 50 |
| Table 3-8 Observed Counts and Confidence Interval (Assumed Model and Proposed<br>Method) for the Warsaw Girl Data..... | 52 |
| Table 3-9 Results from the Known and Proposed Methods for Women's Role Data .....                                      | 53 |
| Table 3-10 Results from the Known Tests for Credit Card Data .....   | 55 |
| Table 3-11 Goodness-of-fit Test Based on Model (3.8).....  | 56 |
| Table 3-12 Parameter Estimates of Model (3.8).....   | 57 |
| Table 3-13 Model Fit Statistics from Two Cases of Model (3.8) $i=2$ .....  | 58 |
| Table 3-14 Parameter Estimates of Model (3.9).....   | 59 |



|   |    |
|---|----|
| Table 4-1 Observed Type I Error Rate for Simulation Study one.....              | 63 |
| Table 4-2 the 95% Confidence Interval for the Type I Error Rate.....            | 64 |
| Table 4-3 Rejection Rates for the Known and Proposed Tests for Study One.....   | 65 |
| Table 4-4 Observed Type I Error Rate for Simulation Study Two.....              | 69 |
| Table 4-5 Rejection Rates for the Known and Proposed Tests for Study Two.....   | 71 |
| Table 4-6 Observed Type I Error Rate for Simulation Study Three.....            | 76 |
| Table 4-7 Rejection Rates for the Known and Proposed Tests for Study Three..... | 77 |
| Table 4-8 Observed Type I Error Rate for Simulation Study Four.....             | 81 |
| Table 4-9 Rejection Rates for the Known and Proposed Tests for Study Four.....  | 82 |
| Table 4-10 Complete Separation Data.....  | 89 |
| Table 4-11 Partial Logistic Iteration Steps Printout.....                       | 91 |
| Table 4-12 Dispersion Matrices on the Selected Iterations.....                  | 92 |

## **Acknowledgements**

My PhD study in the Department of Statistics, Kansas State University could never be fulfilled without the full support from my major advisor, my advisory committee, the department office, my friends and my family. I would like to take this opportunity to thank all of them for their support, inspiring guidance, helpful discussion, beautiful friendship and generous love.

I would like to express my sincere gratitude to my major professor and mentor, Dr. Shie-shien Yang, for his constant support, patience, and encouragement throughout my studying life at Kansas State University. His support of my research and extracurricular activities were invaluable in making my graduate experience a positive and successful one. Without his support, I would not complete this long journey.

I want to thank my graduate committee member Dr. Paul I. Nelson, Dr. Susan Brown, Dr. Haiyan Wang for their warm-hearted help, support, and constructive suggestions, throughout my doctoral program of study either through formal meetings or informal chat.

I would also take this opportunity to thank Dr. John E. Boyer for giving me a chance to be a graduate teaching assistant. Let my appreciations go to Dr. John E. Boyer and Dr. Jim Higgins for their infinite patience and generous help in my teaching. This teaching experience will be the treasure of my whole life.

I would also give my thanks to all the faculty members in the Departments of Statistics for sharing their knowledge with me. With special thanks to Dr. Dallas, E.

Johnson for his very nice help, infinite patience during my studying. His attitude toward students and work will influence me in the future.

The friendship and support from all the graduate students and the department staff will be kept in my heart for the rest of my life. Special thanks to my best friends Lin Xie and Suzhi Wang, you make my studying life at K-state colorful.

Last but not the least, I would like to give my special thanks to my beloved husband and best friend, Dr. Zhiwu Li and an amazing daughter, Yeleeya. Their support throughout this process is immeasurable and I would not be where I am today without them, I am forever humbled and grateful. I would also thank my dear parents, brothers and sisters for their support during my oversea study. Their love is always the motivation to pursue my goals.

# CHAPTER 1 - Introduction

## 1.1 Basic Concepts on Logistic Regression Model

Generalized linear model (GLZ) is a useful generalization of the general linear model. In a GLZ, the dependent or outcome variable  $y$  is assumed to be generated from a particular probability distribution function from the exponential family. Generalized linear models are defined by three components:

1. Random Components: the probability distribution function  $f$  for the dependent variable  $\mathbf{y}=(y_1, y_2, \dots, y_n)'$  is from an exponential family of distributions.
2. Systematic component: a linear predictor  $\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta}$ , where the matrix  $\mathbf{X}$  contains columns of explanatory variables which is called the design matrix,  $\boldsymbol{\beta}$  are the unknown parameters and  $\boldsymbol{\eta}=(\eta_1, \eta_2, \dots, \eta_n)$ .
3. A link function  $g$  such that  $E(\mathbf{y}|\mathbf{X}) = \boldsymbol{\mu} = \mathbf{g}^{-1}(\boldsymbol{\eta})$ , which provides the link between the linear predictor and the mean of the probability distribution function  $f$  of  $y$  (McCullagh and Nelder 1989, Agresti 1990). Let  $E(y_i|\mathbf{x}_i) = \mu_i$ ,  $i=1,2,\dots,N$ . then  $\mathbf{x}_i$  is the  $i^{\text{th}}$  row of  $\mathbf{X}$ ; the explanatory variables associated with the  $i^{\text{th}}$  response  $y_i$  vector of independent variables.

Generally, the exponential family of distributions for the dependent variables  $\mathbf{y}$  are the probability distributions, parameterized by  $\theta$  and  $\phi$  with density functions expressed in the form:

$$f(y; \theta, \phi) = e^{\left\{ \frac{y\theta - b(\theta)}{a(\phi)} \right\} + c(y, \phi)} \quad (1.1)$$

$\theta$  is related to the mean of the distribution with  $E(y | \mathbf{x}) = \mu = b'(\theta)$ .  $\phi$  is the *dispersion parameter*, typically is known and is usually related to the variance of the distribution and  $Var(y | \mathbf{x}) = b''(\theta)a(\phi)$ . The functions  $a$ ,  $b$ , and  $c$ , are known.

The generalized linear models can be summarized in the following table with the following commonly used link functions (McCullagh and Nelder, 1989).

**Table 1-1 Link Functions for the Generalized Linear Models**

| Distribution | Name     | Link Function   | Mean Function   |
|--------------|----------|---|---|
| Normal       | Identity | $\mathbf{x}\beta = \mu$                               | $\mu = \mathbf{x}\beta$   |
| Exponential  | Inverse  | $\mathbf{x}\beta = \mu^{-1}$                          | $\mu = (\mathbf{x}\beta)^{-1}$                                  |
| Gamma        |          |   |   |
| Poisson      | Log      | $\mathbf{x}\beta = \ln(\mu)$                          | $\mu = \exp(\mathbf{x}\beta)$                                   |
| Binomial     | Logit    | $\mathbf{x}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$ | $\mu = \frac{\exp(\mathbf{x}\beta)}{1 + \exp(\mathbf{x}\beta)}$ |
| Multinomial  |          |   |   |

When each of the response variable of  $y$  is a binary or dichotomous outcome (taking on only values 0 and 1), the distribution function is generally chosen to be the Bernoulli distribution. There are several link functions for the Bernoulli data. They are probit link, log-log link and complementary log-log link, and logit link function. In the generalized linear models, the logistic link function is most widely used and the model with this type of link function is called logistic regression model (Nelder and Wedderburn 1972) which was first suggested by Berkson (1944), who showed the model could be

fitted by using iteratively weighted least squares. The logistic regression model is used in social and scientific fields (Dreisitl et al. 2005, Tan et al. 2004, Bedrick and Hill 1996). In some cases, the response variable  $y$  may have more than two outcomes (Tabachnick and Fidell 1996).

In order to simplify the notation, we denote the quantity  $E(y | \mathbf{x}) = \pi(\mathbf{x})$  as the conditional mean of  $y$  given  $\mathbf{x}$  in the logistic regression model which is also the probability of taking the value 1. Thus, the probability of  $y$  taking the value 0 is  $1 - \pi(\mathbf{x})$ .

The logit transformation is central to the formulation of the logistic regression model. It is assumed that the log of the odds  $\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}$  is linearly related to the covariates, that is,

$$\log \left[ \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right] = \text{logit}(\pi(\mathbf{x})) = \mathbf{x}\beta \quad (1.2)$$

The logit,  $\eta = \mathbf{x}\beta$ , may be continuous, and may range from  $-\infty$  to  $+\infty$ .

The logit transformation leads to the following form of the probability of success:

$$\pi(\mathbf{x}) = \frac{e^{\mathbf{x}\beta}}{1 + e^{\mathbf{x}\beta}} \quad (1.3)$$

Two requirements for logistic regression are that the observations are independent and binary, and that the logit of unknown binomial probabilities is linearly related to the covariates. Table 1-2 gives a comparison of the standard linear regression model with the logistic regression model.

**Table 1-2 A Comparison of Linear & Logistic Regression Models**

|                          | Linear regression   | Logistic regression  |
|--------------------------|---|--|
| Model                    | $E(y) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$                      | $\text{logit}(\pi(x)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$   |
| Response                 | Continuous variable   | Binary variable  |
| Covariates               | Continuous/or discrete  | Continuous/or discrete   |
| Meaning of coefficient   | Amount of expected change in the response y per unit change in covariates | Amount of change of log odds per unit change in the covariates   |
| Link function            | Identity  | Logit  |
| Distribution assumptions | Response variables are independent and with constant variance             | Response variables are independent and each has a Bernoulli distribution with probability of event dependent of the covariates |

When the response is binary, the linear regression model

$$E(y | \mathbf{x}) = \pi(\mathbf{x}) = \alpha + \mathbf{x}' \beta \quad (1.4)$$

is also called linear probability model (Agresti 1990). This model belongs to generalized linear model with identity link function. The probability  $\pi(\mathbf{x})$  is assumed to fall between 0 and 1 over a finite range of x values. However, some of estimated probabilities will fall outside the range of 0 and 1 with this model structure.

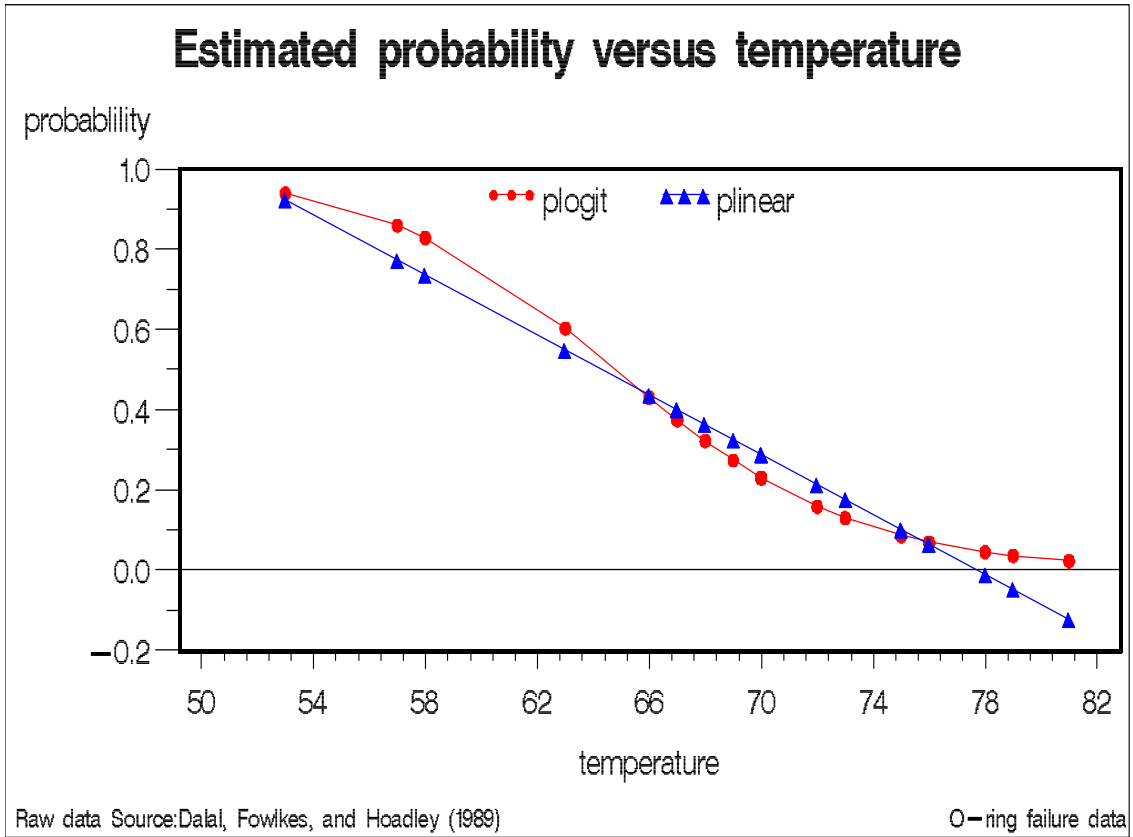
O-Ring failure data (Appendix A) will be used to illustrate that logistic regression model is more suitable to explain the probability of the event than linear probability



model. The space shuttle Challenger accident occurred on January 28, 1986. An O-ring that was used to seal a joint of the booster rockets experienced thermal distress. The Rogers Commission appointed by President R. Reagan to find the cause of the accident, found that the O-ring may not be sealed properly at low temperature. The O-ring failure data consist of 23 pre-accident launches of the space shuttle. Dalal, Fowlkes, and Hoadley (1989) discovered that the O-ring sealing function under the launching conditions is affected by the low temperatures by using the logistic regression model. This data set was also analyzed by Lavine (1991) and Martz and Zimmer (1992) using logistic regression model. The Challenger (space shuttle) was launched at different temperatures to study whether the explosion was due to low temperature condition. The launches are viewed as independent trials.

The graph (Figure1-1) show that the likelihood of failure decreases when the temperature increases with the linear probability model and logistic regression model. For the logistic regression model, all the estimated probabilities fall between 0 and 1. However, for the linear probability model, some of estimated probabilities are less than 0. Obviously, the logistic regression model makes more sense than the linear probability model does for this data set.

**Figure 1-1 Estimated Failure Probability from Logistic Regression Model and Probability Linear Model**



The red dots are the estimate probabilities from the logistic regression model, and the blue triangles are the estimate probabilities from the linear probability model.

## 1.2 The Goodness-of Fit Test for Logistic Regression Model

The goal of logistic regression models is used to model the probability of the occurrence of an event depending on the value of covariates  $\mathbf{x}$ . The model is of the following form:

$$\Pr(y = 1 | \mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{(\mathbf{x}\beta)}}{1 + e^{(\mathbf{x}\beta)}} \quad (1.4)$$

In other words, we want to find a model that fits the observed data well. A model is said to fit poorly if the model's residual variation is large and systematic (Hosmer et al, 1997). This is usually the case when the estimated values produced by the logistic regression model do not accurately reflect the observed values. There are many ways that cause the logistic regression model to fit the data inadequately, The most important of which involves the problem with the linear component (Collett, 1991), such as, omission of higher order terms of covariates, or important covariates related to the response variables from the model. Influential observations and outliers can also lead to a poor fit.

Goodness-of-fit or lack-of-fit tests are designed to determine formally the adequacy or inadequacy of the fitted logistic regression model. A poorly fitted model can give biased or invalid conclusions on the statistical inference based on the fitted model. Therefore, we must test the lack-of-fit of a model before we can use it to make statistic inferences.

Next, we will use a real data set to illustrate the importance of assessing the adequacy of a model. Vasoconstriction data is from Finney (1941) and Pregibon (1981). This data (Appendix B) consist of variables, the occurrence of vasoconstriction in the skin of the fingers, the rate and volume of air breathed. The end point of each test with 1 indicating that constriction occurred, 0 indicating that constriction did not occur. In this controlled experiment, 39 tests under various combinations of rate and volume of air inspired were obtained (Finney 1941). A logistic regression uses Volume and Rate of air breathed as covariates to explain and predict the probability of occurrence of vasoconstriction in the skin of the fingers.

**Table 1-3 Information on Covariates from Logistic Regression**

| <b>Analysis of Maximum Likelihood Estimates</b> |           |                 |                       |                        |                      |
|---|-----------|-----------------|-----------------------|------------------------|----------------------|
| <b>Parameter</b>                                | <b>DF</b> | <b>Estimate</b> | <b>Standard Error</b> | <b>Wald Chi-Square</b> | <b>Pr &gt; ChiSq</b> |
| <b>Intercept</b>                                | 1         | -9.5293         | 3.2331                | 8.6873                 | 0.0032               |
| <b>Rate</b>                                     | 1         | 2.6490          | 0.9142                | 8.3966                 | 0.0038               |
| <b>Volume</b>                                   | 1         | 3.8820          | 1.4286                | 7.3844                 | 0.0066               |

**Table 1-4 Goodness-of-Fit Test**

| <b>Hosmer and Lemeshow Goodness-of-Fit Test</b> |           |                      |
|---|-----------|----------------------|
| <b>Chi-Square</b>                               | <b>Df</b> | <b>Pr &gt; ChiSq</b> |
| 19.1837   | 8         | 0.0139               |

**Table 1-5 Model Fit Statistics from Logistic Regression Model**

|          | Model          | Fit                      | Statistics |
|----------|----------------|--------------------------|------------|
| Criteria | Intercept only | Intercept and Covariates |            |
| AIC      | 56.040         | 35.772                   |            |
| SC       | 57.703         | 40.763                   |            |
| -2 Log L | 54.040         | 29.772                   |            |

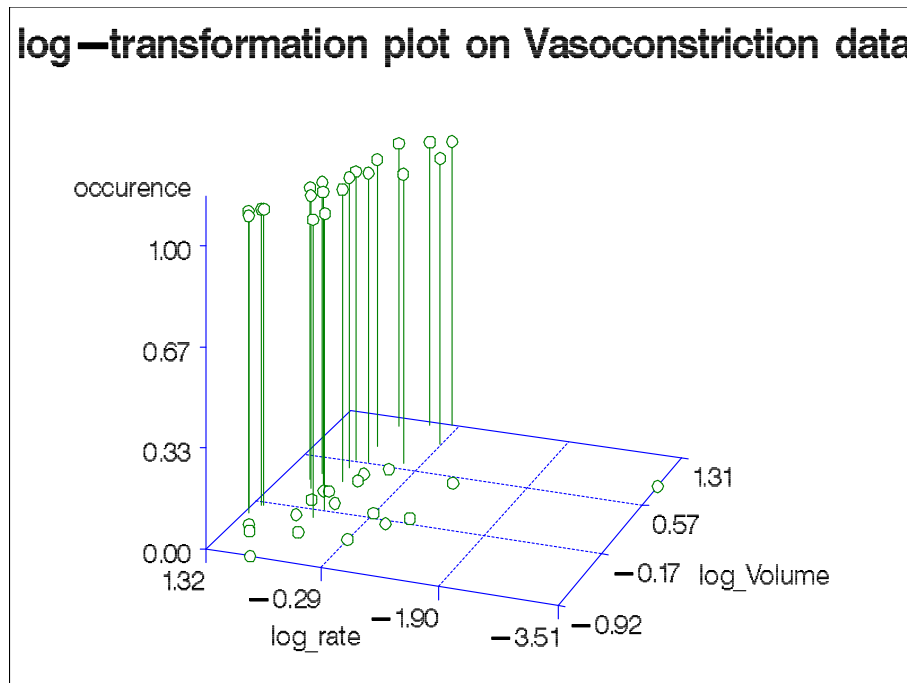
From Table 1-3, we may conclude that these two predictors do have significant effects on the occurrence probability of the vasoconstriction in the skin of the fingers, and we get the following model

$$\text{logit}(\pi(x)) = -9.5293 + 2.6490 * \text{rate} + 3.8820 * \text{volume} \quad (1.5)$$

The goodness-of-fit test result shown in Table 1-4 indicates the presence of lack-of-fit in model (1.5). Therefore, model (1.5) may not be reliable for us to predict the probability of occurrence of vasoconstriction. In order to get a suitable model, a higher order or interaction term may be needed, or latent predictors should be investigated in this experiment.

The natural log-transformation on Rate and Volume was used to search for an adequate model. The Figure 1-2 is the plot on the log-transformed data.

**Figure 1-2 Three Dimensional Plot on Transformed Vasoconstriction Data**



The statistical inference on the model parameters for the log-transformation data shown in Table1-6 suggests the following model

$$\text{logit}(\pi(x)) = -2.8754 + 4.5617 * \log(\text{rate}) + 5.1793 * \log(\text{volume}) \quad (1.6)$$

It fits the Vasoconstriction data better than model (1.5) based on the result of Hosmer and Lemeshow goodness-of-fit test (Table1-8).

AIC and BIC are two useful criteria to select a better model. Akaike Information Criterion (AIC) is proposed by Hirotosuge Akaike (Akaike, 1974), which is usually considered as a good criterion to compare several models.  $AIC = -2 * \ln L + 2 * k$ , where  $L$  is the maximized value of the likelihood function,  $k$  is the number of the parameters in the model (Akaike, 1974). Bayesian Information Criterion (BIC) is another important criterion, which provides more parsimonious model than AIC does. BIC (Bayesian information criterion) is also called SIC or SC (Schwarz information criterion or Schwarz

criterion), because Schwarz (Schwarz 1978) gave a Bayesian argument on it. In the logistic regression model selection,  $BIC = -2 * \ln L + k \ln(n)$ , where  $L$  is the maximized value of the likelihood function,  $k$  is the number of parameters in the model, and  $n$  is the sample size. The smaller the AIC and /or SC is, the better the model is. Model (1.6) is better than model (1.5) based on AIC and SC from Table 1-5 and Table 1-7.

**Table 1-6 Information on Covariates from Logistic Regression with Log-transformation**

| Analysis of Maximum Likelihood Estimates |    |          |                |                 |            |
|--|----|----------|----------------|-----------------|------------|
| Parameter                                | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept                                | 1  | -2.8754  | 1.3208         | 4.7395          | 0.0295     |
| Log_Rate                                 | 1  | 4.5617   | 1.8380         | 6.1597          | 0.0131     |
| Log_Volume                               | 1  | 5.1793   | 1.8648         | 7.7136          | 0.0055     |

**Table 1-7 Model Fit Statistics from Logistic Regression with Log-transformation**

| Model Fit Statistics |                |                          |
|----------------------|----------------|--------------------------|
| Criteria             | Intercept only | Intercept and Covariates |
| AIC                  | 56.040         | 35.227                   |
| SC                   | 57.703         | 40.218                   |
| -2 Log L             | 54.040         | 29.227                   |

**Table 1-8 Goodness-of-Fit Test with Log-transformation**

| <b>Hosmer and Lemeshow Goodness-of-Fit Test</b> |           |                      |
|---|-----------|----------------------|
| <b>Chi-Square</b>                               | <b>Df</b> | <b>Pr &gt; ChiSq</b> |
| 11.0873   | 8         | 0.1986               |

Unlike the linear regression model,  $R^2$  is not a suitable to measure the predictor power of the model in logistic regression, because we can't explain the variation in a binary dependent variable the same way as we do for continuous dependent variables. In many cases the  $R^2$  and /or pseudo  $R^2$  are small even when the model is adequate for the data (Hosmer and Lemeshow, 2000). Pearson's classical Chi-square test and Deviance test are well known, and they work very well when the covariates are categorical. When one or more covariates are continuous, the disadvantages of Pearson chi-square test and Deviance test provide incorrect p-values.

Consequently, many goodness-of-fit tests of logistic regression model are developed. For example, the  $\hat{C}$  and  $\hat{H}$  tests proposed by Hosmer and Lemeshow (1980 and 1982). Pulksteris and Robinson (2002) also proposed two test statistics dealing with the situation in which both discrete and continuous covariates are involved in the logistic regression model. However, their methods can not handle the situation when all the covariates are continuous. Brown (1982), Tsiatis (1980) and Stukel (1988) proposed tests derived from score test, respectively. These score tests are mainly designed to test overall goodness-of-fit for the logistic regression model. Osius and Rojek (1992) proposed normal goodness-of-fit tests for multinomial models with many



degrees of freedom, which can be applied to the binary cases. This test can be viewed as an extension of Pearson Chi-square test. le Cessie and van Houwelingen (1991) proposed a goodness-of-fit test based on the smoothing of the residuals techniques proposed by Copas (1983), Landwehr et al (1984), Fowlkes (1987) and Azzalini et al. (1989). Hosmer et al (1997) developed another smoothed residual test following the step of le Cessie and van Houwelingen (1991). However, Hosmer et al. (1997) showed that smoothed residual tests are not very powerful compared with Stukel's test and Hosmer-Lemeshow's  $\hat{C}$  test.

In spite of the fact that many goodness-of-fit tests have been proposed by researchers in recent decades, none of them can be considered as the universal best one in assessing the adequacy or inadequacy of the logistic regression model. Each proposed test has its advantages or disadvantages.

In this dissertation, we proposed a new goodness-of-fit test for the logistic regression. The proposed test is expected to perform as well as or better than the known tests. The proposed test can identify the nature of the lack-of-fit. The proposed partition logistic regression model can be used as a generalized logistic regression model to improve the fit of a standard logistic regression model.

