Topics in Goodness-of-fit Test for Logistic Regression Models with Continuous Covariates

By

Pengcheng Lu

Submitted to the graduate degree program in Biostatistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Committee members

Jonathan D. Mahnken, Chairperson

Byron J. Gajewski

Jianghua (Wendy) He

John Keighley

Won S. Choi

Date defended: April 23, 2019

The Dissertation Committee for Pengcheng Lu certifies
that this is the approved version of the following dissertation :

Topics in Goodness-of-fit Test for Logistic Regression Models with Continuous Covariates

Jonathan D. Mahnken, Chairperson

Date approved: May 14, 2019

Abstract

There is no phenomenal method practitioners can use as an appropriate tool for model validation when sparse data are presented in multiple logistic regression models. The characteristics of sparsity, i.e. very few number of observations falling in either grouped or individual covariate patterns, will invalidate the asymptotic chi-square distribution which requires large expected frequencies in each group or bin. Among those tests, the Hosmer-Lemeshow (HL) is the most well-known and widely used as the standard test in assessing logistic regression models since its being introduced. However the inefficiencies of Hosmer-Lemeshow test has been pointed out for years, there is no dominant alternative one emerged yet by far, and the research in assessing logistic regression model fit when sparse data are presented is still very active. Two common methods among a few other proposed methods, namely Copas's unweighted residual sum of squares (RSS) and Su and Wei's & Lin's cumulative sums of residuals (CUMSUM), perform seemly better than the HL in some scenarios, however the limitation of those studies are obvious when those alternatives were introduced: (1) the sample size of the simulation is small (up to 500 observations), (2) the design matrix is relatively simple (usually one continuous and one categorical predictor variables), (3) the number of scenarios considered in their studies are limited, (4) the simulation setups are quite subjective. Due to these reasons, there is no well-established guideline on model validation available for statistical practitioners' daily use when fitting a multiple logistic regression model to sparse data. A common approach is suggested to check model validation by investigating all those

existing goodness-of-fit tests to see if they provide similar evidence of lack-of-fit. Therefore, it is crucial to assess the performance of each method through a comprehensive comparative study. We designed the comparison differently in at least four directions as we mentioned above: varied and expanded sample size, relatively complicated design matrix, more scenarios including adding (over-fitting) continuous/categorical predictor variables and omitting (under-fitting) main effect and /or interaction terms, and a more flexible or robust simulation setting in terms of many randomly sampled models rather than very few pre-specified models were investigated. Furthermore, we proposed a goodness-of-fit test by introducing a new method to partition the fitted values based on the commonly known rules for the limiting distribution of chi-square type statistics for grouped data, which to some extend would overcome the disadvantage of the HL test when the expected counts in some bins are small (usually the cut-off is set as less than five). We also conducted the comparative study by including our proposed method. We summarized the varied goodness-of-fit results in terms of empirical level of significance and power and offered recommendations based on our more generalized simulation studies.

Acknowledgements

I would like to express my deepest gratitude to my academic advisor, Dr. Jonathan D. Mahnken. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own, meanwhile provided me sharp criticism and valuable guidance to recover when my steps faltered. For a week long I was enjoying our research conversations after I finished our weekly meeting and walked out of his office during my dissertation research career. His brilliant ideas and great advice consistently encouraged and helped me walk through the last stage of my study at the University of Kansas Medical Center (KUMC). It is impossible for me to finish up my dissertation without Dr. Mahnken's input.

A big THANK YOU to my other advisors, Dr. Byron J. Gajewski, Dr. Jianghua (Wendy) He, Dr. John Keighley and Dr. Won S. Choi, for offering me insights that guided and challenged my thinking and advice that improved my research. I am deeply grateful to all of them for consistently checking my dissertation progress.

I am very grateful to Dr. Matthew S. Mayo and Dr. Jo A. Wick for not only offering me the opportunity to join the Graduate Education Program in the Department of Biostatistics & Data Science but also for their continued support and mentoring me throughout my graduate study at KUMC.

It has been an honor and pleasure to work as a GRA under the supervision of Dr. Brooke L. Fridley and Dr. Jonathan D. Mahnken and to work as a GTA under Dr. Devin C. Koestler's supervision. I am so thankful to them all for mentoring me through my academic services at KUMC. I am also grateful to

Dr. Koestler for offering me the high performance computing facility to meet my intensive computation need, my simulation processing would take much longer time without his support. I would also like to thank three great IT professionals, Dinesh Pal Mudaranthakam in the Department of Biostatistics & Data Science, Riley Epperson and Bradley Fleming in the Center for Research Computing at the University of Kansas, for their timely help and support.