The rest of the chapters are organized as follows. Chapter 2 reviews several primary overall goodness-of-fit tests including their theoretical developments, advantages and drawbacks. Chapter 3 presents the proposed method, and applications of the proposed test to several real data sets, to illustrate the performance of the proposed test. Chapter 4 gives several simulations studies under different setting to evaluate and compare the proposed tests with other known tests. Chapter 5 and

Chapter 6 summarize the finding obtained in this research and some possible extensions of the present work in a future study.

## CHAPTER 2 - Literature Review

Many methods on assessing the goodness-of-fit for logistic regression models have been developed recently (Evans and Li 2005). The current methods can be roughly categorized into four types. (1) The tests are based on covariate patterns, which includes Pearson's Chi-square  $\chi^2$  test (Pearson 1900), Deviance D test, and Osius and Rojek's normal approximation test (Osius and Rojek 1992). (2) Hosmer and Lemeshow's  $\hat{C}$  and Hosmer and Lemeshow's  $\hat{H}$  tests (Hosmer and Lemeshow 1980) use grouping of event estimated probability from the assumed model. (3) Score tests are based on the comparison of two models, where the assumed logistic regression model is embedded in a more general parametric family of models (Prentice 1976). The score tests contains Brown's Score test (Brown 1982), Stukel's Score test (Stukel 1988) and Tsiatis's test (Tsiatis 1980). (4) Smoothed residual tests include le Cessie and van Howelingen's  $\hat{T}_{lc}$  test (Cessie and Howelingen 1991) and Hosmer and Lemeshow's  $\hat{T}_{rc}$  test (Hosmer et al. 1997). These two methods have similar performances, Hosmer et al (1997) indicate they are not better than Stukel's score test,  $\hat{C}$  test, and  $\hat{H}$  test.

In this dissertation, six well known goodness-of-fit tests for logistic regression model will be compared with the proposed test. The six tests are the Pearson Chi-square

$\chi^2$  test (Pearson 1900), Deviance D test, Hosmer and Lemeshow's  $\hat{C}$  test, Hosmer and Lemeshow's  $\hat{H}$  test (Hosmer and Lemeshow 1980), the Osius and Rojek normal approximation test (Osius and Rojek 1988) and Stukel's score test (Stukel 1988) . The first three methods are adopted in several commercial statistical packages such as, SAS, SPSS, GLIM, S-plus/R. The Hosmer and Lemeshow's  $\hat{C}$  , the Osius and Rojek normal approximation test and Stukel's score test are considered as better methods for overall assessment of goodness-of-fit tests when continuous covariates are present in the model. They are recommended by several researchers (Hosmer and Lemeshow 2000).

## 2.1 Pearson Chi-square $\chi^2$ Test and Deviance D Test

In logistic regression, there are several goodness-of-fit tests obtained by comparing the overall difference between the observed and fitted values. Among these tests Pearson Chi-Square  $\chi^2$  and Deviance D test statistic are used most often among. Pearson chi-square  $\chi^2$  goodness-of-fit test was proposed by Karl Pearson (Pearson 1900). This method made a revolutionary impact on the categorical data analysis. We will first define the items needed to describe these two tests.

Suppose that  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  are independent pairs of observations obtained from  $n$  subjects. For the  $i^{\text{th}}$  subject, the response  $y_i$  is a Bernoulli variable with probability success  $\pi_i$  and  $x_i$  is a single set of values of independent (predictor or explanatory) variables called covariates associated with the  $i^{\text{th}}$  response  $y_i$ . The logit of the logistic regression model is the linear predictor:

$$\eta(x_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, i = 1, 2, \dots, n, \quad (2.1)$$

where  $\pi(x) = \Pr(y = 1 | x)$  is the conditional probability that  $y$  is 1 given  $x$ , and

$1 - \pi(x) = \Pr(y = 0 | x)$  is the conditional probability that  $y$  is 0 given  $x$ .  $\frac{\pi(x)}{1 - \pi(x)}$  is called

the odds associated with the set of covariate  $x$  in the logistic regression model.

The likelihood function for the pair  $(x_i, y_i)$  is

$$\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (2.2)$$

The likelihood function and log-likelihood function of  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  are respectively,

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (2.3)$$

$$\text{Log}L(\beta) = \sum_{i=1}^n \{y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\} \quad (2.4)$$

The value of  $\beta$ ,  $\hat{\beta}$  that maximizes the equation (2.3) or (2.4) is called the maximum likelihood estimate of  $\beta$ . The estimated probability  $\hat{\pi}(x_i)$  and estimated logit  $\hat{\eta}(x_i)$  are obtained by the maximum log likelihood method. Thus,  $\hat{\pi}(x_i)$  is the maximum likelihood estimate of  $\pi(x_i)$ , and  $\hat{\eta}(x_i)$  is the maximum likelihood estimate of the linear predictor  $\eta(x_i)$ .

Suppose the fitted model has p-dimensional covariates  $\mathbf{x} = (x_1, x_2, \dots, x_p)$ . A single covariate pattern is defined as a single set of values of the covariates used in the model. For example, if two covariates, gender (Female and Male) and class (Freshmen, Sophomore, Junior, Senior) are used in a data set, then the maximum number of distinct covariate patterns is eight. There are essentially two types of covariate patterns. Type one pattern, there are no tied covariates which indicates that each subject has a unique set of covariate values, and the number of covariate patterns J is equal to the number of subjects, i.e.,  $J=n$ . This is very common, when only continuous covariates are involved in the logistic regression model. The data set is considered to be sparse in this situation (Kuss 2002). For the type two pattern, some subjects have tied covariates, that is, they have the same covariate values, making the number of covariate patterns less than the number of subjects, i.e.,  $J < n$ . In the type two pattern, let

the total number of successes (  $y=1$  ) be  $n_1$  and the total number of failures (  $y=0$  ) be  $n_0$ , and it follows that  $n_1 + n_0 = n$ . Suppose  $m_j$  subjects in the  $j^{\text{th}}$  covariate pattern have the same covariate values  $x_j$ , then,  $\sum_{j=1}^J m_j = n$ . Let the  $y_{j1}$  denote the number of successes in the  $j^{\text{th}}$  group with  $j^{\text{th}}$  covariate pattern. It follows that  $\sum_{j=1}^J y_{j1} = n_1$ . Similarly, let the  $y_{j0}$  denote the number of failures observed by subjects in the  $j^{\text{th}}$  group with  $j^{\text{th}}$  covariate pattern. It follows that  $\sum_{j=1}^J y_{j0} = n_0$ . Under type two covariate pattern, the binary outcomes can be represented by a J by 2 frequency table. The two columns of the table corresponding to the response variable,  $y=1,0$  and the J rows corresponding to J possible covariate patterns.

Let  $\hat{\pi}_j$  be the maximum likelihood estimate of  $\pi_j$  associated with the  $j^{\text{th}}$  covariate pattern, then the expected number of success observed by the subject in the  $j^{\text{th}}$  group with  $j^{\text{th}}$  covariate pattern is

$$\hat{y}_{j1} = m_j \hat{\pi}(x_j)$$

The likelihood function (2.3) and the log-likelihood function (2.4) can be written respectively as below for type two covariate pattern

$$L(\beta) = \prod_{j=1}^J \binom{m_j}{y_{j1}} (\pi(x_j))^{y_{j1}} (1 - \pi(x_j))^{m_j - y_{j1}} \quad (2.5)$$

$$\text{Log}L(\beta) = \sum_{j=1}^J \left\{ \log \binom{m_j}{y_{j1}} + y_{j1} \log \pi(x_j) + (m_j - y_{j1}) \log(1 - \pi(x_j)) \right\} \quad (2.6)$$

The Pearson residual (Hosmer and Lemeshow 1989) is defined as

$$r(y_{j1}, \hat{\pi}(x_j)) = \frac{(y_{j1} - m_j \hat{\pi}(x_j))}{\sqrt{m_j \hat{\pi}(x_j)(1 - \hat{\pi}(x_j))}} \quad (2.7)$$

$$\text{The Pearson Chi-Square test statistic is } \chi^2 = \sum_{j=1}^J r(y_{j1}, \hat{\pi}(x_j))^2 \quad (2.8)$$

If the fitted model is correct, the sampling distribution of the Pearson chi-Square test statistic can be approximated by a Chi-Square distribution with degrees of freedom  $J - (p+1)$ , where  $p$  is the number of the parameters in the model. For the type one pattern case with  $J=n$ . Pearson's chi-square (McCullagh and Nelder 1989) test statistic is not an applicable goodness-of-fit test, because the test statistic will not have a Chi-square distribution (Christensen 1997). This due to the fact that when  $m_j=1$ , one subject for each covariate pattern, the Pearson residual does not asymptotically have a normal distribution for large  $n$ .

The deviance as a measure of goodness-of-fit was first proposed by Nelder and Wedderburn (1972). The deviance residual for the  $j^{\text{th}}$  covariate pattern is defined as (Hosmer and Lemeshow 1989)

$$d(y_{j1}, \hat{\pi}(x_j)) = \pm \{2[y_{j1} \ln(\frac{y_{j1}}{m_j \hat{\pi}(x_j)}) + (m_j - y_{j1}) \ln(\frac{(m_j - y_{j1})}{m_j(1 - \hat{\pi}(x_j))})]\}^{1/2} \quad (2.9)$$

where the sign of the deviance residual is the same as that of  $(y_{j1} - m_j \hat{\pi}(x_j))$ . The

$$\text{Deviance test statistic is } D = \sum_{j=1}^J d(y_{j1}, \hat{\pi}(x_j))^2 \quad (2.10)$$

If the model is correct, the test statistic of Deviance has approximately a Chi-Square distribution with degree of freedom  $J - (p+1)$ .

Under the null model and Type one covariate pattern, the deviance residual is reduced to



$$d(y_{i1}, \hat{\pi}(x_i)) = \pm \{2[y_{i1} \ln(\frac{y_{i1}}{\hat{\pi}(x_i)}) + (1-y_{i1}) \ln(\frac{(1-y_{i1})}{(1-\hat{\pi}(x_i))})]\}^{1/2} \quad (2.11)$$

Deviance residual measures the difference of the log likelihood between the assumed model and the saturated model.  $y_{i1}$  has two values 0 and 1, which indicates that  $y_{i1} \log y_{i1}$  and  $(1-y_{i1}) \log(1-y_{i1})$  will be both 0. Then under Type one covariate pattern the test statistic of Deviance can be expressed as (Collett 1991)

$$D = -2 \sum_{i=1}^n \{ \hat{\pi}(x_i) \log \hat{\pi}(x_i) + (1-\hat{\pi}(x_i)) \log(1-\hat{\pi}(x_i)) \} \quad (2.12)$$

Note that the test statistic (2.12) of Deviance does not involve in the comparison of the observed and fitted frequency of success but involves a comparison of the log of the maximum likelihood of the saturated model and the assumed model.

Hosmer and Lemeshow (Hosmer and Lemeshow 2000) reported that the p-values of the Pearson Chi-square test and Deviance test are not correct under type one covariate pattern for which  $J=n$ . The major advantage of Deviance test and Pearson Chi-square is their elementary calculation of the test statistics and the associated p-value. When the Pearson Chi-square  $\chi^2$  and Deviance D have different conclusions on the data set, extreme caution should be taken, because it may indicate the Chi-square distribution ( $df=J-(p+1)$ ) may not approximate the sampling distribution of Pearson Chi-square and Deviance test statistic (Collett 1991).

Usually, the Deviance test is preferred to the Pearson Chi-square test for the following reasons. (1) The logistic regression model is fitted by the maximum likelihood method, and the maximum likelihood estimates of the success probabilities minimize the test statistic D (Collett 1991). (2) The Deviance test can be used to compare a

sequence of hierarchical logistic models, but the Pearson Chi-square test can not be used in this way (Collett 1991).

## 2.2 The Hosmer and Lemeshow Tests

Hosmer and Lemeshow developed several different methods to test the lack-of-fit of a logistic regression model. We will only consider two widely known tests which group the subjects based on the estimated probabilities of success. The test statistic,  $\hat{H}$ , is calculated based on the fixed and pre-determined cut-off points of the estimated probability of success (Hosmer and Lemeshow 1980). The test statistic,  $\hat{C}$ , is calculated based on the percentiles of estimated probabilities (Lemeshow and Hosmer 1980). Only the test statistic  $\hat{C}$  method is well accepted and is included in several major statistical packages.

The tests proposed by Hosmer and Lemeshow (Hosmer and Lemeshow 1980, Lemeshow and Hosmer 1982) do not require the number of covariate patterns less than the total number of the subjects.

### 2.2.1 Hosmer and Lemeshow's $\hat{C}$

In this method, the subjects are grouped into  $g$  groups with each group containing  $n/10$  subjects. The number of groups  $g$  is about 10 and can be less than 10, due to fewer subjects. Ideally, the first group contains  $n'_1 = n/10$  subjects having the smallest estimated success probabilities obtained from the fitted assumed model. The second group contains  $n'_2 = n/10$  subjects having the second smallest estimated success probabilities, and so on. Let  $\bar{\pi}_k$  be the average estimated success probability based on the fitted model corresponding to the subjects in the  $k^{\text{th}}$  group with  $y=1$ , and

let  $o_k$  be the number of subjects with  $y=1$  in the  $k^{\text{th}}$  group. We have a  $g$  by 2 frequency table with the two columns of the table corresponding to the two values of the response variable,  $y=1,0$  and the  $g$  rows corresponding to the  $g$  groups. The formula of Hosmer and Lemeshow test statistic  $\hat{C}$  is

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (2.13)$$

where  $\sum_{k=1}^g n_k = n$ , and  $n_k$   $k = 1, 2, \dots, g$ , is the total subjects in the  $k^{\text{th}}$  group.

Under the null hypothesis, the test statistic,  $\hat{C}$ , is approximately distributed as a Chi-square distribution with  $g-2$  degrees of freedom (Lemeshow and Hosmer, 1980).

The equation  $n=10*w$  does not always hold, where  $n$  is the total number of subjects and  $W$  is an arbitrary positive integer. There is a tendency to ensure that each group has fair number of subjects. The test statistic,  $\hat{C}$ , may be sensitive to the cut-off points specified to form the groups. To illustrate the sensitivity of  $\hat{C}$ , Hosmer et al. (Hosmer et al. 1997) used 6 different packages to carry out lack-of-fit test on low birth weight data (Hosmer and Lemeshow 1989). The six different packages did produce the same fitted model with same coefficients; however, these different packages provided six different  $\hat{C}$  values associated with six different p-values.

**Table 2-1 the Results of Hosmer-Lemeshow  $\hat{C}$  Test  
from Different Packages**

| Package   | $\hat{C}$ | df | p-value |
|-----------|-----------|----|---------|
| BMDPLR    | 18.11     | 8  | 0.02    |
| LOGXACT   | 13.08     | 8  | 0.109   |
| SAS       | 11.83     | 8  | 0.159   |
| STATA     | 12.59     | 8  | 0.127   |
| STATISTIX | 12.11     | 8  | 0.147   |
| SYSTAT    | 14.70     | 8  | 0.065   |

The above results are not surprising, since different statistical packages have their own algorithms to determine the cut-off points. Bertolini et al. (2000) pointed out that the paradox of Hosmer and Lemeshow's goodness-of-fit test may occur when ties are present in the covariates. They also pointed out that Hosmer and Lemeshow goodness-of-fit test results may be inaccurate, when the number of covariate patterns is less than number of subjects.

### **2.2.2 Hosmer and Lemeshow $\hat{H}$**

This test was developed by Hosmer and Lemeshow(1980), in which the subjects are grouped into 10 groups if the estimated probabilities cover 0-1. The grouping method is described as the following: the subjects are in the first group whose estimated probabilities fall in the range from 0 to 0.1; the subjects are in the second group whose

estimated probabilities fall in the range from 0.1 to 0.2, other groups have the same grouping policy. The number of subjects is not always the same among groups. In many cases, the range of estimated probabilities can only cover a small subset of (0,1), which resulted in the group numbers less than 10.

The Hosmer and Lemeshow  $\hat{H}$  test statistic can be obtained by computing the Pearson chi-square statistic from a g by 2 frequency table considered in the  $\hat{C}$  test. It has exactly the same form as  $\hat{C}$  test:

$$\hat{H} = \sum_{k=1}^g \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (2.14)$$

where  $o_k$ ,  $\bar{\pi}_k$  and  $n'_k$  have the same definition as in (2.13). For large n, the distribution can be approximated by a Chi-square distribution with g-2 degrees of freedom under the null hypothesis.

The simulation study conducted by Hosmer and Lemeshow (Lemeshow and Hosmer 1982) compare their  $\hat{C}$  and  $\hat{H}$  tests. The results seem to suggest that the  $\hat{H}$  test is more powerful than  $\hat{C}$  test, and thus,  $\hat{H}$  test was the preferred test. However, additional study from Hosmer, Lemeshow, and Klar (1988) showed that  $\hat{C}$  test is better than  $\hat{H}$ , especially when there are a lot of estimated probabilities that are less than 0.2 (Hosmer and Lemeshow 2000).

From (2.13) and (2.14), the Hosmer-Lemeshow's  $\hat{C}$  and  $\hat{H}$  test can be considered as extensions of Pearson chi-square test on the basis of merging multiple covariate patterns into one group, which makes the distribution of the Pearson residuals approximately normal.

## 2.3 Osius and Rojek Normal Approximation Test

This method derived by Osius and Rojek (Osius and Rojek 1992), can be applied not only to binary cases but also to multinomial cases. The test statistic is obtained by modification of Pearson Chi-square test statistic, and it is approximately normally distributed when the number of subjects is large.

The procedure to obtain this test statistic is described for the type two covariate pattern (the number of covariate patterns are less than the number of subjects). Let the subjects have  $J$  covariate patterns, and let the estimated probability of the  $j^{\text{th}}$  covariate pattern be  $\hat{\pi}_j$  based on the assumed model,  $j=1,2,\dots,J$ , then the variance of number of successes in  $j^{\text{th}}$  covariate pattern is  $v_j = m_j \hat{\pi}_j (1 - \hat{\pi}_j)$ , where  $m_j$  is the number of subjects in the  $j^{\text{th}}$  covariate pattern.

For  $j=1,2,\dots,J$ , let  $c_j = \frac{(1-2\hat{\pi}_j)}{v_j}$ . The  $v_j$ s are used as the weights to perform an ordinary linear regression of  $c_j$  ( $j=1,2,\dots,J$ ) on the covariates  $x_j$  ( $j=1,2,\dots,J$ ), and calculate the residual sum squares  $RSS$ . In this ordinary weighted linear regression, total sample size equals to number of covariates patterns  $J$  instead of total subjects  $n$ . The test statistic  $z$  is denoted as

$$z = \frac{[\chi^2 - (J - p - 1)]}{\sqrt{A + RSS}} \quad (2.15)$$

where  $A = 2(J - \sum_{j=1}^J \frac{1}{m_j})$  is the correction factor for the variance and  $\chi^2$  is the Pearson

Chi-square test statistic given in (2.8). The test statistic Z approximately follows the standard normal distribution when the sample size is large.

This method is applicable to the type one covariate pattern ( $J=n$ ) and the type two covariate pattern ( $J<n$ ). The correction factor A will be zero when  $J=n$ . The Pearson Chi-square can be inflated by the very small or large estimated probabilities (Windmeijer 1990), thus the test statistic Z is affected by this characteristic as well. Windmeijer (Windmeijer 1990) suggested that the subjects with extreme estimated probabilities should be excluded in order to construct a suitable test statistic.



## 2.4 Stukel's Score Test

Stukel's score test is based on the comparison between the assumed logistic regression model and a more general logistic regression model. This general logistic regression model has two additional parameters (Stukel 1988). Thus, this test is called a two degree-of-freedom test (Hosmer and Lemeshow, 2000). Each one of the two additional parameters independently affects the behavior of one of the tails of the probability of success curve.

Let the assumed logistic regression model is the same as (1.3). Let the maximum likelihood estimates of  $\pi_i = \pi(x_i)$  and  $\eta_i = \eta(x_i)$  based on the assumed model be  $\hat{\pi}_i$  and  $\hat{\eta}(x_i)$   $i = 1, 2, 3, \dots, n$ , respectively. Two new covariates are created as follow:  
 $z_{1i} = 0.5 * \hat{\eta}(x_i) * I(\hat{\pi}_i \geq 0.5)$  and  $z_{2i} = -0.5 * \hat{\eta}(x_i) * I(\hat{\pi}_i < 0.5)$ ,  $i = 1, 2, 3, \dots, n$ , where  $I$  is an indicator function.

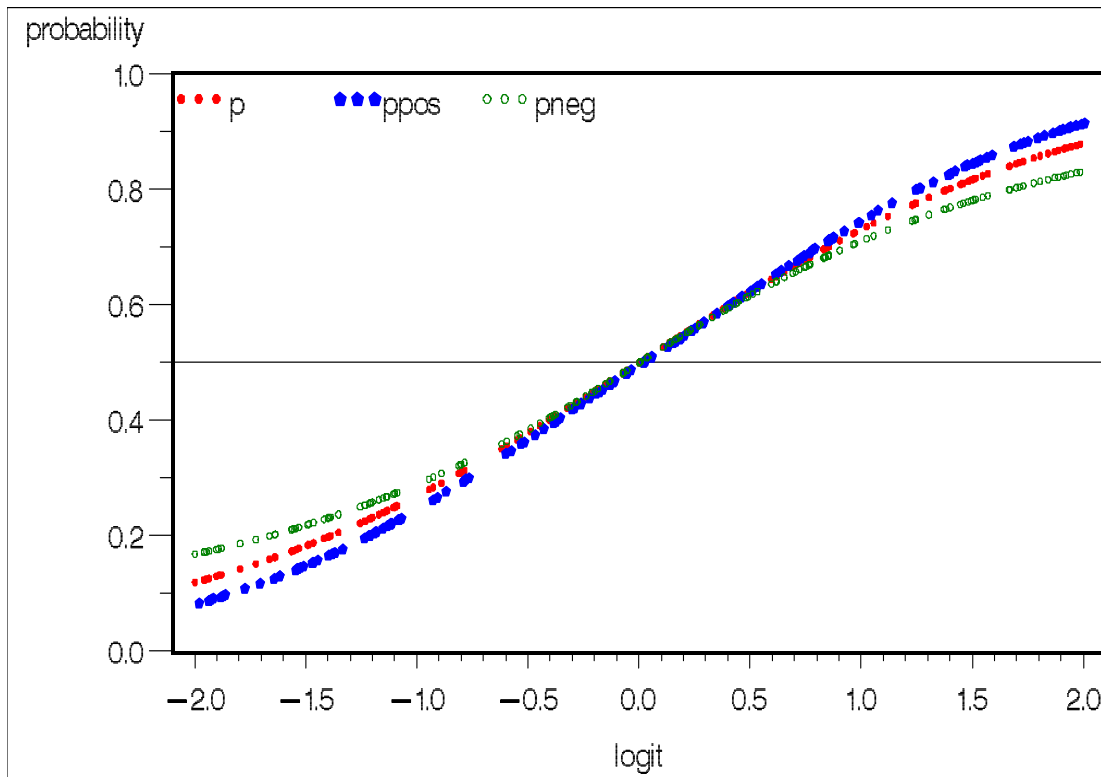
The generalized logistic regression model is obtained by adding the two covariates  $z_{1i}$  and  $z_{2i}$  in the logit of the assumed model. That is,

$$\text{logit}(\pi_i) = \eta(x_i) + \alpha_1 z_{1i} + \alpha_2 z_{2i} \quad (2.16)$$

The null hypothesis is expressed as  $H_0 : \alpha_1 = \alpha_2 = 0$ . If  $\alpha_1, \alpha_2$  are larger than zero, the estimated tail of the successes probabilities curve of model (2.16) converges to their respective asymptotes ( $\pi = 0, \pi = 1$ ) more quickly than the assumed model (1.3), otherwise, estimated tail of successes probabilities curve reach its asymptote more slowly than that of assumed model. If  $\alpha_1 = \alpha_2$ , the curve of estimated probabilities for the generalized logistic regression model will be symmetric about  $\text{logit}=0$  and,

probability= 0.5. By this property, the nature of lack of fit on the assumed logistic regression model can be viewed by plotting of the fitted probability of success curve.

**Figure 2-1 the Plot of  $\pi(x)$  versus  $\eta(x)$  for Stukel's Model**



In Figure 2-1, the Red dot line represents the assumed model with  $\alpha_1 = \alpha_2 = 0$  , the blue diamond line represents model (2.16) with  $\alpha_1 = \alpha_2 = 0.2$  and the green circle line represents the model (2.16) with  $\alpha_1 = \alpha_2 = -0.2$ .