I am thankful for getting to know such an amazing group of people mixed with exceptional faculty, staff, and graduate students in the Department of Biostatistics & Data Science at KUMC. It has been such a great experience for me to be a part of this big family, to get helped and blessed.

Finally I would like to thank my wife, my son and my daughter, my parents and my mother-in-law, it would not be possible for me to push my academic destination this far without their love and patience. I especially want to thank my wife for her love and support throughout my graduate study, it is impossible for me to pursue my Ph.D. degree without her understanding and encouragement in both words and actions even prior to my graduate career at the University of Kansas School of Medicine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Generalized linear models

The generalized linear model (GLM) generalizes ordinary linear regression by allowing for response variables that have distributions in the exponential family (including simply normal distributions), and for an function of the response variable (the link function) to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly).

The class of generalized linear models consists of three common components:

1. Random component: the probability distribution of the response variables $Y_1, \ldots, Y_N$, or $\boldsymbol{Y}$, are assumed to be independent and to share the same distribution from the exponential family, having means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$.

2. Systematic component: specifies the explanatory variables $(X_1, X_2, \ldots, X_p)$ in the model, more specifically their linear combination in creating the linear predictor, $\boldsymbol{\eta} = \boldsymbol{X\beta}$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)^T$, $\boldsymbol{\beta}$ is the vector of unknown parameters, and $\boldsymbol{X}$ is the design matrix.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \\ \vdots \\ \boldsymbol{X}_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & \ldots & \vdots \\ x_{N1} & \ldots & x_{Np} \end{pmatrix}$$

3. Link Function, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$, such that $E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X\beta})$, is monotonic

and differentiable, which specifies the relationship or link between the random and the systematic components.

The outcomes modeled by generalized linear models are assumed to have distributions in the exponential family. The generalized linear models were formulated by Nelder and Wedderburn (1972)[1] as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. Extensive treatment of generalized linear models can be found in McCullagh and Nelder (1989) [2], Dobson and Barnett (2008) [3], and Agresti (2012) [4].

A distribution falls into the exponential family if its distribution function of the outcome random variable, Y, can be written in the form

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} \tag{1.1}$$

where θ is the canonical parameter, which depends on the expected value of y, ϕ is a scale parameter for dispersion, and $b(\theta)$ is the cumulant function, and $c(y,\phi)$ is a normalizing term. When the chosen link function of the GLM is the same function as the canonical parameter θ, then $\theta = \mu$ and the link function is referred to as the canonical link.

The joint density function of an exponential family distribution for a set of outcomes y, is given by

$$f(\boldsymbol{y}|\theta,\phi) = \prod_{i=1}^{N} exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right\} \tag{1.2}$$

Assuming that the observations y_i's are i.i.d., the joint log likelihood for members of the exponential family can be expressed as

$$l = \sum_{i=1}^{N}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right) \tag{1.3}$$

1.2 Logistic regression model in the framework of GLMs

Logistic regression is just similar to multiple linear regression, which is considered as part of GLMs, with the exception that the response variable is binomial (each individual has two outcomes, for example, success or failure, event or non-event, usually numerically coded as 1 or 0, in this case the response variable can be referred to Bernoulli random variable ($n = 1$ in binomial distribution)). A binomial distribution is also be used to describe aggregated data, in case of $n > 1$, for example when the predictor variable is categorical. Logistic regression or a logit model is commonly used to fit binomial data and to explain the relationship between one dependent binomial variable and one or more nominal, ordinal, interval, ratio-level independent variables, or a mixture of them.

Logistic regression is widely used in analyzing binary responses in biomedical research field for many reasons such as ease of interpretation of parameters, possibility of calculating prognoses for the event of interest, and availability of standard software. It has been introduced by many textbooks (Hosmer and Lemeshow, 1989 [5]; Neter et al, 1990 [6]; Glantz and Slinker, 1990 [7]; Montgomery and Peck, 1992 [8]; Woolson and Clarke, 2002 [9]).

At the center of the logistic regression analysis is the task estimating the log odds of success or an event. In logistic regression model, the dependent variable Y_i for the i^{th} observation is often coded as 1 with a probability of success π_i or 0 with a probability of failure $1-\pi_i$, whereas the independent variables $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{ip})$ can take any data type and the relationship between the predictors and the outcome variable is linear through the

logit link function of π_i:

$$\begin{aligned}\eta_i &= g(\pi_i) \\ &= logit(\pi_i) \\ &= ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ &= \boldsymbol{X_i}^T \boldsymbol{\beta}\end{aligned} \tag{1.4}$$

or equivalently,

$$\begin{aligned}\pi_i &\equiv \pi_i(\boldsymbol{\beta}, \boldsymbol{X_i}) \\ &= g^{-1}(\eta_i) \\ &= E[Y_i|\boldsymbol{\beta}, \boldsymbol{X_i}] \\ &= \frac{exp\left\{\boldsymbol{X_i}^T \boldsymbol{\beta}\right\}}{1 + exp\left\{\boldsymbol{X_i}^T \boldsymbol{\beta}\right\}}\end{aligned} \tag{1.5}$$

where $\boldsymbol{X_i} = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ denote the $(p+1) \times 1$ covariate vector for the i^{th} individual in the sample, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$ is the $(p+1) \times 1$ vector of regression parameter, i.e., an intercept and p explanatory variables.

It is straight forward to show that the binomial distribution is an exponential family member. The probability mass function of the binomial distribution is given by

$$f(y|n, \pi) = \binom{n}{x} \pi^y (1-\pi)^{n-y}, \tag{1.6}$$

and it can be rewritten in the form of exponential function as below:

$$f(y|n, \pi) = exp\left\{yln\left(\frac{\pi}{1-\pi}\right) + nln(1-\pi) + ln\binom{n}{x}\right\} \tag{1.7}$$

with the canonical parameter $\theta = ln\left(\frac{\pi}{1-\pi}\right)$, the cumulant function $b(\theta) = -nln(1-\pi)$, the dispersion parameter $\phi = 1$, the normalizing term $c(y, \phi) = 0$. When $n = 1$, it becomes the special case of the binomial distribution, i.e. the Bernoulli distribution.

Two important 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 explanatory variables.

Therefore the outcome variable should be dichotomous in nature such as presence/absence or success/failure, and there should be no high correlations (multicollinearity) among the predictors, this can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013)[10] suggest that as long as the correlation coefficients among independent variables are less than 0.90 the assumption is met.

1.3 The goodness-of-fit test for logistic regression model

The goal of any statistical model processing is to obtain a mathematical model that describes the relationship between observations of the outcome or dependent variable and a collection of independent variables. Hosmer et al. (1991) [11] describe a convenient way to conceptualize the model is to think of the value of the dependent variable as being composed of two parts: (1) the systematic component and (2) the error component. Then, the process of "building" the regression model concentrates on the systematic component, attention turns to the error component after completion of the systematic component, and the process of examining the values of the error component is the so-called assessing the goodness-of-fit of the model. Therefore the goodness-of-fit of a statistical model describes how well it fits a set of observations. Measures of goodness-of-fit typically summarize the discrepancy between observed values and the values expected under the model in question.

Generally a goodness-of-fit statistic tests the following pair of hypotheses:

H_0: the model M_0 exhibits no lack-of-fit to the observed data.

vs.

H_A: the model M_0 exhibits some lack-of-fit to the observed data, should seek alternative model M_A for a better fit.

Most often the observed data represent the fit of the saturated model, the most complex model possible with the given data. Thus, most often the alternative hypothesis H_A will represent the saturated model M_A which fits perfectly because each observation has a separate parameter.

The assessment of model fit is a very important component in any model processing, and this once difficult task of using goodness-of-fit test to assess the adequacy of fitted logistic regression model has become a routine step in model building process after a lot of goodness-of-fit tests being introduced and implemented to statistical software packages. Nowadays any analysis should incorporate a thorough examination of logistic regression diagnostics before reaching a final decision on model adequacy.

There are a number of ways in which a fitted model can be inadequate. For example, the linear systemic component of the model may be incorrectly specified; that is, important covariates that should be included in the model may be omitted. or the function form of the liner predictor may be inappropriate. The transformation of the response probability may not have the desired relationship with the liner predictor: for instance, it may be more appropriate to relate the linear predictor to the complementary log-log of the response probability rather than to the logit of that probability. The assumption that the observed response data come from a particular probability distribution may be wrong. Goodness-of-fit tests try to evaluate how well model-based predicted outcomes coincide with the observed data.

Similarly as when fitting a liner regression model, fitting a logistic regression model requires one to identify potential covariates that will be included in the model. Estimation of the coefficients is usually achieved utilizing the maximum likelihood method (Bickel and Doksum, 1977[12]; Edwards, 1992[13]), followed by formulating and testing a hypothesis to

assess the significance of the covariates in the fitted model.