Let  $l(X)$  be the maximum log-likelihood from the assumed model (1.3), and  $l(X,Z)$  be that from the generalized logistic regression model (2.16). The test statistic

is  $ST = -2l(X) - (-2l(X, Z))$  Under  $H_0$ , the test statistic  $ST$  has a limiting chi-square distribution with 2 degrees of freedom as sample size increases to  $\infty$ .

Hosmer and Lemeshow (1997) indicated that this method strictly speaking is not a goodness-of-fit test, since the test statistic is not based on the comparison of the observed and expected values. However, they do agree that this method in general is more powerful in detecting lack-of-fit than the known methods in many situations and it controls the type I error rate (Hosmer et al 1997).

## 2.5. An Illustrative Example

The data are given in Appendix C on the survival of 33 leukemia patients, which were also analyzed by Feigl and Zelen (1965), Cook and Weisberg (1982) and John (1985) using the logistic regression model. In this data set, the binary response is survival status 52 weeks after the time of diagnosis. “1” is used to denote that the patient is alive, otherwise, “0” is used. Two important covariates of interest are log transform of the white blood cell count and action of the presence or absence of certain morphological characteristic denoted by the two levels (AG+, AG-). A logistic regression model with the follow function is fitted to the data:

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1(\log(\text{cell})) + \beta_2(AG) \quad (2.17)$$

Estimates of the model parameters are given in Table 2-2, and the goodness-of-fit test results from different methods are showed in Table 2-3.

**Table 2-2 Parameter Estimates of Leukemia Data**

| Analysis of Maximum Likelihood Estimates |    |          |                |                 |            |
|--|----|----------|----------------|-----------------|------------|
| Parameter                                | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept                                | 1  | 9.4001   | 4.2076         | 4.9911          | 0.0255     |
| Log(Cell)                                | 1  | -2.5653  | 1.0654         | 5.7981          | 0.0160     |
| AG                                       | 1  | 1.2622   | 0.5462         | 5.3409          | 0.0208     |

**Table 2-3 Result from Six Goodness-of-Fit Tests on Leukemia Data**

| Test           | df | statistic | P-value |
|----------------|----|-----------|---------|
| HL $\hat{C}$   | 9  | 8.0289    | 0.5312  |
| HL $\hat{H}$   | 8  | 9.9491    | 0.2686  |
| Deviance       | 27 | 22.9544   | 0.6875  |
| Pearson        | 27 | 19.74     | 0.8415  |
| Osious & Rojek |    | 0.4357    | 0.6630  |
| Stukel Score   | 2  | 0.4158    | 0.8123  |

From Table 2-3, these six different goodness-of-fit tests yield the same conclusion,

$$\text{logit}(\pi(x)) = 9.4001 - 2.5653 * (\log(\text{cell})) + 1.2622 * AG \quad (2.18)$$

which suggests that model (2.18) fits the data adequately. Clearly, different tests have different p-values. Hosmer-Lemeshow's  $\hat{H}$  test yields the smallest p-value with 0.2686 and the Pearson-chi-square test gives the largest p-value with 0.8415. It is worth noting that Hosmer-Lemeshow  $\hat{C}$  groups the data points into 11 groups instead of 10 groups in order to get equal number of subjects within each group using SAS software. The grouping scheme is showed as Table 2-4.

**Table 2-4 Group Scheme of Hosmer-Lemeshow 's  $\hat{C}$  test with SAS Software on Leukemia Data**

| <b>Partition for the Hosmer and Lemeshow Test</b> |       |          |          |          |          |
|---|-------|----------|----------|----------|----------|
| Group   | Total | y = 1    |          | y = 0    |          |
|   |       | Observed | Expected | Observed | Expected |
| 1   | 3     | 0        | 0.03     | 3        | 2.97     |
| 2   | 3     | 0        | 0.11     | 3        | 2.89     |
| 3   | 3     | 0        | 0.15     | 3        | 2.85     |
| 4   | 3     | 1        | 0.31     | 2        | 2.69     |
| 5   | 3     | 0        | 0.42     | 3        | 2.58     |
| 6   | 3     | 1        | 0.68     | 2        | 2.32     |
| 7   | 3     | 1        | 0.87     | 2        | 2.13     |
| 8   | 3     | 1        | 1.53     | 2        | 1.47     |
| 9   | 3     | 3        | 1.91     | 0        | 1.09     |
| 10  | 3     | 1        | 2.27     | 2        | 0.73     |
| 11  | 3     | 3        | 2.72     | 0        | 0.28     |

## CHAPTER 3 - The Proposed Method

Hosmer and Lemeshow's (1980) and Lemeshow and Hosmer (1982) proposed two goodness-of-fit tests: the  $\hat{C}$  test and the  $\hat{H}$  test based on the grouping of estimated probabilities obtained from the assumed logistic regression model. Hosmer and Lemeshow's  $\hat{C}$  test which is based on the percentile grouping is better than Hosmer and Lemeshow's  $\hat{H}$  test which is based on the predetermined cut-off point especially when many estimated probabilities are small. (Hosmer Lemeshow and Klar 1988).

Stukel's score test was based on the comparison between the assumed logistic regression model and a general logistic model with two additional parameters (Stukel, 1988). But the power was not studied by Stukel. Hosmer and Lemshow (Hosmer and Lemshow, 2000) commented that Stukel's test was a good overall assessment method, besides Hosmer and Lemeshow  $\hat{C}$  test, and Ouis and Rojek approximately normal distribution test.

Pulkstenis and Robinson (Pulkstenis and Robinson, 2002) proposed two goodness-of-fit tests, which are very similar to the Deviance and Pearson chi-square tests. Their different grouping method was based on covariate pattern: covariate patterns are determined only by the categorical covariates. Pulkstenis and Robinson's methods are not applicable to the completely sparse data set in which each observation

has its own covariate pattern (2002), although they showed their proposed method is more powerful than Hosmer and Lemeshow's  $\hat{C}$  test under the type two covariate pattern.

The known goodness-of-fit tests for logistic regression model are not perfect even though they might be adopted methods in commercial software. Bender and Grouven (1996) suggested that medical journals might publish papers about misleading results and misinterpretation of statistical inferences due to lack-of-fit of the logistic regressions. This motivates us to search for a better goodness-of-fit test to assess the adequacy of the logistic regression model in explaining and predicting the responses.

We proposed a partition logistic regression model, which can be used to detect the overall lack-of-fit, between groups lack-of-fit and within groups lack-of-fit. These concepts will be described in the later sections of the chapter. It is recommended that the overall goodness-of-fit test should be performed first, and if the null hypothesis is not rejected, we simply conclude that we fail to detect any lack of fit for the assumed model; otherwise, the between and within group goodness-of-fit test should be carried out in order to figure out which type of lack of fit is present in the assumed model.



### 3.1 Description of the Proposed Test

The response variable could be multinomial case or binary case in the logistic regression model. Here, we will consider only the binary case. Suppose we have  $n$  pairs of observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , where  $y_i$  follows a Bernoulli distribution and  $x_i$  is a set of covariate value associated with  $y_i$ . The assumed logistic regression model is

$$\text{logit}(\pi(x_i)) = x_i' \beta \quad i = 1, 2, \dots, n, \quad (3.1)$$

where  $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  denotes  $p+1$  regression parameters and

$x_i' = (1, x_{1i}, x_{2i}, \dots, x_{pi})$  represents a set of values of the  $p+1$  covariates for the  $i^{\text{th}}$  subject.

The estimates of the parameters are obtained by the maximum likelihood method and denoted by  $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ . The assumed model (3.1) may or may not fit the data very well. We say that lack-of-fit is present in the model (3.1) if the “true” unknown model for the data is

$$\text{logit}(\pi(x_i)) = x_i' \beta + w_i' \delta \quad i = 1, 2, \dots, n, \quad (3.2)$$

Of course,  $w_i' \delta$  is not known to us. We propose to approximate the true model by a partition (segmented) logistic regression model. The approximately true model is constructed as follow

Partition the probability of success into  $M$  mutually disjoint intervals:

$$S_m = (\pi_{m-1}, \pi_m), m = 1, 2, \dots, M (M \geq 2), \text{ where } \pi_M = 1, \text{ and } \pi_0 = 0.$$

Partition the data points  $D = \{(x_i, y_i), i = 1, 2, \dots, n\}$  into  $M$  mutually disjoint groups:

$$D_m = \{(x_{mj}, y_{mj}), j = 1, 2, \dots, n_m, \pi(x_{mj}) \in S_m \ (m = 1, 2, \dots, M)\}, \text{ where } x_{mj} \text{ is the } j^{\text{th}} \text{ set of}$$

covariate value in the  $m^{\text{th}}$  group. Of course, we do not know  $\pi(x_{mj})$ . Hence, we estimate  $\pi(x_{mj})$  by the maximum likelihood method from the assumed model (3.1).

The following family of partition logistic regression model is used to approximate the unknown true model (3.2).

$$\logit(\pi(x_{mj})) = x'_{mj}\beta + z'_{mj}\alpha_m \quad m = 1, 2, \dots, M, \quad j = 1, 2, \dots, n_m \quad (3.3)$$

where  $z_{mj} = x_{mj}I((x_{mj}, y_{mj}) \in D_m)$  and  $\sum_{m=1}^M n_m = n$

The  $x'_{mj}$  in the definition of  $z'_{mj}$  can be replaced by the vector whose components are squared or the cross products of the components of  $x_{mj}$  or any function of the components  $x_{mj}$ .  $\beta$  and  $\alpha_i$  are vectors of the regression coefficients of the components of  $x'_{mj}$  and  $z'_{mj}$ , respectively.  $z'_{mj}$  contains '1' functioning as the covariates associated with the intercept for the  $m^{\text{th}}$  group.

$$\text{Let } X = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ X_M \end{bmatrix} \text{ where } X_m = \begin{bmatrix} x_{m1} \\ x_{m2} \\ \cdot \\ x_{mn_m} \end{bmatrix} \text{ and } m = 1, 2, \dots, M$$

matrix Z consists of  $z'_{mj}$  defined similarly as matrix X. Ideally, we want the column space of (XZ) to contain the column space of (XW), i.e.,  $C(XW) \subseteq C(XZ)$ . Generally, (XZ) is not of full column rank. We assume that for sufficiently large M, the full model (3.3) should approximate the unknown true model (3.2), that is,  $C(XW)$  is approximately equal to  $C(XZ)$ . Hence, in principal we would like to choose M as large as possible with the restriction that the group size is large enough to fit each sub-model of (3.3).

For the logistic regression model with only continuous covariates,  $M=2$  or  $3$ , seems to be sufficient for the data we have considered. If the model contains both continuous and categorical covariates, we partition each category group defined by the categorical variables into two to three subgroups.

Let  $L(X, Z)$  and  $L(X)$  be the maximum likelihood obtained from the partition model (3.3) and the assumed model (3.1), respectively. If the assumed model (3.1) is the true model--that is,  $\delta = 0$  in the true model (3.2), then  $-2 \log L(X) - (-2 \log L(X, Z))$  follows an asymptotic Chi-square distribution with degrees of freedom  $df = \text{rank}(XZ) - \text{rank}(X)$ . This is an overall goodness-of-fit test for the assumed logistic regression model (3.1).

Consider a partition model for which models from different groups apart from having different intercepts, they all have the same regression coefficients

$$\text{logit}(\pi(x_{mj})) = x'_{mj} \beta + u_m \quad m = 1, 2, \dots, M, \quad j = 1, 2, \dots, n_m \quad (3.4)$$

Let  $L(X, u)$  be the maximum likelihood of the model (3.4). Write the overall lack-of-fit test statistic as

$$\begin{aligned} & -2 \log L(X) - (-2 \log L(X, Z)) \\ &= -[2 \log L(X) - 2 \log L(X, u)] + \{-[2 \log L(X, u) - 2 \log L(X, Z)]\} \end{aligned}$$

We call the first term the log-likelihood of between-group lack-of-fit and the second term the log-likelihood of within-group lack-of-fit. If the assumed model is correct, then lack of these terms have approximately a Chi-square distribution with degrees of freedom  $df = \text{rank}(X, u) - \text{rank}(X)$  and  $df = \text{rank}(X, u) - \text{rank}(X, Z)$ , respectively. If the overall lack-of-fit test is significant, we can carry out the between-group and within-group lack-of-fit tests to determine whether the lack-of-fit is due to

between-group or within-group. For example, if the former is significant but the latter is not, we may conclude that the lack-of-fit may be due to difference of levels between groups and model (3.4) may be sufficient for fitting the data. However, if the latter is significant but the former is not, then it is an indication that model (3.4) is not sufficient to explain the data and we need a separate model for each group—that is, a partition model.

The proposed method is applicable to the type one covariate pattern ( $J=n$ ) and type two covariate pattern ( $J<n$ ), and we will use several illustrative examples to show this point. Next, the known test and the proposed tests are summarized in the following table.

**Table 3-1 Summary Table for Known and Proposed Tests**

| Name of the test | Test Statistic   | Distribution | df                  |
|------------------|--|--------------|---------------------|
| Pearson          | $\chi^2 = \sum_{j=1}^J r(y_{j1}, \hat{\pi}(x_j))^2$  | Chi-Square   | n-p-1               |
| Deviance D       | $D = \sum_{j=1}^J d(y_{j1}, \hat{\pi}(x_j))^2$   | Chi-Square   | n-p-1               |
| H-L $\hat{C}$    | $\hat{C} = \sum_{k=1}^g \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$ | Chi-Square   | g-2                 |
| H-L $\hat{H}$    | $\hat{H} = \sum_{k=1}^g \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$ | Chi-Square   | g-2                 |
| Osius& Rojek     | $z = \frac{[\chi^2 - (J - p - 1)]}{\sqrt{A + RSS}}$  | Normal       |                     |
| Stukel's Score   | $ST = -2l(X) - (-2l(X, Z))$  | Chi-Square   | 2                   |
| New_Overall      | $-2l(X) - (-2l(X, Z))$   | Chi-Square   | rank(XZ)-rank(X)    |
| New_Between      | $-2l(X) - (-2l(X, u))$   | Chi-Square   | rank(X,u)-rank(X)   |
| New_Within       | $-2l(X, u) - (-2l(X, Z))$  | Chi-Square   | rank(X,Z)-rank(X,u) |

### 3.2 The First Example

The beetle data from Bliss (1935) gives the number of beetles killed after 5 hours exposure to gaseous carbon disulphide at 8 different concentrations. In this toxicological experiment, the concentrations were log-transformed. The data set is given in Appendix D.

In this data set, there are 481 subjects, and 8 unique covariate patterns, that is,  $J \ll n$ . The problem of interest is to fit a model treating the log dosage as an explanatory variable. The assumed model is of the form given below:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_i \quad i = 1, 2, \dots, 481 \quad (3.5)$$

**Table 3-2 Results from the Known and Proposed Methods for Beetle Data**

| Method      | df | Statistic | p-value | AIC                     |
|-------------|----|-----------|---------|-------------------------|
| Deviance    | 6  | 11.2322   | 0.0815  | 376.471 (Assumed Model) |
| Pearson     | 6  | 10.0253   | 0.1236  | 376.471                 |
| $\hat{C}$   | 6  | 10.0253   | 0.1236  | 376.471                 |
| $\hat{H}$   | 4  | 8.3506    | 0.0795  | 376.471                 |
| O.S.        |    | 0.6918    | 0.4891  | 376.471                 |
| ST.         | 2  | 8.1863    | 0.0167  | 372.284 (Full mode)     |
| New_overall | 2  | 8.399     | 0.0150  | 372.072 (Full model)    |
| New_within  | 1  | 3.723     | 0.0537  |                         |
| New_between | 1  | 4.676     | 0.0306  |                         |

The two-group partition (by the median of the estimated probabilities) model was used. All tests (Table 3-2) except Stukel's score test and the proposed method can not reject the null hypothesis that the assumed model fit and explain the data adequately. Pearson chi-square test and Hosmer and Lemeshow's  $\hat{C}$  test have the same results. P-

values from Deviance test and Pearson Chi-square test are close. P-value of Stukel's test is less than 0.05, which indicates a better model is required for this data. The proposed overall goodness-of-fit test showed that the assumed model is not adequate for the beetle data. Between group and within group test results showed that the lack of fit was from between group and within group.

The 95% prediction intervals on the number of killed beetles were used to show the proposed partition logistic regression model does a better job in predicting the number of killed beetles than the assumed model does.

**Table 3-3 Prediction interval of Number of Beetles Killed by CS<sub>2</sub>**

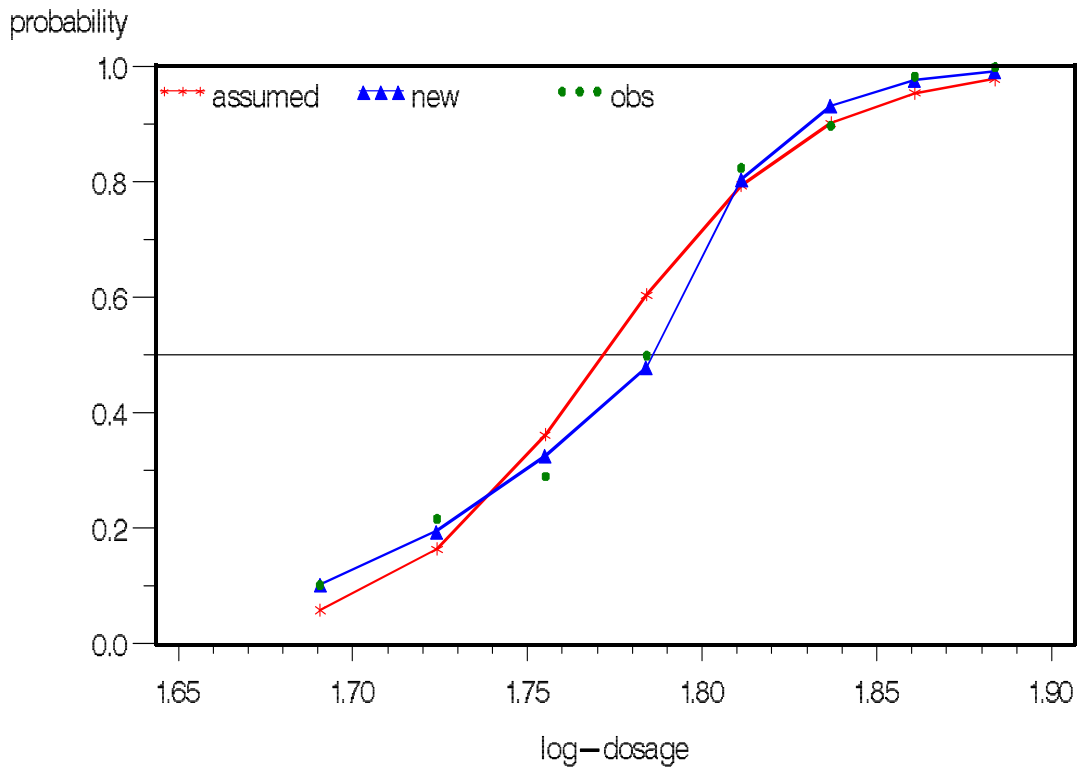
| # of Killed | 95% Confidence Interval for Number of Beetles Killed $n_j \hat{\pi}_j$ |       |                |       |
|-------------|--|-------|----------------|-------|
|             | Assumed Model  |       | Proposed Model |       |
| 6           | 2.02   | 5.81  | 3.40           | 10.57 |
| 13          | 6.96   | 13.61 | 8.52           | 15.81 |
| 18          | 18.50  | 26.73 | 16.32          | 24.61 |
| 28          | 30.36  | 37.25 | 20.82          | 33.10 |
| 52          | 46.53  | 53.05 | 43.82          | 55.65 |
| 53          | 50.72  | 55.12 | 52.18          | 56.77 |
| 61          | 57.40  | 60.34 | 58.07          | 61.52 |
| 60          | 57.63  | 59.34 | 57.71          | 59.91 |

For the assumed model, five out of eight covariate patterns have the observed number of dead beetles falling outside the 95% confidence intervals. However, for the proposed method, only the last covariate pattern lies outside of 95% C.I. (Table3-3).

The estimated probabilities from the assumed model and proposed model are showed in Figure3-1. The line connected blue triangles represents the proposed partition logistic regression model. The line connected the red asterisks represents the assumed model. The observed data are represented by green dots. Figure3-1 indicates that the upper part of both models fit the data very well. The estimated probabilities of

the upper part of the assumed model and the full model of the proposed method are very similar. Figure3-1 also indicates that the proposed model seems to fit the data better than the assumed model when  $\log(\text{dosage}) < 1.8$ .

**Figure 3-1 Plot of Estimated probabilities from the Assumed and Proposed Models for the Beetle Data**



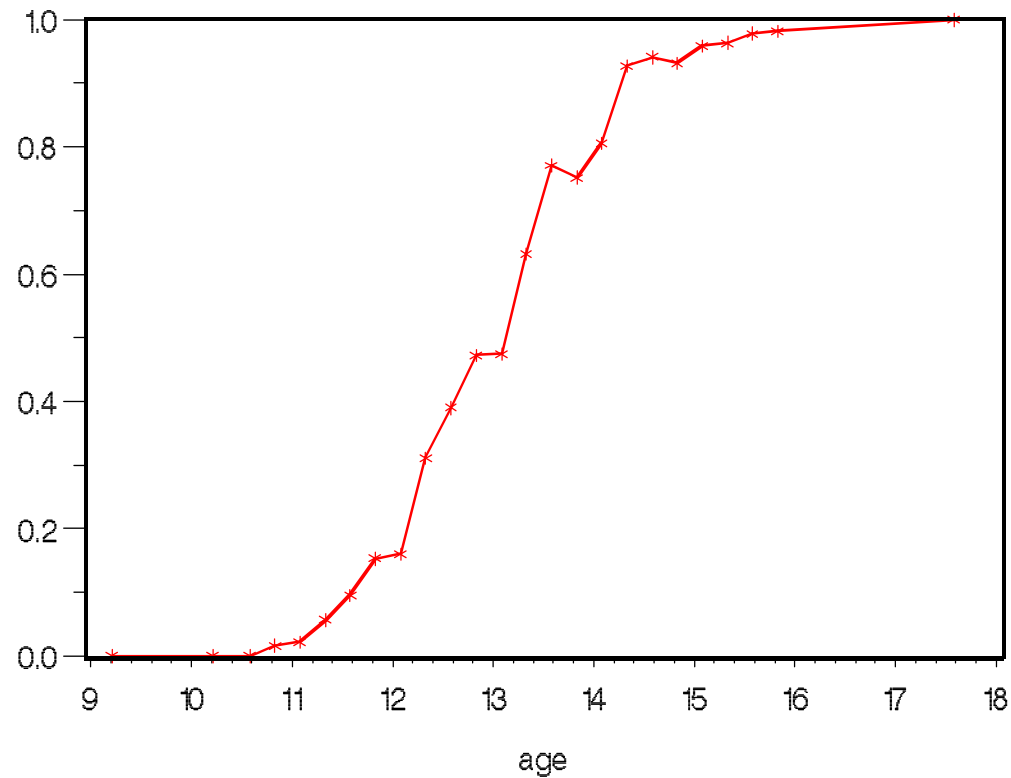
Prentice (1976) showed there is an asymmetric curve departure from the fitted model and suggested that a complementary log-log link function should be used in fitting this data. Stukel (1988) used one additional parameter to show that an asymmetric probability curve is more plausible.



### 3.3 The Second Example

This data is from Milicer and Szczotka (1966) on the age of menarche in 3918 Warsaw girls. The data are given in Appendix E and the observed probability versus age in Figure 3-4. The Warsaw girls are from 9.21 to 17.58. If she reached menarche, denote 1, otherwise, denote 0.

**Figure 3-2 the Observed Probability of Menstruating on Warsaw Girls**



The problem of interest is to fit a model treating the age as the predictor variable and the status of menstruating (binary) as the response variable. The assumed model is the following form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_i \quad i = 1, 2, \dots, 3918 \quad (3.6)$$

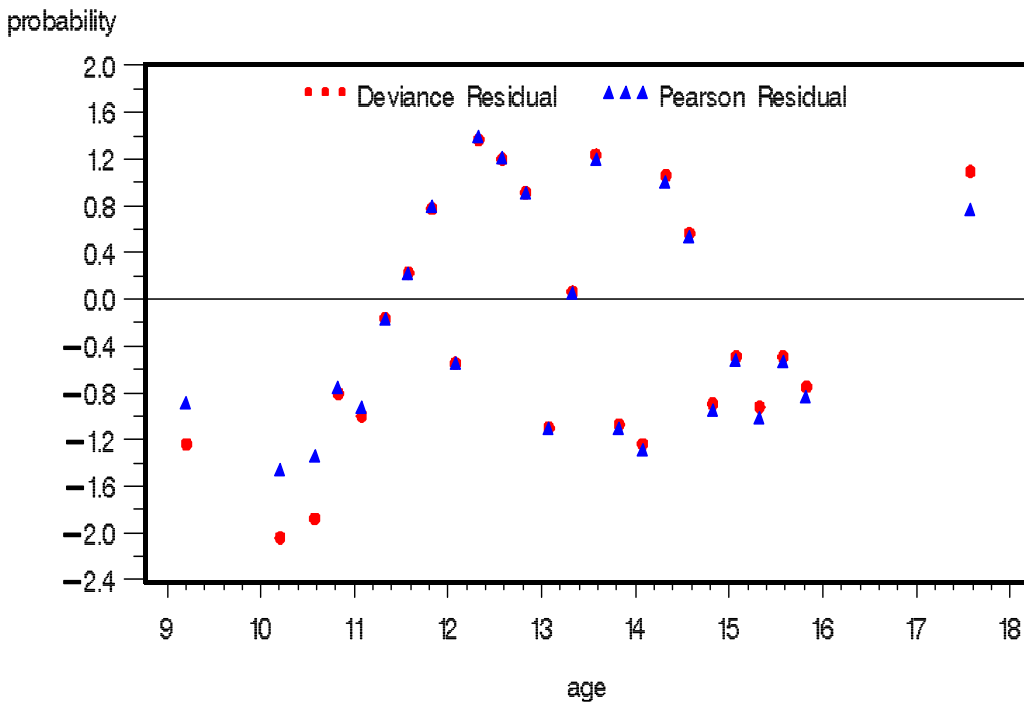
The results from the known methods are displayed in Table 3-4.

**Table 3-4 Results from the Known for Warsaw Girls Data**

| Method    | df | Statistic | p-value | AIC      |
|-----------|----|-----------|---------|----------|
| Deviance  | 23 | 26.7035   | 0.2688  | 1643.305 |
| Pearson   | 23 | 21.8684   | 0.5282  |          |
| $\hat{C}$ | 6  | 9.7756    | 0.1344  |          |
| $\hat{H}$ | 8  | 8.4069    | 0.3948  |          |
| O.S.      |    | 0.0067    | 0.9884  |          |
| ST.       | 2  | 11.282    | 0.0035  | 1636.012 |

From the Table3-4, the p-value of Stukel's score test is less than 0.05 for the known tests, which indicates a better model is needed for this data set. Osius and Rojek's approximate normality method provides a very big p-value. Based on AIC value, it is possible to improve the fit by Stukel's model.

**Figure 3-3 the Plot of Residual versus Age**



The residual plot (Figure3-3) also shows that an improvement of fit over the assumed model is possible.

For the proposed test, the following five different grouping schemes used to form the partition model are considered. For grouping scheme I, the data points are partitioned into two groups with boundary at 25<sup>th</sup> percentile of the estimated probabilities from fitting the assumed model. Grouping scheme II and III are constructed the similar way with boundary at 35<sup>th</sup> percentile and 50<sup>th</sup> percentile, respectively. For grouping scheme IV, the data points are partitioned into three groups by the group boundaries 35<sup>th</sup> percentile and 65<sup>th</sup> percentile of the estimated probabilities from fitting the assumed model. Grouping scheme V is constructed the same way with boundaries at 25<sup>th</sup> percentile and 75<sup>th</sup> percentile.

In Table3-5, the p-values of all five overall proposed tests are less than 0.05, which indicates that lack-of-fit is present in the assumed model. The lack of fit may come from within group or between group or both with different grouping schemes. The proposed overall test is not sensitive to the grouping schemes.

**Table 3-5 Results from Different Grouping Schemes of the Proposed Method**

| grouping    | df | Statistic | p-value  | AIC      |
|-------------|----|-----------|----------|----------|
| I_overall   | 2  | 11.401    | 0.0033   | 1635.940 |
| I_bewteen   | 1  | 8.697     | 0.0032   |          |
| I_within    | 1  | 2.722     | 0.9897   |          |
| II_overall  | 2  | 10.807    | 0.004501 | 1636.498 |
| II_bewteen  | 1  | 1.134     | 0.2869   |          |
| II_within   | 1  | 9.673     | 0.00197  |          |
| III_overall | 2  | 8.5943    | 0.0136   | 1638.710 |
| III_bewteen | 1  | 3.827     | 0.05043  |          |
| III_within  | 1  | 4.768     | 0.02899  |          |
| IV_overall  | 4  | 11.795    | 0.0189   | 1639.510 |
| IV_bewteen  | 2  | 1.202     | 0.5483   |          |
| IV_within   | 2  | 10.597    | 0.005    |          |
| V_overall   | 4  | 12.949    | 0.0115   | 1638.356 |
| V_bewteen   | 2  | 9.726     | 0.0077   |          |
| V_within    | 2  | 3.223     | 0.1996   |          |

The model with grouping scheme I has the smallest AIC value among the assumed model, Stukel's model and the proposed partition models with different grouping schemes. The estimated probabilities from the assumed model and proposed model with grouping scheme I are in the Figure3-4. The line connected orange triangles represents the assumed logistic regression model. The line connected the blue circles represents the proposed model. The observed data are represented by asterisks. Figure3-4 shows that the upper parts of both models fit the data very well. The estimated probabilities upper part of the assumed model and full model of the proposed method are very close. The Figure3-4 also shows a slight improvement of model with grouping scheme I in fitting the lower part of the proposed model. From above, two-

group partition model with grouping scheme I is more suitable for this data and the model obtained from Table3-6 is

$$\text{logit}(\pi(x)) = \begin{cases} -20.0488 + 1.5447 * x & x \in G_1 \\ -54.7125 + 4.6185 * x & x \notin G_1 \end{cases} \quad (3.6)$$

**Table 3-6 the Parameter Estimates for Model with Grouping Scheme I**

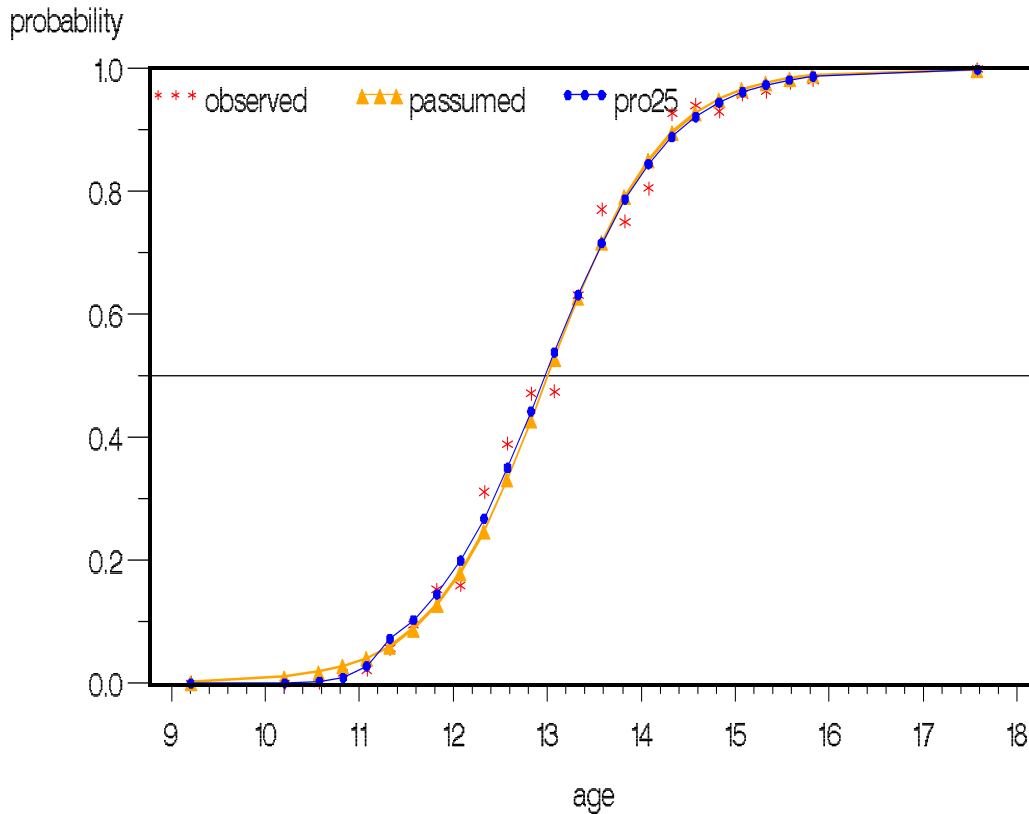
| Analysis of Maximum Likelihood Estimates |    |          |                |                 |            |
|--|----|----------|----------------|-----------------|------------|
| Parameter                                | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| grp1                                     | 1  | -20.0488 | 0.8424         | 566.4574        | <.0001     |
| g1x1                                     | 1  | 1.5447   | 0.0639         | 583.9186        | <.0001     |
| grp2                                     | 1  | -54.7125 | 27.6778        | 3.9076          | 0.0481     |
| g2x1                                     | 1  | 4.6185   | 2.5265         | 3.3416          | 0.0675     |

**Table 3-7 Model Fit Statistics of Model 3.6**

| Model Fit Statistics |          |
|----------------------|----------|
| AIC                  | 1635.940 |
| SC                   | 1661.033 |
| -2 Log L             | 1627.940 |

Stukel (1988) also showed the poor fit of the assumed model due to the deviation from the logistic curve on the lower tail. A probit link function provides a mild improvement on these data by Finney (1971) and Aranda-Ordaz (1981). The model with probit link (AIC 1639.489) is not better than partition models (Table 3-5) and Stukel's general logistic regression model (Table 3-4).

**Figure 3-4 Probability versus Age on Warsaw Girl Data**



From Table3-8, the 95% prediction interval for the observed counts indicates that an improvement over the assumed model made by the proposed partition model is possible, particularly on the lower tail. When the 95% critical limits do not include the observed count, the proposed partition model gives a much closer bound than the assumed model does.

**Table 3-8 Observed Counts and Confidence Interval (Assumed Model and Proposed Method) for the Warsaw Girl Data**

| # of menstruating | 95% Confidence Interval for Number of menstruating $n_j \hat{\pi}_j$ |         |                             |         |
|-------------------|--|---------|-----------------------------|---------|
|                   | Assumed Model  |         | Proposed Model (Grouping I) |         |
| 0                 | 0.48   | 1.21    | 0                           | 11.15   |
| 0                 | 1.46   | 2.91    | 0.002                       | 4.6     |
| 0                 | 1.28   | 2.35    | 0.03                        | 2.13    |
| 2                 | 2.54   | 4.40    | 0.34                        | 3.4     |
| 2                 | 2.90   | 4.75    | 0.8                         | 7.7     |
| 5                 | 4.30   | 6.65    | 5.02                        | 8.10    |
| 10                | 7.70   | 11.25   | 8.79                        | 13.31   |
| 17                | 12.05  | 16.65   | 13.46                       | 19.13   |
| 16                | 15.74  | 20.64   | 17.22                       | 23.02   |
| 29                | 20.68  | 25.81   | 22.15                       | 27.80   |
| 39                | 30.36  | 36.31   | 31.89                       | 38.37   |
| 51                | 42.98  | 49.62   | 44.39                       | 51.30   |
| 47                | 49.40  | 55.50   | 50.29                       | 56.39   |
| 67                | 63.44  | 69.79   | 63.87                       | 70.08   |
| 81                | 72.37  | 78.28   | 72.29                       | 78.06   |
| 88                | 89.69  | 95.62   | 89.90                       | 94.56   |
| 79                | 81.24  | 85.52   | 80.15                       | 95.05   |
| 90                | 85.08  | 88.59   | 84.31                       | 88.01   |
| 113               | 109.55   | 113.03  | 108.60                      | 112.39  |
| 95                | 95.78  | 98.07   | 95.04                       | 97.62   |
| 117               | 116.82   | 118.91  | 116.06                      | 118.48  |
| 107               | 107.74   | 109.16  | 107.17                      | 108.87  |
| 92                | 92.10  | 92.99   | 91.71                       | 92.79   |
| 112               | 112.41   | 113.20  | 112.04                      | 113.03  |
| 1049              | 1047.97  | 1048.65 | 1047.48                     | 1048.51 |

### 3.4 The Third Example

The data of this example are from two general social surveys carried out by the National Opinion Research Center, University of Chicago, Illinois, in 1974 and 1975. Part of these two surveys was concerned with the relationship of education and gender with the attitudes towards the role of women in society. Each respondent was asked if he or she agreed or disagreed with ‘Women should take care of running their homes and leave running the country up to men.’ The data were combined from two sets, since no individual was involved two surveys and short time interval between two surveys. Among 2927 participators, three did not give their education levels and fifty three were not sure whether they agreed or not the statement provided. The responses from 1305 males and 1566 females were given by Haberman (1978) and showed in Appendix E.

The problem of interest is to fit a model treating the years of education and gender as the predictor variables and the attitude of the statement (binary) as the response variable, where “1” is used to denote the agreement on the statement in the survey. The assumed model is of the following form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \quad i = 1, 2, \dots, 2871 \quad (3.6)$$

**Table 3-9 Results from the Known and Proposed Methods for Women’s Role Data**

| Method      | df | Statistic | p-value | AIC      |
|-------------|----|-----------|---------|----------|
| Deviance    | 38 | 64.0066   | 0.0052  | 3354.330 |
| Pearson     | 38 | 72.6678   | 0.0006  | 3354.330 |
| $\hat{C}$   | 6  | 2.9756    | 0.8142  | 3354.330 |
| $\hat{H}$   | 8  | 12.444    | 0.01345 | 3354.330 |
| O.S.        |    | 2.4770    | 0.0067  | 3354.330 |
| ST.         | 2  | 10.1506   | 0.0062  | 3348.179 |
| New_overall | 6  | 15.5117   | 0.0166  | 3348.818 |
| New_within  | 3  | 5.3832    | 0.1458  |          |
| New_between | 3  | 10.085    | 0.01786 |          |



In this example, four groups were used in the partition model, the data were partitioned into two groups by gender, and then each group was divided into two sub-groups by the median of estimated probabilities. From Table 3-9, all tests showed that the assumed model is not adequate for the data except Hosmer-Lemeshow's  $\hat{C}$  tests. The proposed test showed the presence of lack-of-fit between group. The proposed partition model and the Stukel's model give similar AIC values.

Collett (1991) showed the assumed model is not adequate to the data and suggested that two different logistic regression models for males and females, respectively, should be used to fit model:

$$\text{Males: } \log\left(\frac{\pi}{1-\pi}\right) = 2.098 - 0.2346 * (\text{yearsedu.})$$

$$\text{Females: } \log\left(\frac{\pi}{1-\pi}\right) = 3.003 - 0.315 * (\text{yearsedu.})$$

### 3.5 The Fourth Example

The data (Appendix G) of this example was from Luigi Bocconi, a cura di R. Piccarreta (1993) The subjects of this example were randomly selected in Italian in order to study the relation between the annual income and whether one possesses a travel credit card. In this data set, the number of cases represents the number of subjects at the income level. If the subject has at least one travel credit card, the credit card was denoted as “1”, otherwise, “0”.

In this date set, there are total 100 subjects, and 31 unique covariate patterns, that is,  $J < n$ . The problem of interest is to fit a model treating the annual income (millions of Lira) as explanatory variable and credit card as a binary response variable. The assumed model is of the form given below:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_i \quad i = 1, 2, \dots, 100 \quad (3.7)$$

**Table 3-10 Results from the Known Tests for Credit Card Data**

| Method    | df | Statistic | p-value | AIC     |
|-----------|----|-----------|---------|---------|
| Deviance  | 29 | 45.9211   | 0.0239  | 100.963 |
| Pearson   | 29 | 41.0674   | 0.0679  |         |
| $\hat{C}$ | 7  | 8.4355    | 0.2958  |         |
| $\hat{H}$ | 8  | 17.9802   | 0.0214  |         |
| O.S.      |    | 1.6364    | 0.1018  |         |
| ST.       | 2  | 0.261     | 0.8776  | 104.702 |

In this example, Stukel’s score test gives the largest p-value among six known tests. P-values of Deviance test and Hosmer-Lemeshow  $\hat{H}$  test are less than 0.05,

which indicate the presence of lack-of-fit in the assumed model. Hosmer-Lemeshow  $\hat{H}$  test and Hosmer-Lemeshow  $\hat{C}$  test give contradicting conclusion, Hosmer-Lemeshow  $\hat{C}$  test indicates the absence of lack-of-fit in the assumed model. The p-value of Pearson chi-square is just a little bit larger than 0.05.

Since different conclusions are obtained from different tests, it is worthy to investigate the nature of lack-of-fit. There are three parts—upper tail (above 75<sup>th</sup> percentile of estimated probabilities from the assumed model), middle part (between 25<sup>th</sup> percentile and 75<sup>th</sup> percentile) and lower tail (below 25<sup>th</sup> percentile of the estimated probabilities from the assumed model) will be considered. The data are partitioned into three groups as below:

$$G_1 = \{(x_i, y_i) : \hat{\pi}(x_i) < 25^{\text{th}} \text{ percentile}\}$$

$$G_2 = \{(x_i, y_i) : 25^{\text{th}} \text{ percentile} \leq \hat{\pi}(x_i) \leq 75^{\text{th}} \text{ percentile}\}$$

$$G_3 = \{(x_i, y_i) : \hat{\pi}(x_i) > 75^{\text{th}} \text{ percentile}\}$$

First, we will detect whether the lack-of-fit occurs at the lower tail, middle part or upper tail of  $\hat{\pi}(x)$  by fitting respectively the following two-group partition models:

$$\text{logit}(\pi(x)) = \begin{cases} x' \beta + z' \alpha_i & x \in G_i \\ x' \beta & x \notin G_i \end{cases} \quad (i=1,2,3) \quad (3.8)$$

**Table 3-11 Goodness-of-fit Test Based on Model (3.8)**

| Model                                 | df | Test statistic | p-value |
|---------------------------------------|----|----------------|---------|
| Model 1 of 3.8 (without second order) | 2  | 0.998          | 0.6071  |
| Model 1 of 3.8 (with second order)    | 3  | 2.508          | 0.4738  |
| Model 2 of 3.8 (without second order) | 2  | 6.152          | 0.04614 |
| Model 2 of 3.8 (with second order)    | 3  | 13.376         | 0.00389 |
| Model 3 of 3.8 (without second order) | 2  | 4.838          | 0.089   |

The parameter estimates of model 3.8 under different cases considered is shown in the Table 3-12.

**Table 3-12 Parameter Estimates of Model (3.8)**

|                     | Parameter   | Estimate | Std.err | p-value |
|---------------------|-------------|----------|---------|---------|
| G1<br>(lower tail)  | Intercept   | -3.6308  | 0.9793  | 0.0002  |
|                     | X           | 0.0544   | 0.0164  | 0.0009  |
|                     | grp1        | 4.5688   | 7.8076  | 0.5584  |
|                     | X_low       | -0.2194  | 0.2648  | 0.4073  |
| G1<br>(lower tail)  | Intercept   | -3.6308  | 0.9793  | 0.0002  |
|                     | X           | 0.0544   | 0.0164  | 0.0009  |
|                     | grp1        | -137.6   | 180.9   | 0.4468  |
|                     | X_low       | 9.6927   | 12.3771 | 0.4336  |
|                     | X_low*x_low | -0.1721  | 0.2115  | 0.4158  |
| G2<br>(middle part) | Intercept   | -3.3142  | -0.8143 | <.0001  |
|                     | X           | 0.0423   | 0.0122  | 0.0005  |
|                     | grp2        | -6.6457  | 1.8299  | 0.0003  |
|                     | X_middle    | 0.1178   | 0.0346  | 0.0007  |
| G2<br>(middle part) | Intercept   | -3.3142  | -0.8143 | <.0001  |
|                     | X           | 0.0423   | 0.0122  | 0.0005  |
|                     | grp2        | 17.3259  | 9.5534  | 0.0697  |
|                     | X_mid       | -0.8878  | 0.4096  | 0.0302  |
|                     | X_mid*X_mid | 0.0101   | 0.0042  | 0.0165  |
| G3<br>(upper tail)  | Intercept   | -4.9787  | 1.0884  | <.0001  |
|                     | X           | 0.0877   | 0.0226  | 0.0001  |
|                     | grp3        | -5.0788  | 2.7945  | 0.0691  |
|                     | X_upper     | 0.0612   | 0.0316  | 0.0529  |

From Table 3-11, for model 2 of 3.8 with first order, the test statistic of the proposed method is 6.512 with 2 degrees of freedom and gives the p-value=0.04614, which indicates that a better model is needed for this data set. For model 2 of 3.8 with first and second order, the test statistic of the proposed method is 13.376 with 3 degrees of freedom and gives the p-value=0.00389, which indicates that the assumed model is not adequate for this data set.

**Table 3-13 Model Fit Statistics from Two Cases of Model (3.8) i=2**

| Criteria | With first order | With first and second order |
|----------|------------------|-----------------------------|
| AIC      | 98.811           | 93.587                      |
| SC       | 109.232          | 106.883                     |
| -2 Log L | 90.811           | 83.587                      |

From Table 3-13, we may conclude model 2 of 3.8 with first and second order is more suitable than the model without second order, since the model with first and second order has the smaller AIC and SC.

Second, the two three-group partition models will be used to detect further on the lack-of-fit and model building. The three-group partition models are expressed as:

$$\logit(\pi(x)) = \begin{cases} x' \beta + z' \alpha_2 & x \in G_2 \\ x' \beta + z' \alpha_1 & x \in G_1 \\ x' \beta & x \in G_3 \end{cases} \quad (3.9)$$

$$\text{and } \logit(\pi(x)) = \begin{cases} x' \beta + z' \alpha_2 & x \in G_2 \\ x' \beta + z' \alpha_3 & x \in G_3 \\ x' \beta & x \in G_1 \end{cases} \quad (3.10)$$

We are interested in testing  $H_0: \alpha_2 = 0$  versus  $H_a: \alpha_2 \neq 0$ . The test statistic was calculated from model 3.9 and model 1 of 3.8 with second order is 13.499 along with  $df=3$ , which yield the p-value=0.0367. We can conclude that the lack-of-fit occurs in the middle part. Similarly, for the testing on  $H_0: \alpha_1 = 0$  versus  $H_a: \alpha_1 \neq 0$ , the test statistic was calculated from model 3.9 and model 2 of 3.8 with second order is 2.631 with  $df=3$ , which yield the p-value=0.4521. Thus, we can conclude that the lack-of-fit occurs in the middle part.

**Table 3-14 Parameter Estimates of Model (3.9)**

| Parameter   | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-------------|----------|----------------|-----------------|------------|
| grp1        | -137.6   | 180.9          | 0.5787          | 0.4468     |
| X_low       | 9.6927   | 12.3771        | 0.6133          | 0.4336     |
| X_low*x_low | -0.1721  | 0.2115         | 0.6620          | 0.4158     |
| grp2        | 17.3262  | 9.5535         | 3.2892          | 0.0697     |
| X_mid       | -0.8878  | 0.4096         | 4.6986          | 0.0302     |
| X_mid*X_mid | 0.0101   | 0.00420        | 5.7478          | 0.0165     |
| grp3        | -5.0788  | 2.7945         | 3.3031          | 0.0691     |
| X_upper     | 0.0612   | 0.0316         | 3.7470          | 0.0529     |

For the model 3.10, here is a note:

WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Thus, we can declare that the model 3.10 is not applicable for this data set.

From first and second step, we may claim that the final model is model 2 of (3.8) with first and second order:

$$\text{logit}(\pi(x)) = \begin{cases} 17.3259 - 0.8878 * x + 0.0101x^2 & x \in G_2 \\ -3.3142 - 0.0423 * x & x \notin G_2 \end{cases}$$

## CHAPTER 4 - Simulation Study

In this chapter, four different simulation studies were conducted to evaluate the adequacy of the proposed null distributions in approximating the sampling distribution of our test statistics when the assumed model is the correct model and to compare the power of the proposed test with six known tests to detect a variety of departures from the true models. The six known goodness-of-fit test statistics used to compared with the proposed test statistic in the simulation study are Pearson Chi-square statistic,  $\chi^2$ ; Deviance test statistic, D; Hosmer and Lemeshow's decile of risk statistic, that is the test statistic  $\hat{C}$  with 10 groups; Hosmer and Lemeshow's predetermined cutoff point statistic,  $\hat{H}$ , with up to 10 groups; the Osius and Rojek normal approximation test statistic,  $Z_{OR}$ ; and Stukel's score test statistic,  $ST$ . All simulations were performed on a Dell Inspiron 6000 PC using SAS Version 9.1.

In this Chapter, all simulation studies considered here will include type one covariate pattern (J=n) to examine the performance of the proposed test, since the standard goodness-of-fit tests are unsatisfactory for this covariate pattern. Further more, data with type one covariate pattern are frequently seen in application. The covariates in the first, third, and the fourth simulation study are continuous, and the covariates in the second simulation are a mixture of continuous and categorical covariates.



## 4.1 The First Simulation Study

In this section, we will use a simple logistic regression model as our assumed model in which only one predictor variable is used. We will first check whether the proposed model controls the type I error rate, when the data are generated from the following logistic model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 \quad (4.1)$$

where  $\beta_0 = -0.3$  and  $\beta_1 = 1.3$ ,

In this simulation study, 1000 random samples of size  $n=100$ , 200 or 500 were respectively generated from the above model with covariate  $x$  independently generated from a uniform distribution over the interval  $(-2,2)$ . The response variable  $y$  associated with the covariate  $x$  was generated as follow: A random variable  $u$  is generated from a uniform distribution over the interval  $(0,1)$ . The response  $y$  associated with  $x$  was assigned a value 1, if  $\pi(x) \geq u$ , and 0 otherwise.

The proposed 2-group partition test with cut-off point set at 50<sup>th</sup> percentile, 3-groups partition test with cut-off points set at 25<sup>th</sup> percentile and 75<sup>th</sup> percentile, and the six known tests mentioned above were applied to the random samples generated from model (4.1) at significance level set at  $\alpha = 0.05, 0.1$ . The observed rejection rates are presented in Table 4-1, where  $\hat{C}$  is the Hosmer and Lemeshow's  $\hat{C}$  test;  $\hat{H}$  is the Hosmer and Lemeshow's  $\hat{H}$  test; D is the Deviance test; P is the Pearson Chi-square test; O.S. is the Osius and Rojek's normal approximately normal test; ST is the Stukel's

score test; New2 is the proposed method with 2 groups; New3 is the proposed method with 3 groups.

**Table 4-1 Observed Type I Error Rate for Simulation Study one**

| Sample Size | Method    | $\alpha=0.05$ | $\alpha=0.10$ |
|-------------|-----------|---------------|---------------|
| 100         | $\hat{C}$ | 0.037         | 0.082         |
|             | $\hat{H}$ | 0.043         | 0.079         |
|             | D         | 0.05          | 0.031         |
|             | P         | 0.014         | 0.023         |
|             | O.S.      | 0.06          | 0.098         |
|             | ST.       | 0.054         | 0.102         |
|             | New2      | 0.063         | 0.104         |
|             | New3      | 0.06          | 0.101         |
| 200         | $\hat{C}$ | 0.036         | 0.087         |
|             | $\hat{H}$ | 0.052         | 0.092         |
|             | D         | 0.006         | 0.029         |
|             | P         | 0.008         | 0.016         |
|             | O.S.      | 0.041         | 0.084         |
|             | ST.       | 0.047         | 0.102         |
|             | New2      | 0.052         | 0.104         |
|             | New3      | 0.06          | 0.102         |
| 500         | $\hat{C}$ | 0.036         | 0.086         |
|             | $\hat{H}$ | 0.052         | 0.085         |
|             | D         | 0.006         | 0.041         |
|             | P         | 0.008         | 0.006         |
|             | O.S.      | 0.041         | 0.078         |
|             | ST.       | 0.047         | 0.106         |
|             | New2      | 0.052         | 0.101         |
|             | New3      | 0.055         | 0.105         |

Given the specified significance level  $\alpha$  and sample size  $n$ , the 95% likely interval for type I error rate is:

$$\left( \alpha - 1.96 * \sqrt{\frac{\alpha(1-\alpha)}{n}}, \alpha + 1.96 * \sqrt{\frac{\alpha(1-\alpha)}{n}} \right)$$

**Table 4-2 the 95% Confidence Interval for the Type I Error Rate**

| Sample size   | Lower limit | Upper limit |
|---------------|-------------|-------------|
| $\alpha=0.05$ | 0.0365      | 0.0635      |
| $\alpha=0.10$ | 0.0814      | 0.1186      |

Based on the Table 4-2, the type I error rates of Stukel's score test and the proposed tests all fall within 95% confidence interval, under different sample size and different significance levels. Type I error rates of Pearson Chi-square test are less than lower limit of 95% confidence interval. Some type I error rates of Deviance test, Hosmer-Lemeshow's tests and Osius and Rojek's test fall outside 95% confidence interval.

Next, the random samples were generated from the following model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \quad (4.2)$$

where  $\beta_2$  takes on the values 0, 0.2, 0.4, 0.6, and 0.8, respectively. The proposed 2-group and 3-group partition tests and the six known texts were applied to the random samples generated from (4.2) to test the goodness-of-fit of the assumed model (4.1) at level of significance  $\alpha = 0.05$ . The rejection rates of these tests were presented in Table 4-3 and graphically in Figure 4-1 to 4-3.

**Table 4-3 Rejection Rates for the Known and Proposed Tests for Study One**

| size | $\beta_2$ | $\hat{C}$ | $\hat{H}$ | D     | P     | O.S.  | ST.   | New2  | New3  |
|------|-----------|-----------|-----------|-------|-------|-------|-------|-------|-------|
| 100  | 0         | 0.037     | 0.043     | 0.005 | 0.014 | 0.061 | 0.054 | 0.063 | 0.06  |
|      | 0.2       | 0.06      | 0.057     | 0.007 | 0.014 | 0.104 | 0.106 | 0.11  | 0.1   |
|      | 0.4       | 0.119     | 0.143     | 0.012 | 0.012 | 0.13  | 0.292 | 0.272 | 0.239 |
|      | 0.6       | 0.271     | 0.328     | 0.057 | 0.003 | 0.097 | 0.604 | 0.55  | 0.498 |
|      | 0.8       | 0.519     | 0.595     | 0.138 | 0     | 0.089 | 0.876 | 0.829 | 0.778 |
| 200  | 0         | 0.036     | 0.052     | 0.006 | 0.008 | 0.041 | 0.047 | 0.052 | 0.06  |
|      | 0.2       | 0.101     | 0.085     | 0.01  | 0.006 | 0.091 | 0.171 | 0.163 | 0.13  |
|      | 0.4       | 0.273     | 0.301     | 0.027 | 0.004 | 0.114 | 0.556 | 0.517 | 0.448 |
|      | 0.6       | 0.646     | 0.682     | 0.137 | 0     | 0.116 | 0.903 | 0.872 | 0.808 |
|      | 0.8       | 0.896     | 0.919     | 0.383 | 0     | 0.397 | 0.991 | 0.986 | 0.969 |
| 500  | 0         | 0.045     | 0.039     | 0.011 | 0     | 0.035 | 0.042 | 0.047 | 0.055 |
|      | 0.2       | 0.172     | 0.152     | 0.019 | 0.007 | 0.117 | 0.345 | 0.332 | 0.265 |
|      | 0.4       | 0.716     | 0.709     | 0.081 | 0.001 | 0.125 | 0.914 | 0.887 | 0.836 |
|      | 0.6       | 0.984     | 0.985     | 0.453 | 0     | 0.261 | 0.999 | 0.998 | 0.996 |
|      | 0.8       | 1         | 1         | 0.849 | 0     | 0.904 | 1     | 1     | 1     |

**Figure 4-1 Plots of Rejection Rate for Study One (n=100)**

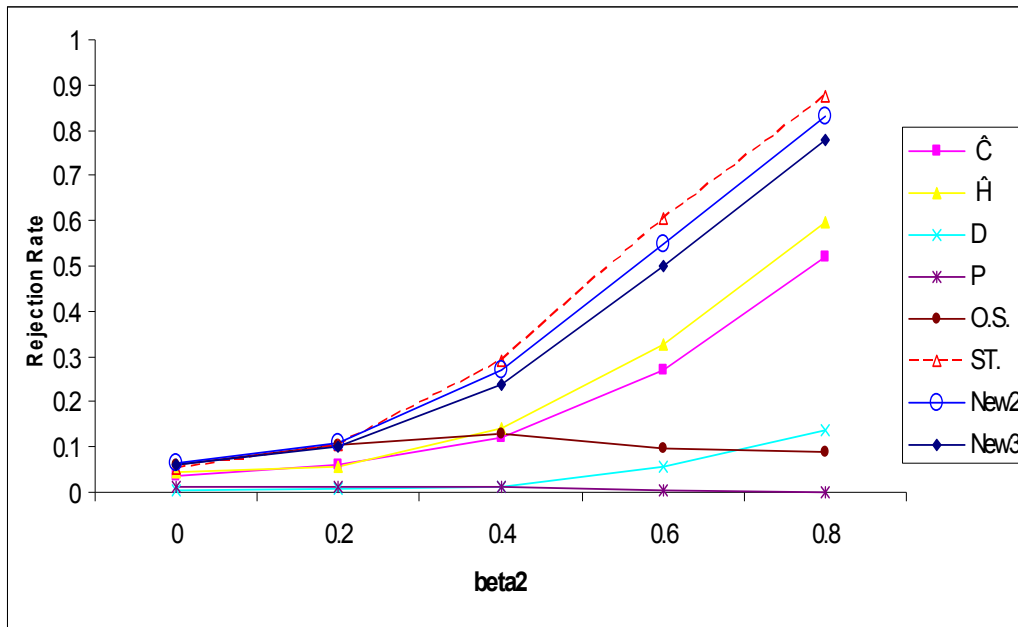


Figure 4-2 Plots of Rejection Rate for Study One (n=200)

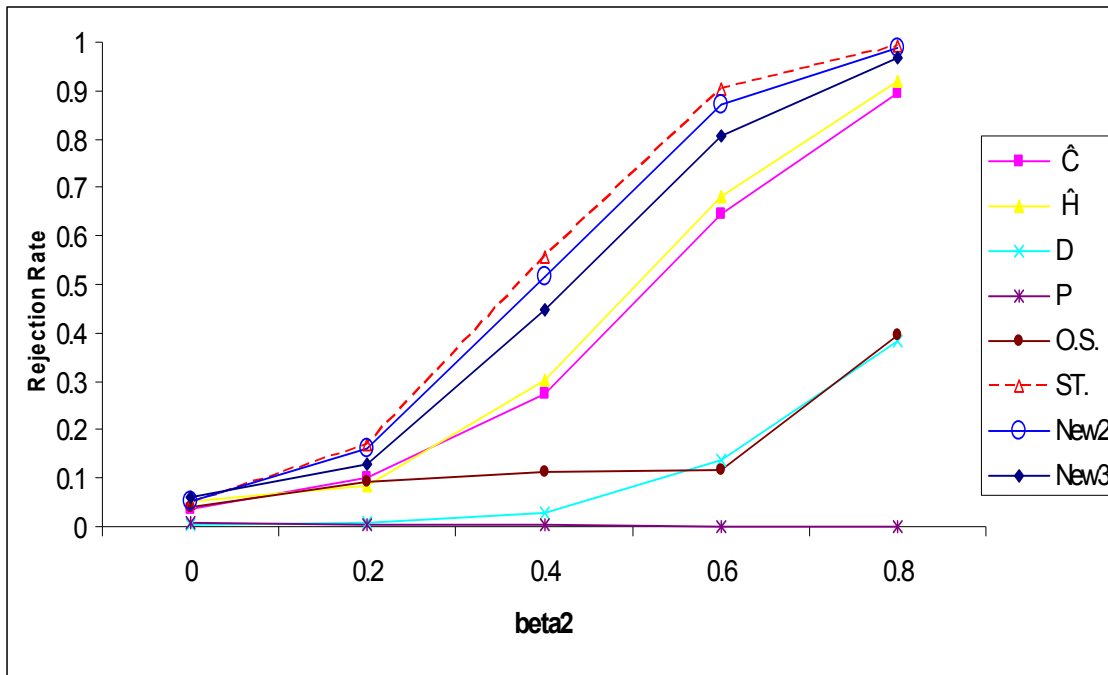
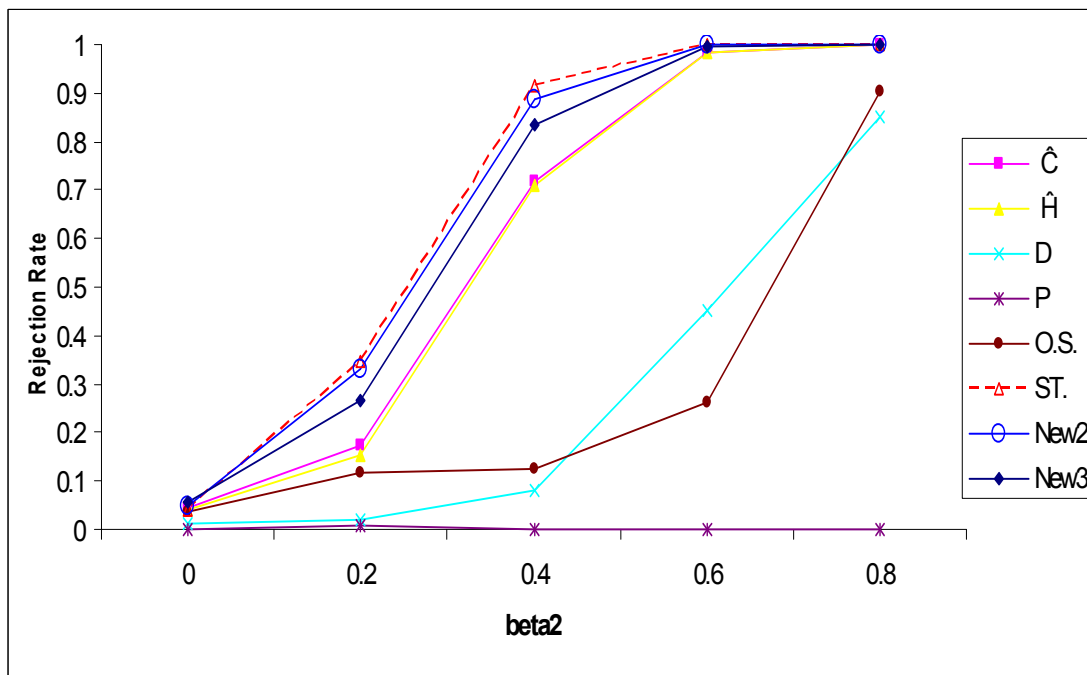


Figure 4-3 Plots of Rejection Rate for Study One (n=500)



The simulation results suggest that, the rejection rates of the Pearson Chi-square test and Osius and Rojek normal approximate test decrease as the assumed model departs further away from the true linear logistic regression model when  $n=100$ . When sample size are 200 and 500, the rejection rates of all tests except the Pearson chi-square test increase as the departure of the assumed model from the true model increases (Figure 4-2, 4-3). The Hosmer and Lemeshow  $\hat{C}$  and  $\hat{H}$  tests have similar power functions for different level of sample sizes. The proposed (two-group and three-group) partition tests and Stukel's score test have nearly the same rejection rate with the latter having slightly higher power when the assumed model departs from the true model. With the exception of Pearson square test, the performances of all tests improve when sample size increases. For example, when  $\beta_2 = 0.4$ , the proposed test, Stukel's score test, Hosmer-Lemeshow  $\hat{C}$  test, Hosmer-Lemeshow  $\hat{H}$  test, Deviance test have 61.5, 62.2, 59.7, 56.5, 6.9 per cent more chance of rejecting the assumed model when  $n=500$  than that when  $n=100$ . The simulation results investigated by Hosmer, Hosmer, le Cessie and Lemeshow (1997) indicated that all of the overall goodness-of-fit tests are not powerful for small to moderate sample sizes ( $n < 400$ ). Table 4-3 and Figure 4-1, 4-2 and 4-3 showed that Stukel's score test is more powerful than the other tests, followed by the new proposed method, and the Hosmer and Lemeshow  $\hat{C}$  and  $\hat{H}$  tests. Pearson Chi-square test has very poor performance, the rejection rate remains at very low level regardless of how big the sample size is and how far the assumed model is away from the true model.

## 4.2 The Second Simulation Study

In this section, we will use a mixture logistic regression model as our assumed model in which one predictor is a continuous variable and the other is a categorical variable. We will first check whether the proposed method controls the type I error rate, when the data are generated from the following logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 d \quad (4.3)$$

where  $\beta_0 = -0.25$ ,  $\beta_1 = 0.6$  and  $\beta_2 = 0.3$ ,  $x_1$  and  $d$  are defined as below

In this simulation study, 1000 random samples of size  $n=100,200,500$  were respectively generated from model (4.3) with covariate  $x_1$  independently generated from an uniform distribution over the interval  $(-3,3)$ , and the covariate  $d$  independently generated from a Bernoulli distribution with probability 0.5. The response variable  $y$  associated with the covariate  $x_1$  and  $d$  was generated as follow. A random variable  $u$  is generated from a uniform distribution over the interval  $(0,1)$ . The response  $y$  was assigned a value 1, if  $\pi(x) \geq u$ , and 0 otherwise.

The proposed 4-group partition test partitions each of the two categorical group defined by  $d$  into two groups at cut off points set at the 50<sup>th</sup> percentiles. The six known tests mentioned before were applied to the random samples generated from model (4.3) at the level of significance  $\alpha = 0.05, 0.1$ . The observed rejection rates are presented in Table4-4, where  $\hat{C}$  is the Hosmer and Lemeshow's  $\hat{C}$  test;  $\hat{H}$  is the Hosmer and Lemeshow's  $\hat{H}$  test;  $D$  is the Deviance test;  $P$  is the Pearson Chi-square test; O.S. is

the Osius and Rojek's normal approximately normal test; ST. is the Stukel's score test, New is the proposed test with 4 groups.

**Table 4-4 Observed Type I Error Rate for Simulation Study Two**

| Sample Size | Methods   | $\alpha=0.05$ | $\alpha=0.10$ |
|-------------|-----------|---------------|---------------|
| 100         | $\hat{C}$ | 0.04          | 0.089         |
|             | $\hat{H}$ | 0.042         | 0.09          |
|             | D         | 0.221         | 0.454         |
|             | P         | 0.003         | 0.006         |
|             | O.S.      | 0.154         | 0.212         |
|             | ST.       | 0.064         | 0.138         |
|             | New       | 0.065         | 0.128         |
| 200         | $\hat{C}$ | 0.047         | 0.106         |
|             | $\hat{H}$ | 0.063         | 0.107         |
|             | D         | 0.492         | 0.731         |
|             | P         | 0             | 0             |
|             | O.S.      | 0.093         | 0.134         |
|             | ST.       | 0.064         | 0.109         |
|             | New       | 0.064         | 0.126         |
| 500         | $\hat{C}$ | 0.059         | 0.119         |
|             | $\hat{H}$ | 0.062         | 0.119         |
|             | D         | 0.929         | 0.981         |
|             | P         | 0             | 0             |
|             | O.S.      | 0.037         | 0.079         |
|             | ST.       | 0.06          | 0.106         |
|             | New       | 0.052         | 0.11          |

In this simulation study, the Deviance test can not control type I error rate.

When the assumed model is correct model, the chance of rejecting the null hypothesis increases as the sample size gets larger and larger, which indicate that the large sample size does not help Deviance test to control type I error rate. In addition, all type I error rates of Deviance test are larger than the upper bound of 95% confidence interval. Type I error rates of Pearson Chi-square test are smaller than the lower bound of 95% confidence interval under different sample sizes, even zero per cent chance to reject



null hypothesis when the assumed model is true with  $n=200, 500$ . The ability of controlling type I error rate of Osius and Rojek's normal approximation test can be improved by larger sample size, for example, under the null hypothesis using  $\alpha=0.05$ , it has 3.7 per cent to reject the null hypothesis when  $n=500$  instead of 15.4 per cent when  $n=100$ . Type I error rate of Osius and Rojek's normal approximately normal test are greater than the upper bound of 95% confidence interval under  $\alpha=0.05$   $n=100$ ,  $\alpha=0.05$   $n=200$ ,  $\alpha=0.10$   $n=100$ . Based on Table 4-2, almost all type I error rates of Hosmer and Lemeshow's  $\hat{C}$  test, Hosmer and Lemeshow's  $\hat{H}$ , Stukel's score test and the proposed test fall within 95% confidence interval, under different sample size and different significance levels.

Next, the random samples were generated from the following model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 d + \beta_3 x_1 d \quad (4.4)$$

where  $\beta_3$  takes on the values 0, 0.1, 0.2, ..., 0.8, respectively. The proposed 4-group partition test and the six known tests were applied to the random sample generated from (4.4) to test the goodness-of-fit of the assumed model (4.3) at the level of significance  $\alpha = 0.05$ . The rejection rates of these tests were presented in Table 4-5 and graphically shown in Figure 4-4 to 4-6.

**Table 4-5 Rejection Rates for the Known and Proposed Tests for Study Two**

| Sample Size | $\beta_3$ | $\hat{C}$ | $\hat{H}$ | D     | P     | O.S.  | ST.   | New   |
|-------------|-----------|-----------|-----------|-------|-------|-------|-------|-------|
| 100         | 0         | 0.04      | 0.042     | 0.221 | 0.003 | 0.154 | 0.064 | 0.065 |
|             | 0.1       | 0.044     | 0.054     | 0.132 | 0.003 | 0.114 | 0.072 | 0.088 |
|             | 0.2       | 0.045     | 0.05      | 0.066 | 0.005 | 0.088 | 0.076 | 0.099 |
|             | 0.3       | 0.045     | 0.045     | 0.05  | 0.01  | 0.08  | 0.074 | 0.125 |
|             | 0.4       | 0.046     | 0.048     | 0.031 | 0.017 | 0.072 | 0.071 | 0.156 |
|             | 0.5       | 0.048     | 0.044     | 0.018 | 0.025 | 0.07  | 0.081 | 0.21  |
|             | 0.6       | 0.056     | 0.05      | 0.008 | 0.031 | 0.071 | 0.09  | 0.261 |
|             | 0.7       | 0.06      | 0.045     | 0.001 | 0.037 | 0.073 | 0.093 | 0.301 |
|             | 0.8       | 0.068     | 0.055     | 0     | 0.046 | 0.081 | 0.096 | 0.381 |
| 200         | 0         | 0.047     | 0.063     | 0.492 | 0     | 0.093 | 0.064 | 0.064 |
|             | 0.1       | 0.049     | 0.051     | 0.329 | 0     | 0.076 | 0.061 | 0.079 |
|             | 0.2       | 0.047     | 0.046     | 0.183 | 0.003 | 0.074 | 0.067 | 0.117 |
|             | 0.3       | 0.052     | 0.054     | 0.089 | 0.005 | 0.058 | 0.07  | 0.174 |
|             | 0.4       | 0.05      | 0.058     | 0.055 | 0.007 | 0.056 | 0.076 | 0.232 |
|             | 0.5       | 0.058     | 0.06      | 0.036 | 0.011 | 0.069 | 0.079 | 0.333 |
|             | 0.6       | 0.061     | 0.056     | 0.024 | 0.016 | 0.072 | 0.088 | 0.422 |
|             | 0.7       | 0.066     | 0.059     | 0.013 | 0.022 | 0.079 | 0.082 | 0.537 |
|             | 0.8       | 0.073     | 0.065     | 0.01  | 0.025 | 0.089 | 0.082 | 0.623 |
| 500         | 0         | 0.059     | 0.062     | 0.929 | 0     | 0.037 | 0.06  | 0.052 |
|             | 0.1       | 0.067     | 0.064     | 0.787 | 0     | 0.033 | 0.061 | 0.068 |
|             | 0.2       | 0.08      | 0.063     | 0.529 | 0     | 0.036 | 0.074 | 0.147 |
|             | 0.3       | 0.078     | 0.076     | 0.304 | 0     | 0.033 | 0.077 | 0.324 |
|             | 0.4       | 0.081     | 0.075     | 0.16  | 0     | 0.04  | 0.076 | 0.546 |
|             | 0.5       | 0.087     | 0.103     | 0.063 | 0.002 | 0.053 | 0.09  | 0.738 |
|             | 0.6       | 0.092     | 0.1       | 0.026 | 0.004 | 0.063 | 0.099 | 0.861 |
|             | 0.7       | 0.093     | 0.094     | 0.014 | 0.005 | 0.075 | 0.114 | 0.941 |
|             | 0.8       | 0.1       | 0.099     | 0.008 | 0.014 | 0.092 | 0.143 | 0.979 |

Figure 4-4 Plots of Rejection Rate for Study Two (n=100)

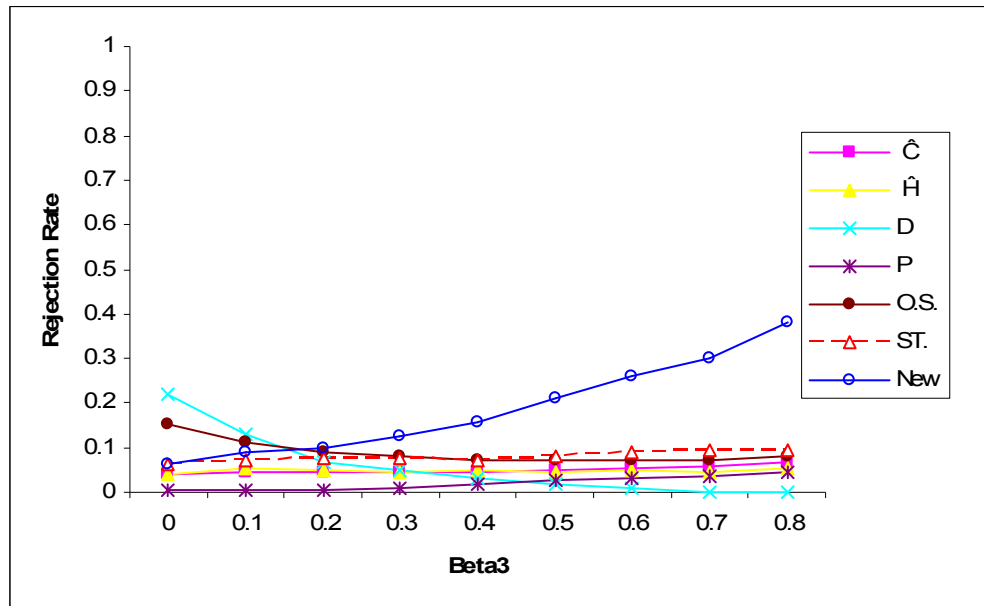
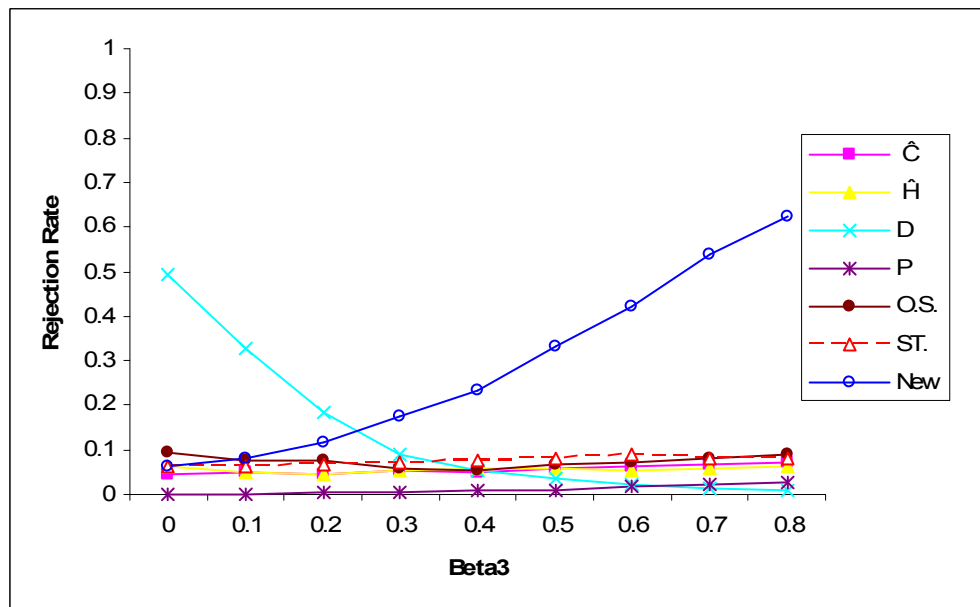
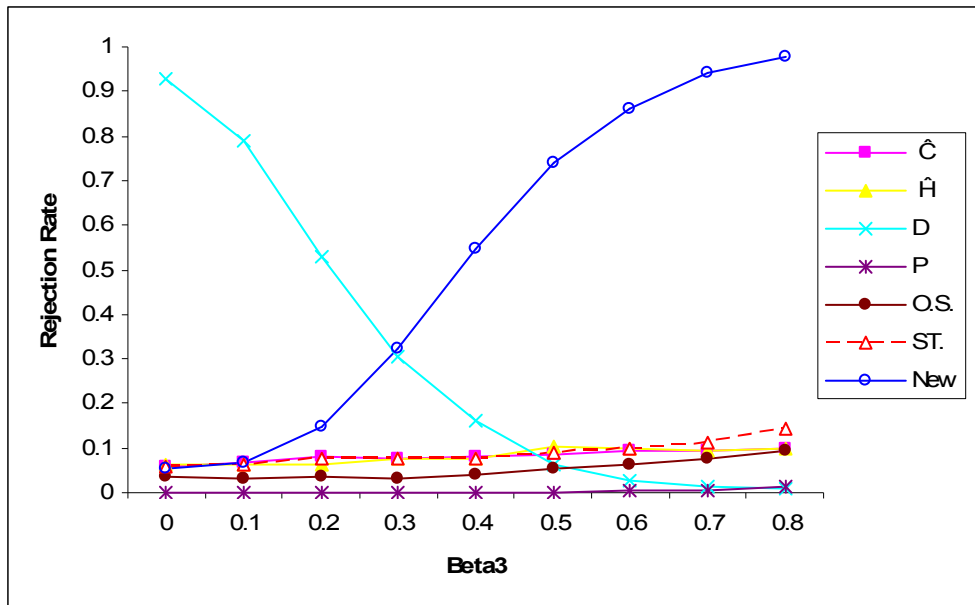


Figure 4-5 Plots of Rejection Rate for Study Two (n=200)



**Figure 4-6 Plots of Rejection Rate for Study Two (n=500)**



This simulation study results suggest that, the power of Deviance test is getting poorer and poorer when the assumed model departs further and further from the true model. For Pearson Chi-square test, bigger sample size and further departure from true model can not enhance its power to detect the lack of fit. This simulation study showed that the Deviance test and Pearson chi-square test are not applicable to the type one covariate pattern ( $J=n$ ). For Osius and Rojek's normal approximation test, larger sample sizes improve the ability to control the type one error rate. However, neither larger sample sizes nor further departure from the true model can improve the power of detecting lack-of-fit. Stukel's score test and Hosmer and Lemeshow's  $\hat{C}$  and  $\hat{H}$  tests have very similar performance: They control type I error rate, but the power of detecting lack of fit can not be improved much by larger sample size and further departure away from the true model. The proposed test has the best performance among these tests at

different sample sizes and degree of departure from the true model. The ability of controlling type I error rate and power of detecting lack of fit are improved by increasing the sample sizes. In summary, the proposed method is the best test in this simulation study.

### 4.3 The Third Simulation Study

In this section, we will use a more complicated logistic regression model than the previous two in which three predictor variables used in the model. We will first check whether the proposed test controls the type I error rate, when the data are generated from the following logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (4.5)$$

where  $\beta_0 = -1.3$ ,  $\beta_1 = 0.26$ ,  $\beta_2 = 0.26$ ,  $\beta_3 = 0.23$

In this simulation study, 1000 random samples of size  $n=100,200,500$  were respectively generated from model (4.5) with covariate  $x_1$  independently generated from an uniform distribution over the interval  $(-2,2)$ , the covariate  $x_2$  independently generated from the standard normal distribution with mean 0 and standard deviation 1, and the covariate  $x_3$  independently generated from the Chi-square distribution with degrees of freedom 4. The response variable  $y$  associated with the covariate  $x_1$ ,  $x_2$  and  $x_3$  was generated as follow: A random variable  $u$  is generated from a uniform distribution over the interval  $(0,1)$ . The response  $y$  was assigned a value 1, if  $\pi(\mathbf{x}) \geq u$ , and 0 otherwise.

The proposed 2-group partition test with cut-off point set at 50<sup>th</sup> percentile of the estimated probability from the assumed model and the six known tests mentioned before were applied to the random samples generated from model (4.5) at level of significance  $\alpha = 0.05, 0.1$ . The observed rejection rates are presented in Table 4-6, where the same labels used to identifying various tests in section 4.2 are used here.

**Table 4-6 Observed Type I Error Rate for Simulation Study Three**

| Sample Size | Methods   | $\alpha=0.05$ | $\alpha=0.10$ |
|-------------|-----------|---------------|---------------|
| 100         | $\hat{C}$ | 0.04          | 0.095         |
|             | $\hat{H}$ | 0.072         | 0.119         |
|             | D         | 0.537         | 0.771         |
|             | P         | 0.006         | 0.007         |
|             | O.S.      | 0.436         | 0.478         |
|             | ST.       | 0.087         | 0.146         |
|             | New       | 0.064         | 0.127         |
| 200         | $\hat{C}$ | 0.041         | 0.089         |
|             | $\hat{H}$ | 0.054         | 0.096         |
|             | D         | 0.894         | 0.972         |
|             | P         | 0.001         | 0.004         |
|             | O.S.      | 0.288         | 0.351         |
|             | ST.       | 0.066         | 0.119         |
|             | New       | 0.061         | 0.128         |
| 500         | $\hat{C}$ | 0.044         | 0.095         |
|             | $\hat{H}$ | 0.074         | 0.116         |
|             | D         | 1             | 1             |
|             | P         | 0.001         | 0.001         |
|             | O.S.      | 0.139         | 0.189         |
|             | ST.       | 0.052         | 0.105         |
|             | New       | 0.055         | 0.108         |

From Table 4-6, when the assumed model is the correct model, the Deviance test can not control type I error rate with these three different sample sizes, and the rejection rate increases as the sample size increases. For example, when  $n=500$ , there are 100 per cent chance to reject the null hypothesis with  $\alpha=0.05$  and  $\alpha=0.10$ . The larger sample sizes help Osius and Rojek's approximately normal test to reduce the rejection rate under the null hypothesis. Unfortunately, Osius and Rojek's approximately normal test can not control the type I error rate in this simulation study. Type I error rates of Pearson Chi-square test are less than the lower bound of 95% confidence

interval (Table 4-2) with different level of significance level and sample sizes. Type I error rate of Hosmer and Lemeshow's  $\hat{H}$  test, Hosmer-Lemeshow's  $\hat{C}$ , Stukel's score test and proposed test are very close to or fall within 95% confidence interval when sample sizes are 100, 200 and 500.

Next, the random samples were generated by the following model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2x_3 \quad (4.6)$$

where  $\beta_4$  takes the values 0, 0.2, 0.4, 0.6, 0.8, respectively. The proposed 2-group partition test and the six known tests were applied to the random sample generated from (4.6) to test the goodness-of-fit of the assumed model (4.5) at the level of significance  $\alpha = 0.05$ . The rejection rates of these tests were presented in the Table 4-7 and graphically in Figure 4-4 to 4-6.

**Table 4-7 Rejection Rates for the Known and Proposed Tests for Study Three**

| Sample Size | $\beta$ | $\hat{C}$ | $\hat{H}$ | D     | P     | O.S.  | ST.   | New   |
|-------------|---------|-----------|-----------|-------|-------|-------|-------|-------|
| 100         | 0       | 0.04      | 0.072     | 0.537 | 0.006 | 0.436 | 0.087 | 0.064 |
|             | 0.2     | 0.051     | 0.097     | 0.669 | 0.001 | 0.628 | 0.135 | 0.239 |
|             | 0.4     | 0.092     | 0.156     | 0.792 | 0.003 | 0.785 | 0.217 | 0.483 |
|             | 0.6     | 0.154     | 0.196     | 0.865 | 0.002 | 0.844 | 0.279 | 0.607 |
|             | 0.8     | 0.186     | 0.232     | 0.901 | 0     | 0.88  | 0.337 | 0.673 |
| 200         | 0       | 0.041     | 0.054     | 0.894 | 0.001 | 0.288 | 0.066 | 0.061 |
|             | 0.2     | 0.074     | 0.12      | 0.959 | 0.002 | 0.633 | 0.128 | 0.354 |
|             | 0.4     | 0.146     | 0.209     | 0.984 | 0.004 | 0.799 | 0.242 | 0.638 |
|             | 0.6     | 0.203     | 0.251     | 0.995 | 0.002 | 0.874 | 0.324 | 0.75  |
|             | 0.8     | 0.251     | 0.284     | 0.997 | 0     | 0.916 | 0.379 | 0.81  |
| 500         | 0       | 0.044     | 0.074     | 1     | 0.001 | 0.139 | 0.052 | 0.055 |
|             | 0.2     | 0.085     | 0.181     | 1     | 0.001 | 0.573 | 0.178 | 0.642 |
|             | 0.4     | 0.178     | 0.285     | 1     | 0     | 0.819 | 0.336 | 0.885 |
|             | 0.6     | 0.267     | 0.347     | 1     | 0     | 0.905 | 0.386 | 0.905 |
|             | 0.8     | 0.308     | 0.373     | 1     | 0     | 0.932 | 0.439 | 0.906 |



Figure 4-7 Plots of Rejection Rate for Study Three (n=100)

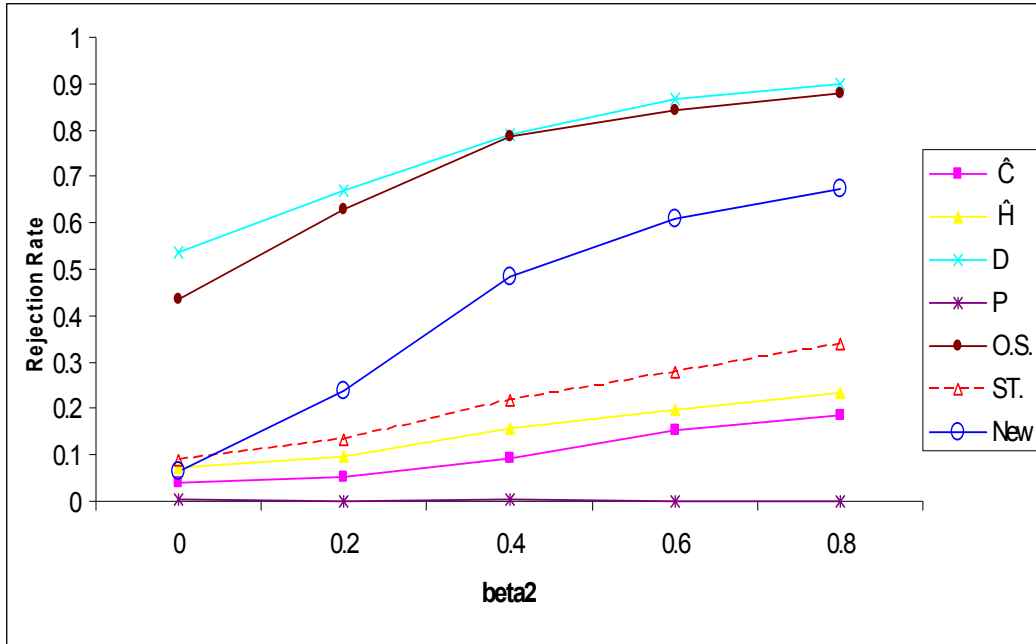
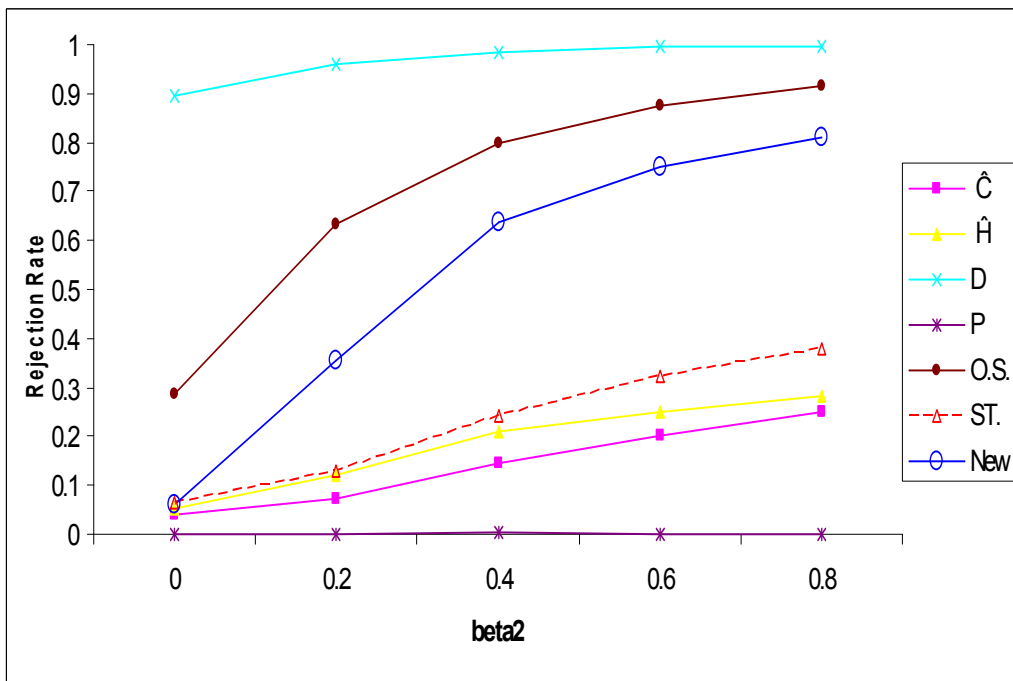
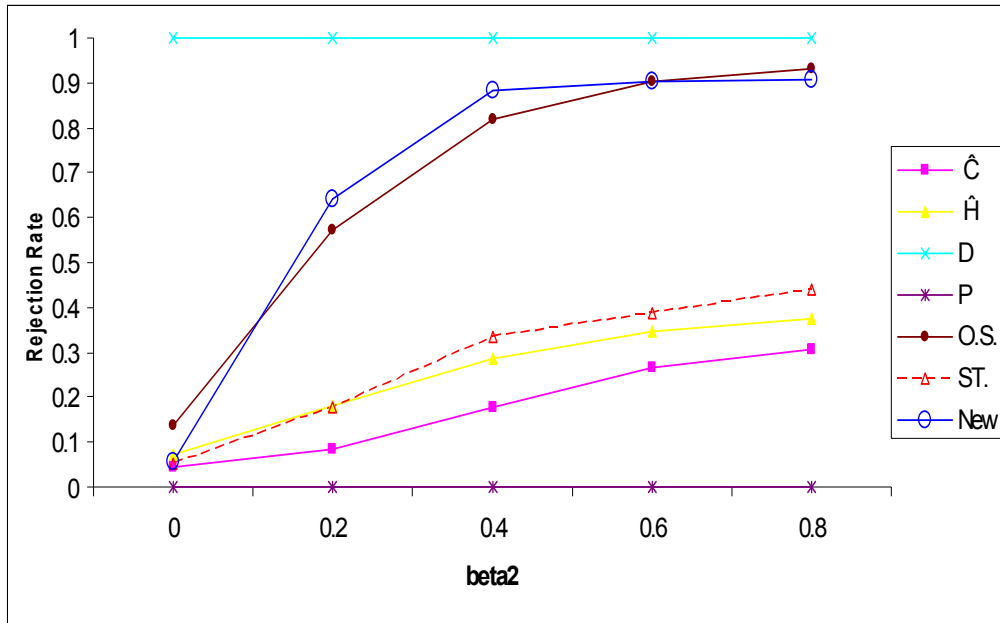


Figure 4-8 Plots of Rejection Rate for Study Three (n=200)



**Figure 4-9 Plots of Rejection Rate for Study Three (n=500)**



The simulation results suggest that the power of detecting lack of fit for Pearson Chi-square test is poor for different sample sizes and degree of departure from the true model. Since Deviance test has very high type I error rate, it is not necessary to compare its power with other tests. Once again, this simulation study gives a strong evidence that the Pearson Chi-square test and Deviance test are not applicable to the Type one covariate pattern ( $J=n$ ). Osius and Rojek's approximately normal test is powerful to reject the null hypothesis when the assumed model is not correct one. In this simulation study, the rejection rate of Stukel's score test, Hosmer- Lemeshow  $\hat{H}$  and  $\hat{C}$  tests are almost same or nearly close when the assumed model is not the true model. The proposed test is the best test among the seven tests because it controls the type one error rate and has strong power to reject the assumed model when it is not true one. In addition, the power increases when the sample size increases.

## 4.4 The Fourth Simulation Study

In the previous three simulation studies, the terms omitted from the generating model are associated with the covariates in the assumed model. In this simulation study, the covariate in the assumed model is not related to the added terms in the generating model.

The null hypothesis is that the following model adequately fits the data.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 \quad (4.7)$$

where  $\beta_0 = -0.3$ ,  $\beta_1 = 1.3$ ,  $x_1 \sim u(-6,6)$

Let the generating model be

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (4.8)$$

Where  $x_2 \sim \text{Bernulli}(0.5)$  and  $\beta_2$  takes on the values 0, 2, ..., 10.

In this simulation study, 1000 random samples of size  $n=100,200,500$  were respectively generated from model (4.5) with covariate  $x_1$  independently generated from a uniform distribution over the interval  $(-6, 6)$ , the covariate  $x_2$  independently generated from the Bernoulli distribution with probability 0.5. The response variable  $y$  associated with the covariate  $x_1, x_2$  was generated the same way as in the previous simulation studies.

The proposed 2-group partition test with cut-off point set at the 50<sup>th</sup> percentile of the estimated probability from the assumed model and the six known tests mentioned before were applied to the random samples generated from assumed model (4.7) at

level of significance  $\alpha = 0.05, 0.1$ . The observed rejection rates are presented in Table 4-7, the labels of various tests are same as those used in the previous simulation studies.

**Table 4-8 Observed Type I Error Rate for Simulation Study Four**

| Sample Size | Method    | $\alpha=0.05$ | $\alpha=0.10$ |
|-------------|-----------|---------------|---------------|
| 100         | $\hat{C}$ | 0.04          | 0.067         |
|             | $\hat{H}$ | 0.041         | 0.096         |
|             | D         | 0             | 0             |
|             | P         | 0.082         | 0.093         |
|             | O.S.      | 0.002         | 0.004         |
|             | ST.       | 0.081         | 0.154         |
|             | New       | 0.063         | 0.131         |
| 200         | $\hat{C}$ | 0.078         | 0.099         |
|             | $\hat{H}$ | 0.036         | 0.085         |
|             | D         | 0             | 0             |
|             | P         | 0.14          | 0.15          |
|             | O.S.      | 0.005         | 0.007         |
|             | ST.       | 0.064         | 0.128         |
|             | New       | 0.058         | 0.106         |
| 500         | $\hat{C}$ | 0.087         | 0.107         |
|             | $\hat{H}$ | 0.057         | 0.118         |
|             | D         | 0             | 0             |
|             | P         | 0.193         | 0.204         |
|             | O.S.      | 0.006         | 0.01          |
|             | ST.       | 0.073         | 0.12          |
|             | New       | 0.067         | 0.116         |

The results in Table 4-8 indicate that under the null hypothesis, the rejection rate of Pearson Chi-square test increases as the sample size increases which indicate it can not control type I error rate. Type I error rates of Chi-square test are greater than the upper bound of 95% confidence interval (Table 4-2) when  $n=200$  and  $n=500$ . The rejection rate of deviance test is 0 per cent when the null hypothesis is true regardless of the sample size and level of significance. The type I error rates of Stukel's score test are higher than the nominal value except  $n=200$  and  $\alpha=0.05$ . All type I error rates of Osius and Rojek's approximately normal test are lower than the lower bound of 95% confidence interval in this simulation study. All the type I error rates of the proposed test

and Hosmer-Lemeshow's  $\hat{H}$  test lie within the 95% of confidence interval or very close to the limits.

**Table 4-9 Rejection Rates for the Known and Proposed Tests for Study Four**

| size | $\beta_4$ | $\hat{C}$ | $\hat{H}$ | D | P     | O.S.  | ST.   | New   |
|------|-----------|-----------|-----------|---|-------|-------|-------|-------|
| 100  | 0         | 0.053     | 0.041     | 0 | 0.082 | 0.002 | 0.081 | 0.063 |
|      | 2         | 0.044     | 0.036     | 0 | 0.084 | 0.002 | 0.095 | 0.066 |
|      | 4         | 0.023     | 0.052     | 0 | 0.041 | 0.001 | 0.154 | 0.079 |
|      | 6         | 0.028     | 0.096     | 0 | 0.009 | 0     | 0.237 | 0.164 |
|      | 8         | 0.08      | 0.194     | 0 | 0.002 | 0.002 | 0.41  | 0.395 |
|      | 10        | 0.126     | 0.289     | 0 | 0.001 | 0.016 | 0.559 | 0.544 |
| 200  | 0         | 0.078     | 0.036     | 0 | 0.141 | 0.005 | 0.064 | 0.058 |
|      | 2         | 0.037     | 0.037     | 0 | 0.123 | 0.001 | 0.08  | 0.055 |
|      | 4         | 0.02      | 0.065     | 0 | 0.042 | 0     | 0.16  | 0.079 |
|      | 6         | 0.072     | 0.169     | 0 | 0.011 | 0.001 | 0.4   | 0.275 |
|      | 8         | 0.21      | 0.389     | 0 | 0     | 0.037 | 0.732 | 0.692 |
|      | 10        | 0.397     | 0.6       | 0 | 0     | 0.19  | 0.876 | 0.842 |
| 500  | 0         | 0.087     | 0.057     | 0 | 0.193 | 0.006 | 0.073 | 0.067 |
|      | 2         | 0.037     | 0.055     | 0 | 0.136 | 0.004 | 0.084 | 0.057 |
|      | 4         | 0.05      | 0.148     | 0 | 0.029 | 0.001 | 0.309 | 0.166 |
|      | 6         | 0.309     | 0.447     | 0 | 0.001 | 0.002 | 0.787 | 0.646 |
|      | 8         | 0.76      | 0.858     | 0 | 0     | 0.475 | 0.986 | 0.982 |
|      | 10        | 0.943     | 0.969     | 0 | 0     | 0.881 | 0.999 | 0.998 |

**Figure 4-10 Plots of Rejection Rate for Study Four (n=100)**

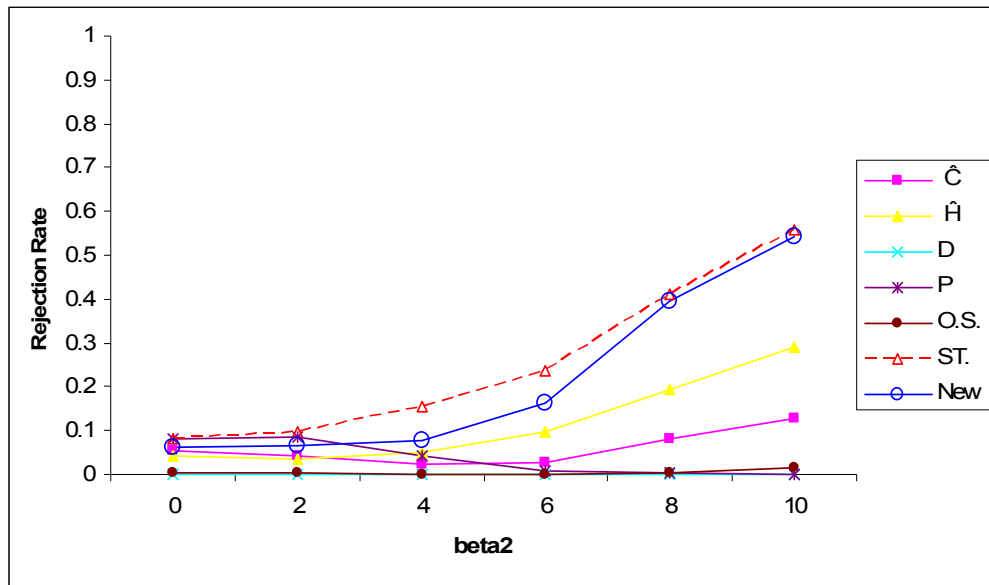


Figure 4-11 Plots of Rejection Rate for Study Four (n=200)

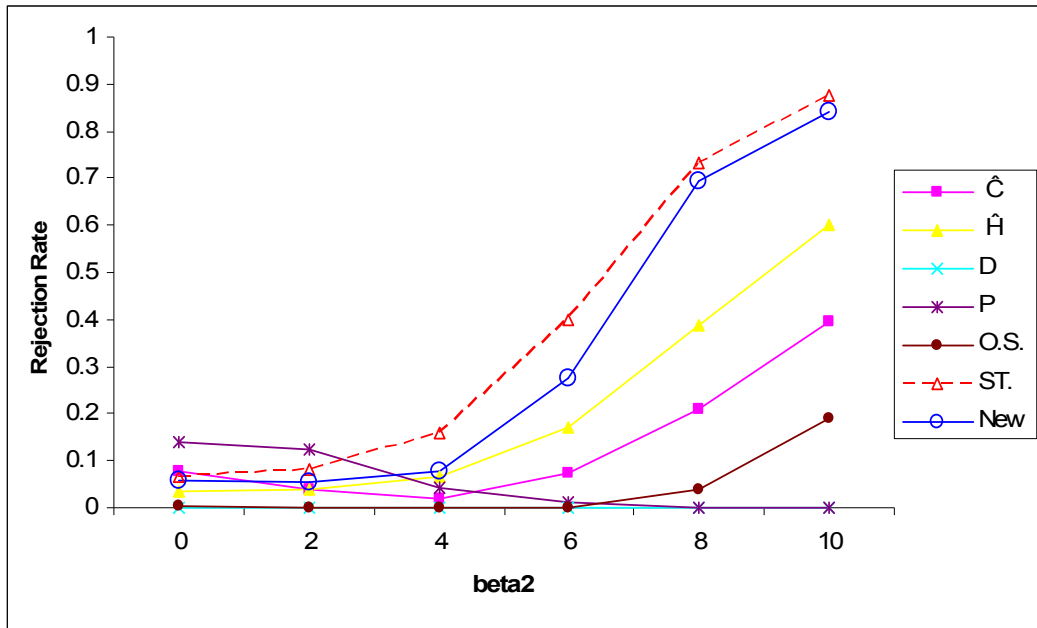
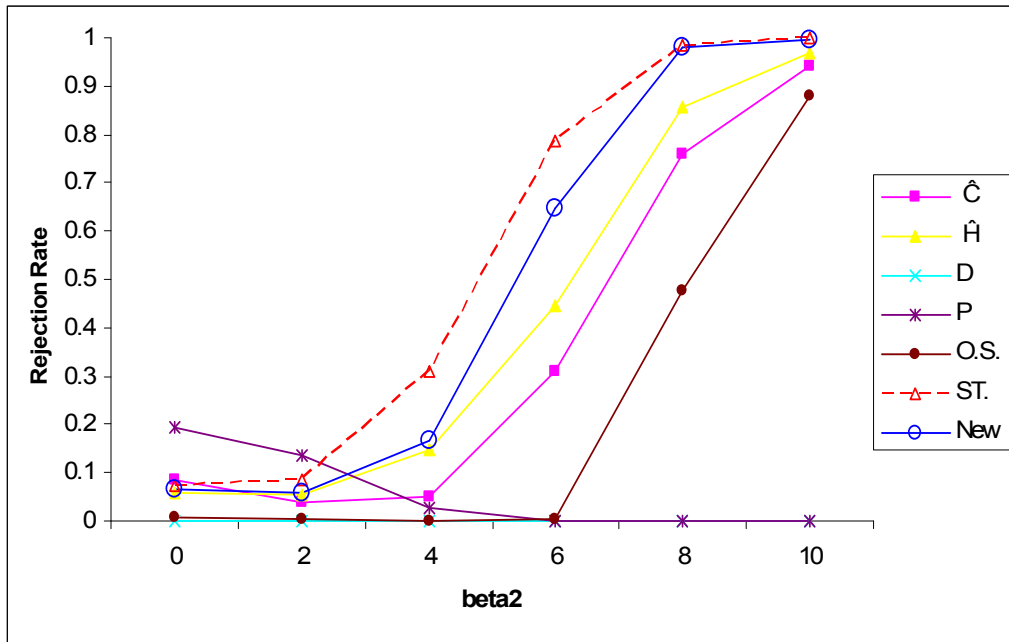


Figure 4-12 Plots of Rejection Rate for Study Four (n=500)



The results in Table 4-9 indicate that the power of Pearson chi-square test Deviance test are poor. In addition, their power of detecting lack of fit can not be

improved by larger sample sizes or further departure from the true model. The Osius and Rojek's approximately normal test has poor power of detecting the lack of fit, when  $n=100$  and  $n=200$ . The power is improved by the larger sample size and further departure from the true model. The performance of Hosmer-Lemeshow  $\hat{H}$  test is better than that of Hosmer-Lemeshow's  $\hat{C}$  test. Hosmer-Lemeshow  $\hat{H}$  test control type I error rate well and has a higher power of detecting the lack of fit when  $n=100$ ,  $n=200$  and  $500$  under the alternative hypothesis. Stukel's score test followed by the proposed test has the highest power of detecting the lack of fit under the alternative model. The proposed method had good performance in this simulation in controlling type I error rate well and in detecting lack-of-fit when it is present in the assumed model.

## 4.5 Summary on Simulation Studies

These four simulations cover four different situations. These four different simulation studies provide evidence again that the goodness-of-fit tests of Pearson chi-square and deviance are not suitable when the number of unique covariate patterns is equal or almost equal to the number of subjects.

Hosmer-Lemeshow  $\hat{H}$  test and proposed test always control type I error rate well in these four studies. Hosmer-Lemeshow  $\hat{C}$  test and Stukel's score test control type I error in the first three simulation studies; in the fourth study, they can not control type I error rate when  $\alpha=0.05$  and  $n=500$ . The Osius and Rojek's approximately normal test controls type I error rate in the first simulation study. In the other three studies, the ability of controlling type I error rate of the Osius and Rojek's approximately normal test is affected by the sample size and the degree of departure from the true model.

The first and forth simulations involve a simple logistic regression model (assumed model has only one predictor variable). When the assumed model is a simple model, Stukel's score has higher power of detecting lack of fit than others. The order of power is, Stukel > proposed test > Hosmer-Lemeshow tests > Osius and Rojek's approximately normal test. Hosmer-Lemeshow  $\hat{C}$  test is better than Hosmer-Lemeshow  $\hat{H}$  test in the first simulation; however, Hosmer-Lemeshow  $\hat{H}$  test is better than Hosmer-Lemeshow  $\hat{C}$  test in the fourth simulation.

The second and third simulations considered multiple logistic regression model (assumed model has more than one predictor variables). The proposed test performs best among all tests considered. It controls type I error rate well and has higher power to detect lack-of-fit.



## 4.6 Convergence Problem in the Simulation Studies

The parameters of logistic regression models are estimated by the maximum likelihood method. The log-likelihood is a strictly concave function and the maximum likelihood estimates of the parameters exist and are unique (Wedderburn 1976). Two general iterative algorithms are usually used to obtain the maximum likelihood estimates and are available in SAS for a logistic regression model: the Newton-Raphson algorithm and the Fisher-scoring algorithm. The Fisher-scoring algorithm is the default method in SAS PROC LOGISTIC, which is equivalent to fitting the model by iteratively reweighted least squares. Same parameter estimates and a slightly different estimated covariance matrices are given by these two algorithms, the reason is that the Fisher-scoring algorithm is based on the expected information matrix and the Newton-Raphson algorithm is based on the observed information matrix.

Let the first derivatives and second derivatives of the log likelihood function be, respectively

$$\mathbf{U}(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_i \mathbf{x}_i y_i - \sum_i \mathbf{x}_i y_i$$

$$\mathbf{I}(\beta) = \frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} = \sum_i \mathbf{x}_i \mathbf{x}_i' y_i (1 - y_i)$$

Then the iterative equation of the Newton-Raphson algorithm and Fisher-Scoring algorithm are, respectively,

$$\beta^{(t+1)} = \beta^{(t)} - \mathbf{I}^{-1}(\beta^{(t)}) \mathbf{U}(\beta^{(t)})$$

$$\beta^{(t+1)} = \beta^{(t)} + E[\mathbf{I}^{-1}(\beta^{(t)})] \mathbf{U}(\beta^{(t)})$$

For large iteration  $t$  and  $j^{\text{th}}$  ( $j=1, 2, \dots, p$ ) parameter, the convergence satisfies,

$$\left| \beta_j^{(t+1)} - \hat{\beta}_j \right| \leq c \left| \beta_j^{(t)} - \hat{\beta}_j \right| \quad \text{for some } c > 0$$

For the logistic regression, fortunately, its log likelihood is globally concave, which indicate it has at most one maximum (Amemiya, 1985). In fact, convergence works most of the time, but not every time. If the Hessian matrix (the partial second derivatives) is singular or near singular, a convergence problem occurs, the estimated logistic regression model is questionable and not reliable (Sherrod 2003). When the convergence problem occurs, Tian and Liu (2006) believe that there exists not only a numerical problem but also an indication that the model is not suitable for fitting the data.

Albert and Anderson (1984) and Santner and Duffy (1985) pointed out the existence of maximum likelihood estimates for the logistic model depends on how the sample points are distributed. The existence of data separation will lead to failure to converge (Prusseuw and Christmann, 2003, Altman, Gill and McDonald, 2004). Data separation can be categorized into three types: complete separation, quasi-complete separation, and overlap. The data are completely separated, if there exists a vector  $\mathbf{b}$  that allocates the observations to their two response groups in the following manner:

If  $\mathbf{b}'\mathbf{x}_i > 0$  then  $y_i = 1$ ,

If  $\mathbf{b}'\mathbf{x}_i < 0$  then  $y_i = 0$ .

Under this case, the maximum likelihood estimate for any covariates and covariate patterns does not exist (in another words, the likelihood equation does not have a finite

solution). The data is quasi-completely separated, if there exists a coefficient vector  $\mathbf{b}$  such that

If  $\mathbf{b}'\mathbf{x}_i \geq 0$  then  $y_i = 1$ ,

If  $\mathbf{b}'\mathbf{x}_i < 0$  then  $y_i = 0$ .

Under this case, the maximum likelihood estimate also does not exist. Under the case of quasi-complete separation, as the number of iterations increases, the log-likelihood does not reach to 0 as in the complete separation case and the dispersion matrix is unbounded. Complete separation and quasi-complete separation also lead to convergence failure with other link function for binary response variable, such probit link, log-log link. The data which are not of complete or quasi-complete separation type is called overlap data. The maximum likelihood estimate exists and is unique for overlap data.

The empirical method of Albert and Anderson (1984) is built in the PROC LOGISTIC in SAS. It can help us detect the complete separation and quasi-complete separation data. Albert and Anderson (1984) gave the following three steps in detecting data separation:

1. If the convergence criteria is satisfied with eight iteration, one declares that there is no problem.
2. After eight iterations, the probability of observed response predicted for the  $i^{\text{th}}$  subject is obtained by

$$\hat{y}_i = \frac{1}{1 + e^{[(2y_i - 1)\hat{\beta}\mathbf{x}_i]}}$$

If the predicted probability is 1 for all subjects after the 8<sup>th</sup> iteration, we declare that complete separation occurs and stop the iteration.

3. If the observed probabilities for some subjects are large than 0.95, then examine estimate standard errors for that iteration. If they exceed some specified value, we declare that quasi-complete separation occurs and stop the iteration.

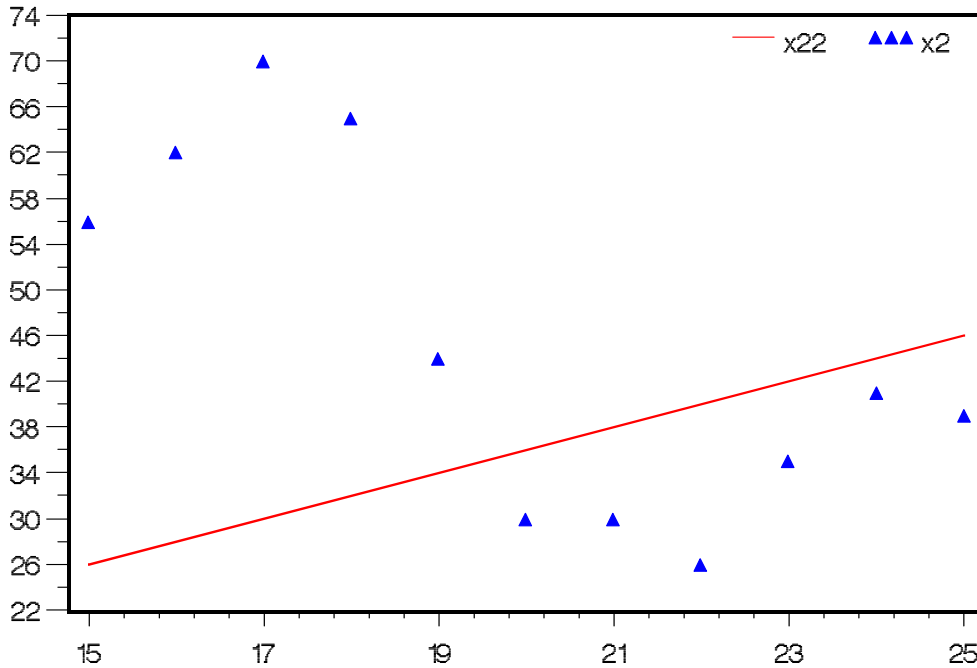
The following example illustrates that the Newton-- Raphson algorithm fail to converge for a complete separation data with eleven separation data points (Table 4-10).  $y$  is the response variable and  $x_1$  and  $x_2$  are two covariates.

**Table 4-10 Complete Separation Data**

| obs | x1 | x2 | x22 | b'x | y |
|-----|----|----|-----|-----|---|
| 1   | 15 | 56 | 26  | 30  | 0 |
| 2   | 16 | 62 | 28  | 34  | 0 |
| 3   | 17 | 70 | 30  | 40  | 0 |
| 4   | 18 | 65 | 32  | 33  | 0 |
| 5   | 19 | 44 | 34  | 10  | 0 |
| 6   | 20 | 30 | 36  | -6  | 1 |
| 7   | 21 | 30 | 38  | -8  | 1 |
| 8   | 22 | 26 | 40  | -14 | 1 |
| 9   | 23 | 35 | 42  | -7  | 1 |
| 10  | 24 | 41 | 44  | -3  | 1 |
| 11  | 25 | 39 | 46  | -7  | 1 |

Figure 4-13 shows that the vector  $b=(4,-2,1)'$  completely separates the data points into two response groups; that is, all observations with the same response lie on the same side of the line  $x_{22}=2x_1-4$ .

**Figure 4-13 Scatter Plot Sample Points with Complete Separation Data**



The iterative history of fitting a logistic regression model on the above data is shown in the Table 4-10. Notice that the Log-likelihood decreases as the iteration increases. The estimated covariance matrices for iterations 0, 5, 15, and 25 are shown in the Table 4-11 which indicates that the likely larger variance of pseudo-estimates may occur with the number of iteration increases.

**Table 4-11 Partial Logistic Iteration Steps Printout**

| Iteration History For Parameter Estimates |       |               |           |           |           |
|---|-------|---------------|-----------|-----------|-----------|
| Iter                                      | Ridge | LogLikelihood | Intercept | X1        | X3        |
| 0   | 0     | -1.8635644    | 3.1567224 | -0.342496 | 0.0771645 |
| 1   | 0     | -1.0049775    | 6.0528044 | -0.56796  | 0.1163513 |
| 2   | 0     | -0.5248162    | 9.2794008 | -0.832715 | 0.172829  |
| 3   | 0     | -0.2087828    | 11.199218 | -1.124201 | 0.2791172 |
| 4   | 0     | -0.0759306    | 12.549511 | -1.419777 | 0.400694  |
| 5   | 0     | -0.0279972    | 14.076113 | -1.720906 | 0.5193476 |
| 6   | 0     | -0.0103793    | 15.755734 | -2.028493 | 0.6368277 |
| 7   | 0     | -0.0038496    | 17.541144 | -2.341719 | 0.7541899 |
| 8   | 0     | -0.0014261    | 19.397439 | -2.659346 | 0.8718078 |
| 9   | 0     | -0.0005275    | 21.300264 | -2.980215 | 0.9897722 |
| 10  | 0     | -0.0001949    | 23.233159 | -3.303393 | 1.108065  |
| 11  | 0     | -0.0000719    | 25.185276 | -3.628187 | 1.2266328 |
| 12  | 0     | -0.0000265    | 27.149593 | -3.954098 | 1.3454174 |
| 13  | 0     | -9.7666E-6    | 29.121624 | -4.280775 | 1.4643671 |
| 14  | 0     | -3.5969E-6    | 31.09852  | -4.607978 | 1.58344   |
| 15  | 0     | -1.3243E-6    | 33.078479 | -4.935542 | 1.7026037 |
| 16  | 0     | -4.8745E-7    | 35.060364 | -5.263352 | 1.8218338 |
| 17  | 0     | -1.7939E-7    | 37.043459 | -5.591331 | 1.941112  |
| 18  | 0     | -6.6014E-8    | 39.027313 | -5.919427 | 2.0604251 |
| 19  | 0     | -2.429E-8     | 41.011641 | -6.247603 | 2.1797634 |
| 20  | 0     | -8.937E-9     | 42.996265 | -6.575835 | 2.2991199 |
| 21  | 0     | -3.2881E-9    | 44.981071 | -6.904105 | 2.4184894 |
| 22  | 0     | -1.2097E-9    | 46.965989 | -7.232402 | 2.5378684 |
| 23  | 0     | -4.45E-10     | 48.950975 | -7.560718 | 2.6572541 |
| 24  | 0     | -1.637E-10    | 50.936002 | -7.889046 | 2.7766446 |
| 25  | 0     | -6.023E-11    | 52.921053 | -8.217383 | 2.8960387 |

WARNING: Convergence not attained in 25 iterations.

WARNING: The procedure is continuing but the validity of the model fit is questionable.

WARNING: The specified model did not converge.

WARNING: Iteration limit exceeded.

**Table 4-12 Dispersion Matrices on the Selected Iterations**

| <b>Estimated Covariance Matrix</b> |           |           |           |
|------------------------------------|-----------|-----------|-----------|
| Iter=0 Loglikelihood=-1.8635644    |           |           |           |
|                                    | Intercept | X1        | X2        |
| Intercept                          | 136.81    | -4.88128  | -0.88115  |
| X1                                 | -4.88128  | 0.18724   | 0.02601   |
| X2                                 | -0.88115  | 0.02601   | 0.008331  |
| Iter=5 Loglikelihood=-0.0279972    |           |           |           |
| Intercept                          | 9143.30   | -351.02   | -54.06806 |
| X1                                 | -351.02   | 15.69193  | 0.95600   |
| X2                                 | -54.06806 | 0.95600   | 0.90934   |
| Iter=15 Loglikelihood=-1.3243E-6   |           |           |           |
| Intercept                          | 162328548 | -6318660  | -910109   |
| X1                                 | -6318660  | 295466    | 10448.52  |
| X2                                 | -910109   | 10448.52  | 18198.23  |
| Iter=25 Loglikelihood=-6.023E-11   |           |           |           |
| Intercept                          | -5.72E-13 | 5.075E-11 | -3.55E-10 |
| X1                                 | 6.023E-11 | 1.2061E-9 | 2.3201E-9 |
| X2                                 | 1.2061E-9 | 2.4307E-8 | 4.6367E-8 |

Complete separation and quasi-complete separation often occur in small sparse data set. The complete separation and quasi-complete separation data belong to sparse data where each covariate pattern has very few subjects (Derr 2000, Kuss 2002). In simulation studies, we did encounter the non-convergence problem. For example, in the

first simulation, if we use  $n=50$  with 1000 replications, non-convergence exist in 17 replications; if we use  $n=100$ , the non-convergence problem does not occur any more. The smaller the data set, the higher the chance of occurrence of complete separation or quasi-complete separation.



## CHAPTER 5 - Summary

Most known tests perform well in detecting the overall lack of fit for the type two covariate pattern ( $J \ll n$ ). However, for the type one covariate pattern ( $J = n$ ), Pearson Chi-square test and Deviance test perform poorly for which they either can not control type I error or have weak power in detecting the lack-of-fit of the assumed model. Hosmer-Lemeshow tests are recommended to solve this problem by combining multiple unique covariate patterns into one group. For the simple logistic regression model in the first simulation study, our work confirm that all tests with the exception of Pearson-square test and Deviance test have good performance. For the simple logistic regression model in the fourth simulation study, the proposed partition test, Stukel's score test and Hosmer-Lemeshow's tests have similar performance as in the first simulation study. Osius and Rojek's approximately normal test can not control the type I error rate with small data set. Stukel's score test is a little bit better than the proposed test for the simple logistic regression model.

For the multiple logistic regression model (the second and third simulation studies), the proposed method has good control of the Type I error rate, and it has higher power in detecting lack-of-fit than the known tests mentioned earlier under

various settings considered. The proposed method is more sensitive than other known tests in detecting the departure of the assumed model from the true model.

The proposed test has the best overall performance. It seems to perform well with overall steady rejection rate, with the type I rates lying within 95% confidence interval in all situations considered. For  $J=n$  and multiple logistic regression, the result of other known tests seemed not reliable, they are unable to control type I error rate or have poor power in detecting lack-of-fit.

In chapter 3, the illustrative examples showed that the proposed method performs fairly well in detecting overall lack-of-fit. If the overall goodness-of-fit test is significant, we can also determine the nature of lack-of-fit, for example, checking which part of the model of  $\pi(x)$  fitting the data inadequately. This could be useful for the researchers who want to build a better model and want to know which parts of his/her model fit the data and which parts do not. In this dissertation we also propose a forward stepwise procedure to build a partition logistic regression model, which can improve the fit of the standard logistic regression model.

The proposed method has some weakness, although it is generally performs better than other known tests. There may be converge problem with the number of partition groups increases. If the assumed model is totally incorrect and there is huge difference in the success probabilities between the assumed and the true model, the partition method may not produce a model that fits better than the assumed model.

## CHAPTER 6 - Extension and Future Study

In the near future, we will try to do the following:

- (1) Extend the proposed idea of testing goodness-of-fit to other generalized linear models, such as, log-linear models, or multinomial logistic regression models.
- (2) Prove that the sampling distribution of the proposed test statistics can be approximated by a chi-square distribution with appropriate degrees of freedom when the sample size is large.
- (3) Extend the proposed idea of testing goodness-of-fit to correlated data sets for logistic regression and other generalized linear models, such as, log-linear models, multinomial logistic regression models.

## CHAPTER 7 - References

Agresti, A., (1990). *Categorical data analysis*. Wiley interscience.

Albert A. and Anderson, J.A. (1984), On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**,1-10.

Altman, M., Gill, J and McDonald, M.P. (2004) *Numerical issues in Statistical Computing for the Social Scientist*. John Wiley & Sons, Inc.

Akaike, Hirotugu (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6): 716–723.

Coull, B. A. and Agresti, A. (2000) Random effects modeling of multiple binary responses using the multivariate binomial logit-normal distribution, *Biometrics* **56**: 73-80.

Amemiya, Y., 1985 What should be done when an estimated between-group covariance-matrix is not nonnegative definite. *Am. Stat.* **39**: 112–117.

Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. *Biometrics* **68**, 357-363.

Azzalinni, A., Bowman, A.W., Härdle.,W (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**:1-12.

- Bedrick, E.J., and Hill, J. R., (1996). Assessing the fit of logistics model to individual matched sets of case control data. *Biometrics* 52, 1-9.
- Berkson, J. (1944) Application of logistic function to bio-assay. *Journal of the American Statistical Association* 39, 357-365.
- Bender, R, Grouven, U., (1996) Logistic regression models used in medical research are poorly presented [letter]. *BMJ*, 313,628.
- Bertolini, G. et al. (2000). One mdel, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for logistic regression model. *Journal of Epidemiology and Biostatistics* 5, 251-253.
- Bliss, C.I. (1935), the calculation of the Dosage-Mortality curve. *Annals of Applied Biology* 22, 134-167.
- Brown, C.C. (1982). On a goodness-of-fit test for the logistic model based on score statistics. *Communications in statistics* 11, 1087-1105.
- Christensen, R. (1997). Log-linear models and logistic regression. 2<sup>nd</sup> ed. NY: Springer.
- Collet, D. (1991). Modeling binary data. Chapman and Hall, London.
- Cook, R.D. and Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.
- Copas, J.B. (1983). Plotting P against x. *Aplied Statsitics* 32:25-31.

Dalal, S.R., Fowlkes, E.B., and Hoadley, B. (1989) Risk analysis of the space shuttle: Pre—Challenger predication of failure. *Journal of the American Statistical Association* 84, 954-957.

Derr, R.E. (2000). Performing exact logistic regression with the **SAS** system, **SAS** Paper P254-25, 1–10.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). Analysis of longitudinal data. New York: Oxford University Press.

Dreiseitl, S. et al. (2005). Nomographic representation of logistic regression models: a case study using patient self-assessment data. *Journal of Biomedical Informatics* 38, 389-394

Evans, SR, and Li, L. (2005). A Comparison of Goodness of Fit Tests for the Logistic GEE Model, *Statistics in Medicine* 24:1245-1261.

Fiegl, P. and Zelen, M. (1965). Estimation of exponential probabilities with concomitant information. *Biometrics* 21, 826-838.

Finney, D.J. (1941). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* 34, 320-324.

Finney, D.J. (1971). Probit analysis (3<sup>rd</sup> ed.), Cambridge, U.K.:Cambridge University press.

Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* 74: 503-515.

Haberman, S.J. (1978) Analysis of Qualitative Data, 1: Introductory Topics, Academic Press, New York.

Hosmer, D. W. and Lemeshow, S. (1989). Applied logistic regression. 1<sup>st</sup> ed. NY: Wiley and Sons.

Hosmer, D. W. and Lemeshow, S. (2000). Applied logistic regression. 2<sup>nd</sup> ed. NY: Wiley and Sons.

Hosmer, D.W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for logistic regression model. *Statistics in medicine* 16:9, 965-80.

Hosmer, D.W., Lemeshow, S., and Klar, J. (1988). Goodness-of-fit testing for the multiple logistic regression analysis when the estimated probabilities are small. *Biometrical Journal*, **30**, 911-924.

Hosmer, D.W., Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics* **A10**, 1043-1069.

Johnson, W. (1985). Influence measurement for logistic regression: Another point of view. *Biometrika* **72**, 59-65.

Kuss, O. (2002). Global goodness-of-fit tests in logistic regression with sparse data. *Statistics in Medicine* 21, 3789-3801.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models (with discussion). *Journal of the American Statistical Association* 79: 61-83.

Lavine, M. (1991), Problems in extrapolation illustrated with space shuttle o-ring data (with discussion), *Journal of the American Statistical Association* **86**, 9191-921.

Le Cessie,S., and van Houwelingen, J.C. (1991). A goodness of fit test for binary regression models based on smoothing methods. *Biometrics* **47**:1267-1282.

Lemeshow, S., and Hosmer, D. W.(1982). The use of goodness-of-fit statistics in the development of logistics regression models. *American Journal of Epidemiology* **115**, 92-106.

Luigi Bocconi, a cura di R. Piccarreta. (1993). Categorical Data Analysis. Quaderni del Corso Estivo di Statistica e Calcolo delle Probabilita;, n. 4., Istituto di Metodi Quantitativi, Universita.

Martz, H.F. and Zimmer, W.J., (1992), The risk of catastrophic failure of the solid rocket booster on the space shuttle. *The American Statistician* **46**, 42-47.

McCullagh, P.and Nelder,J.A. (1989). Generalized linear model. London: Chapman and Hall.

Milicer, H., and Szczotka, F. (1996), Age at Menarche in Warsaw Girls in 1965. *Human biology* **38**, 199-203.

Nelder, J., and R.W. M. Wedderburn. (1972). Generalized linear models. *J. Roy. Statist. Soc.* **A135**: 370-384.

Osborne, M.R. (1992). Fisher's method of scoring. *International Statistical Review / Revue Internationale de Statistique*, **60**: 99-117.

Osius, G., and Rojek,D. (1992). Normal Goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association* **87**: 1145-1152.



Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag., series 5*, **50**:157-175.

Pregibon, Daryl (1981). Logistic regression diagnostics. *Annals of Statistics* **9**, 705-724.

Prentice, R.L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics* **32**: 761-768.

Pulkstenis, E., Robinson, T.J. (2002) Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine* **21**, 79-93.

Rousseeuw, P.J., Christmann, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis* **43**, 315-332.

SAS/STAT software (2003). SAS Institute, Cary, NC.

SAS document v8 <http://v8doc.sas.com/sashtml/>

Santner T.J. and Duffy, E.D. (1986). A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73**, pp. 755-758.

Schwarz, G., (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

- Sherrod, P.H., DTREG predictive modeling software (2003), DTREG. inc.
- Stukel, T.A. (1988). Generalized logistic models. *Journal of American Statistical Association* **83**, 426-431.
- Tabachnick, B., and Fidell, L. (1996). Using multivariate statistics, third edition. Harper Collins.
- Tian, L and Liu, L (2006). A Letter to the Editor RE: "EASY SAS CALCULATIONS FOR RISK OR PREVALENCE RATIOS AND DIFFERENCES" *American Journal of Epidemiology* 163(12):1157-1158.
- Tan, Q., et.al (2004). Haplotype effects on human survival: logistic regression models applied to unphased genotype data. *Annals of human genetics* 68,168-175.
- Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistics regression model. *Biometrika* **67**, 250-251.
- Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**: 27-32.
- Whittaker, E. T. and Watson, G. N. (1990). A course in modern analysis. 4 th ed. Cambridge, England: Cambridge University Press.

## Appendix A - Data on O-Ring Failure

| Case | Flight | Failure | Success | Temperature |
|------|--------|---------|---------|-------------|
| 1    | 14     | 1       | 0       | 53          |
| 2    | 9      | 1       | 0       | 57          |
| 3    | 23     | 1       | 0       | 58          |
| 4    | 10     | 1       | 0       | 63          |
| 5    | 1      | 0       | 1       | 66          |
| 6    | 5      | 0       | 1       | 67          |
| 7    | 13     | 0       | 1       | 67          |
| 8    | 15     | 0       | 1       | 67          |
| 9    | 4      | 0       | 1       | 68          |
| 10   | 3      | 0       | 1       | 69          |
| 11   | 8      | 0       | 1       | 70          |
| 12   | 17     | 0       | 1       | 70          |
| 13   | 2      | 1       | 0       | 70          |
| 14   | 11     | 1       | 0       | 70          |
| 15   | 6      | 0       | 1       | 72          |
| 16   | 7      | 0       | 1       | 73          |
| 17   | 16     | 0       | 1       | 75          |
| 18   | 21     | 1       | 0       | 75          |
| 19   | 19     | 0       | 1       | 76          |
| 20   | 22     | 0       | 1       | 76          |
| 21   | 12     | 0       | 1       | 78          |
| 22   | 20     | 0       | 1       | 79          |
| 23   | 18     | 0       | 1       | 81          |

## Appendix B - Data on Vasoconstriction

| Constriction | Volume | Rate |
|--------------|--------|------|
| 1            | 0.825  | 3.7  |
| 1            | 1.09   | 3.5  |
| 1            | 2.5    | 1.25 |
| 1            | 1.5    | 0.75 |
| 1            | 3.2    | 0.8  |
| 1            | 3.5    | 0.7  |
| 0            | 0.75   | 0.6  |
| 0            | 1.7    | 1.1  |
| 0            | 0.75   | 0.9  |
| 0            | 0.45   | 0.9  |
| 0            | 0.57   | 0.8  |
| 0            | 2.75   | 0.55 |
| 0            | 3      | 0.6  |
| 1            | 2.33   | 1.4  |
| 1            | 3.75   | 0.75 |
| 1            | 1.64   | 2.3  |
| 1            | 1.6    | 3.2  |
| 1            | 1.415  | 0.85 |
| 0            | 1.06   | 1.7  |
| 1            | 1.8    | 1.8  |
| 0            | 2      | 0.4  |
| 0            | 1.36   | 0.95 |
| 0            | 1.35   | 1.35 |
| 0            | 1.36   | 1.5  |
| 1            | 1.78   | 1.6  |

|   |       |      |
|---|-------|------|
| 0 | 1.5   | 0.6  |
| 1 | 1.5   | 1.8  |
| 0 | 1.9   | 0.95 |
| 1 | 0.95  | 1.9  |
| 0 | 0.4   | 1.6  |
| 1 | 0.75  | 2.7  |
| 0 | 0.03  | 2.35 |
| 0 | 1.83  | 1.1  |
| 1 | 2.2   | 1.1  |
| 1 | 2     | 1.2  |
| 1 | 3.33  | 0.8  |
| 0 | 1.9   | 0.95 |
| 0 | 1.9   | 0.75 |
| 1 | 1.625 | 1.3  |

## Appendix C - Data on Leukemia Survival

| Survival | Cell count | Log (cell count) | AG |
|----------|------------|------------------|----|
| 1        | 2,300      | 3.36             | +  |
| 1        | 750        | 2.87             | +  |
| 1        | 4,300      | 3.63             | +  |
| 1        | 2,600      | 3.41             | +  |
| 0        | 6,000      | 3.78             | +  |
| 1        | 10,500     | 4.02             | +  |
| 1        | 10,000     | 4                | +  |
| 0        | 17,000     | 4.23             | +  |
| 0        | 5,400      | 3.73             | +  |
| 1        | 7,000      | 3.84             | +  |
| 1        | 9,400      | 3.97             | +  |
| 0        | 32,000     | 4.51             | +  |
| 0        | 35,000     | 4.54             | +  |
| 0        | 52,000     | 4.71             | +  |
| 0        | 100,000    | 5                | +  |
| 0        | 100,000    | 5                | +  |
| 1        | 100,000    | 5                | +  |
| 1        | 4,400      | 3.63             | -  |
| 1        | 3,000      | 3.48             | -  |
| 0        | 4,000      | 3.6              | -  |
| 0        | 1,500      | 3.18             | -  |
| 0        | 9,000      | 3.95             | -  |
| 0        | 5,300      | 3.72             | -  |
| 0        | 10,000     | 4                | -  |
| 0        | 19,000     | 4.28             | -  |
| 0        | 27,000     | 4.43             | -  |
| 0        | 28,000     | 4.45             | -  |

|   |         |      |   |
|---|---------|------|---|
| 0 | 31,000  | 4.49 | — |
| 0 | 26,000  | 4.41 | — |
| 0 | 21,000  | 4.32 | — |
| 0 | 79,000  | 4.89 | — |
| 0 | 100,000 | 5    | — |
| 0 | 100,000 | 5    | — |

## Appendix D - Beetle Data Set

| Concentration | Killed | Exposed |
|---------------|--------|---------|
| 1.6907        | 6      | 59      |
| 1.7242        | 13     | 60      |
| 1.7552        | 18     | 62      |
| 1.7842        | 28     | 56      |
| 1.8113        | 52     | 63      |
| 1.8369        | 53     | 59      |
| 1.861         | 61     | 62      |
| 1.8839        | 60     | 60      |



## Appendix E - Warsaw Girl Data

| Age   | Number of Menstruating | Number of Interviewed |
|-------|------------------------|-----------------------|
| 9.21  | 0                      | 376                   |
| 10.21 | 0                      | 200                   |
| 10.58 | 0                      | 93                    |
| 10.83 | 2                      | 120                   |
| 11.08 | 2                      | 90                    |
| 11.33 | 5                      | 88                    |
| 11.58 | 10                     | 105                   |
| 11.83 | 17                     | 111                   |
| 12.08 | 16                     | 100                   |
| 12.33 | 29                     | 93                    |
| 12.58 | 39                     | 100                   |
| 12.83 | 51                     | 108                   |
| 13.08 | 47                     | 99                    |
| 13.33 | 67                     | 106                   |
| 13.58 | 81                     | 105                   |
| 13.83 | 88                     | 117                   |
| 14.08 | 79                     | 98                    |
| 14.33 | 90                     | 97                    |
| 14.58 | 113                    | 120                   |
| 14.83 | 95                     | 102                   |
| 15.08 | 117                    | 122                   |
| 15.33 | 107                    | 111                   |
| 15.58 | 92                     | 94                    |
| 15.83 | 112                    | 114                   |
| 17.58 | 1049                   | 1049                  |

## Appendix F - Data set For Illustrative Example 3

| year | Gender | agree | disagree | year | Gender | agree | disagree |
|------|--------|-------|----------|------|--------|-------|----------|
| 0    | 1      | 4     | 2        | 0    | 2      | 4     | 2        |
| 1    | 1      | 2     | 0        | 1    | 2      | 1     | 0        |
| 2    | 1      | 4     | 0        | 2    | 2      | 0     | 0        |
| 3    | 1      | 6     | 3        | 3    | 2      | 6     | 1        |
| 4    | 1      | 5     | 5        | 4    | 2      | 10    | 0        |
| 5    | 1      | 13    | 7        | 5    | 2      | 14    | 7        |
| 6    | 1      | 25    | 9        | 6    | 2      | 17    | 5        |
| 7    | 1      | 27    | 15       | 7    | 2      | 26    | 16       |
| 8    | 1      | 75    | 49       | 8    | 2      | 91    | 36       |
| 9    | 1      | 29    | 29       | 9    | 2      | 30    | 35       |
| 10   | 1      | 32    | 45       | 10   | 2      | 55    | 67       |
| 11   | 1      | 36    | 59       | 11   | 2      | 50    | 62       |
| 12   | 1      | 115   | 245      | 12   | 2      | 190   | 403      |
| 13   | 1      | 31    | 70       | 13   | 2      | 17    | 92       |
| 14   | 1      | 28    | 79       | 14   | 2      | 18    | 81       |
| 15   | 1      | 9     | 23       | 15   | 2      | 7     | 34       |
| 16   | 1      | 15    | 110      | 16   | 2      | 13    | 115      |
| 17   | 1      | 3     | 29       | 17   | 2      | 3     | 28       |
| 18   | 1      | 1     | 28       | 18   | 2      | 0     | 21       |
| 19   | 1      | 2     | 13       | 19   | 2      | 1     | 2        |
| 20   | 1      | 3     | 20       | 20   | 2      | 2     | 4        |

Note: gender "1" denote the male and "2" denote the female.

## Appendix G - Data set for Illustrative Example 4

| Income(Millions of Lira) | Number of Cases | Credit Cards |
|--------------------------|-----------------|--------------|
| 24                       | 1               | 0            |
| 27                       | 1               | 0            |
| 28                       | 5               | 2            |
| 29                       | 3               | 0            |
| 30                       | 9               | 1            |
| 31                       | 5               | 1            |
| 32                       | 8               | 0            |
| 33                       | 1               | 0            |
| 34                       | 7               | 1            |
| 35                       | 1               | 1            |
| 38                       | 3               | 1            |
| 39                       | 2               | 0            |
| 40                       | 5               | 0            |
| 41                       | 2               | 0            |
| 42                       | 2               | 0            |
| 45                       | 1               | 1            |
| 48                       | 1               | 0            |
| 49                       | 1               | 0            |
| 50                       | 10              | 2            |
| 52                       | 1               | 0            |
| 59                       | 1               | 0            |
| 60                       | 5               | 2            |
| 65                       | 6               | 6            |
| 68                       | 3               | 3            |
| 70                       | 5               | 3            |
| 79                       | 1               | 0            |
| 80                       | 1               | 0            |
| 84                       | 1               | 0            |
| 94                       | 1               | 0            |
| 120                      | 6               | 6            |
| 130                      | 1               | 1            |

## Appendix H - SAS Program Code for Simulation Study

\*\*\*\*\*

The following SAS code was used in the first simulation study and it was used to detect the Type I error rate at the level of significance 0.05 with sample size 200. The SAS code for other cases in the first simulation study and the second, third, and fourth simulation study are omitted due to similarities.

\*\*\*\*\*/

```
options nodate pageno=1 linesize=80 pagesize=60;
options nonotes nosource;                /*output format*/

/* define macro LOFtest*/

%macro LOFtest (howmany, quantile, condition);
%do iter = 1 %to &howmany;

data set1;                                /* create covariates*/
do i =1 to 200;
beta0=-0.3; beta1=1.3;beta2=0;
x1=-2+4*ranuni(&iter+100);
x2=x1*x1;
output; end; run;

data set2; set set1;                      /*create the true
                                           probability*/
p=(exp(beta0+beta1*x1+beta2*x2))/(1+exp(beta0+beta1*x1+beta
2*x2));
do i =1 to 200;
yy=ranuni(&iter+100);
end; output; run;

data set3 ;set set2;                      /*create the response
                                           variable*/

if p>=yy then y=1;
if p<yy then y=0;
keep y x1 ;run;
```

```

proc logistic data=set3 desc; /*fit the assumed model*/
model y=x1/scale=none aggregate lackfit;
output out=pha p=phat ;
ods output FitStatistics=rLL2;
ods output GlobalTests=dfr;
ods output LackFitChiSq=HL;
ods output goodnessoffit= pearchi;
ods listing exclude all;
run;

data HL1; set HL; /*hosmer-lemeshow's c
hat method*/

rename chisq= statistic probchisq=pvalue;run;
data HL2; set HL1;
cri=cinv(&quantile, df);
if &condition then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
keep statistic df pvalue reject;
run;
proc append base=HLall data=HL2 force; run;

data hlhat;set pha; /*hosmer-lemeshow h hat method*/
qhat=1-phat;
if (phat ge 0.0 and phat le 0.1) then decile='D01';
if (phat ge 0.1 and phat le 0.2) then decile='D02';
if (phat ge 0.2 and phat le 0.3) then decile='D03';
if (phat ge 0.3 and phat le 0.4) then decile='D04';
if (phat ge 0.4 and phat le 0.5) then decile='D05';
if (phat ge 0.5 and phat le 0.6) then decile='D06';
if (phat ge 0.6 and phat le 0.7) then decile='D07';
if (phat ge 0.7 and phat le 0.8) then decile='D08';
if (phat ge 0.8 and phat le 0.9) then decile='D09';
if (phat ge 0.9 and phat le 1.0) then decile='D10';
run;

proc sort data=hlhat; by decile; run;

proc summary data=hlhat;
by decile; var y phat qhat;
output out=hlhat1 sum=ysm phatsum qhatsum n;
run;

data hlhat2; set hlhat1;
yysum=_freq_ysm;

```

```

o1=(ysm-phatsum)**2/phatsum;
o2=(yysm-qhatsum)**2/qhatsum; run;
proc summary data=hlhat2;
var o1 o2;
output out=hlhat3 sum= olsum o2sum n;run;

data hlhat4; set hlhat3;
df=_freq_-2;
statistic=olsum+o2sum;
if statistic>cinvt(&quantile, df) then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
run;
data HLhat5; set hlhat4;
keep statistic df pvalue reject;run;

proc append base=Hhatal1 data=HLhat5 force; run;

data Deviance; set pearchi; /*Deviance test*/
rename chisq=statistic;
run;
data definal; set deviance;
_ obs _ = _ n _ ;
if _ obs _ ='2' then delete;
cri=cinvt(&quantile, df);
if &condition then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
keep statistic df pvalue reject; run;
proc append base=Deall data=definal force; run;

data pefinal; set deviance; /*pearson chisquare*/
_ obs _ = _ n _ ;
if _ obs _ ='1' then delete;
cri=cinvt(&quantile, df);
if &condition then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
keep statistic df pvalue reject; run;
proc append base=PEall data=pefinal force; run;

data pearchis; set pearchi; /*OS method*/
_ obs _ = _ n _ ;
if _ obs _ ='1' then delete;

```

```

keep chisq df; run;
run;

data os1; set pha;
v=phat*(1-phat);
c=(1-2*phat)/v; run;
proc reg data=os1;
model c=x1;
output out=os2 r=res;
run;
data os3; set os2;
rss=v*res**2; run;
proc means data=os3;
var rss; output out=os4 mean=rssi n=m; run;
data os5; set os4;
rsstot=rssi*m; run;
data os6; merge os5 pearchis;run;
data os7; set os6;
z=(chisq-df)/sqrt(rsstot);
statistic=abs(z);
df=0;
pvalue=2*(1-probnorm(statistic));
if pvalue< 1-&quantile then reject=1;
else reject=0;
keep statistic df pvalue reject ;
run;
proc append base=OSall data=OS7 force; run;

data reduce; /*stukel method*/
merge rLL2 dfr;
keep criterion interceptandcovariates df;
rename interceptandcovariates=r2ll df=rdf ;
run;
data reduce1; set reduce;
obs=_n_;
if obs<3 then delete;run;

data phat ; set pha;
g=log(phat/(1-phat));
if phat>=0.5 then z1=0.5*g*g;
else z1=0;
if phat<0.5 then z2=-0.5*g*g;
else z2=0;
run;

```

```

proc logistic data=phat desc;
model y=x1 z1 z2 /scale=none aggregate;
ods output FitStatistics=fLL2s;
ods output GlobalTests=dffs;
ods listing exclude all;
run;

data fulls;
merge fLL2s dffs;
keep criterion interceptandcovariates df;
rename interceptandcovariates=f211 df=fdf ;
run;

data full1s;set fulls;
obs=_n_;
if obs<3 then delete; run;
data finalST; merge reduce1 full1s;
statistic=r211-f211; df=2;
cri=cinv(&quantile, df);
if &condition then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
keep statistic df pvalue reject;
run;

proc append base=STall data=finalST force; run;

proc summary data=pha; /*proposed method*/
var phat; output out=medout median=mdphat;run;
proc print data=medout; run;

data medp; set medout;
mdphat=mdphat;
do i=1 to 200;
output; end;run;

data medpall; merge medp pha;run;

data ready; set medpall;
if phat<=mdphat then g1x1=x1;
else g1x1=0;
if phat<=mdphat then grp1=1;
else grp1=0;

if phat> mdphat then g2x1=x1 ;
else g2x1=0;

```



```

if phat>mdphat then grp2=1;
else grp2=0;run;

proc logistic data=ready desc;          /*the proposed
                                         method*/
model y=grp1 grp2 g1x1 g2x1 / noint scale=none aggregate
lackfit;
ods output FitStatistics=fLL2;
ods output GlobalTests=df;
ods listing exclude all;
run;

data full;
merge fLL2 df;
keep criterion withcovariates df;
rename withcovariates=f2ll df=fdf ;
run;

data full1;set full;
obs=_n_;
if obs<3 then delete; run;
data finalYL; merge reduce1 full1;
statistic=r2ll-f2ll; df=fdf-rdf-1;
cri=cinv(&quantile, df);
if &condition then reject=1;
else reject=0;
pvalue=1-probchi(statistic, df);
keep statistic df pvalue reject;
run;

proc append base=YLall data=finalYL force; run;

%end; %mend LOFtest;          /* end macro LOG test*/

%LOFtest (1000, 0.95, statistic>cri); /* test lack of
fit*/

```

```

/*another macro to find the reject rate*/

%macro prob (inputname, outputname);
proc means data=&inputname noprint;
var statistic df pvalue reject;
output out =&outputname mean=N=&inputname std=ss1 dd1 pp1
rr1; run;

proc print data=&outputname; run;
%mend prob;

%prob (HLall, out1)          /* call macro prob for results */
%prob(Hhatall, out2)
%prob (Deall, out3)
%prob(Peall, out4)
%prob (OSall, out5)
%prob (Stall, out6)
%prob (YLall, out7)

```