Maximizing the likelihood of regular exponential family for a linear model (e.g. linear or logistic regression) is equivalent to obtaining solutions to their score equations.

$$0 = \sum_{i=1}^{n} S_i(\alpha, \beta) = \frac{\partial}{\partial \beta} logL(\beta, \alpha, X, Y) = X^T(Y - g^{-1}(X\beta)). \tag{1.8}$$

Where Y_i has expected value $g^{-1}(\eta) = g^{-1}(X_i\beta)$. In GLM estimation, $g^{-1}()$ is the inverse of a link function. A generalized estimating equation approach would specify linear models in the following way:

$$0 = \frac{\partial}{\partial \beta} \boldsymbol{V}^{-1}(\boldsymbol{Y} - g^{-1}(\boldsymbol{X}\boldsymbol{\beta})) \tag{1.9}$$

With $\boldsymbol{V}$ a matrix of variances based on the fitted value (mean) given by $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$. In logistic regression $g()$ would be the logit link function, and V_{ii} would be given by $\pi_i(1-\pi_i) = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta})(1-g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}))$. The solutions to this estimating equation, obtained by Newton-Raphson, will yield the $\hat{\boldsymbol{\beta}}$ for the logistic regression.

There are many asymptotically equivalent methods for performing the significance test on the estimated parameter $\hat{\beta}_i$, among which the likelihood ratio test, Wald test and the score test are the most popular (Cox and Snell, 1989[14]).

One should distinguish the significance test for each coefficient from the goodness-of-fit test. The test of $\boldsymbol{\beta}$ parameters provides information about covariate significance in the model, relative to overall variability seen in the dependent variable, whereas the latter provides evidence to answer the question of whether the predicted values are an accurate representation of the observed value.

Both significance test and the goodness-of-fit test should also be distinguished from model selection procedure where different fitted models that are provided or explored by inverstigators are compared. While model selection procedure selects one most fitted model among a group of potential models, the goodness-of-fit tests answer to the question: does this model fit? In practice both of these methods should be used for model checking and model se-

lection, because (1) even the best fitted model among the potential models may not fit the data well; (2) even the goodness-of-fit test concludes that no evidence of the specified model lacking of fit, there may be better models available for exploration.

For example, in general, adding independent variables to a logistic regression model will increase the amount of variance explained in the log odds, typically denoted by pseudo R^2. Hemmert et al. confirm that an increasing number of independent variables increases the values of all type of pseudo-R^2 [15], and Allison (2014, [16]) addresses that it is possible that adding a variable to the model could reduce the Tjur R^2, note this type of R^2 is recommended by the author but it is not included in Hemmert's study. However, adding more variables to the model can result in overfitting, which would reduce the generalizability of the assumed model. If one claims a good model purely based on the pseudo-R^2, the risks are over-fitting issues and also the model may not be a good fit to data. In addition, since pseudo-R^2 is not a statistical test, how large is the pseudo-R^2 must be to ensure a good fit is not easy to determine (Cox and Wermuth, 1992 [17]).

Furthermore, it is worth noting that Hosmer and Lemeshow [5] argued that R^2-type measures are based on various comparisons of the predicted values from the fitted model to those from the base model (intercept only model) and, as a result, do not assess goodness-of-fit, and a true measure of fit should be one based on a comparison of observed to predicted values from the fitted model. They do not recommend routine publishing of pseudo-R^2 values with results from fitted logistic models.

Unfortunately, pseudo-R^2 serving as a versatile goodness-of-fit indicator for logit models can be widely founded in empirical researches. Hoetker (2007, [18]) asserts that almost all papers published after 2000 reporting pseudo-R^2 as a measure of model fit.

A better approach is to present any of the goodness-of-fit tests available, and the measures of goodness of fit is strictly for summarizing the discrepancy between observed values and the values expected under the model in question. Classical goodness-of-fit tests are readily available for logistic regression when the data can be aggregated or grouped into unique

"profiles". Profiles are groups of cases that have exactly the same values on the predictors. For example, based on UCLA tutorial (Example 2, [19]), suppose we are interested in how the factors, GPA and the tier of prestige of the undergraduate institution, affect admission into graduate school. The response variable ADMIT, is a binary variable with values admit/don't admit. The predictor variable GPA is a dummy variable, i.e. 1/0 for undergraduate GPA greater than 3.5/not greater than 3.5, while the other predictor variable TIER takes on the values 1 through 4. Institutions with a TIER of 1 have the highest prestige, while those with a TIER of 4 have the lowest. There are then eight profiles, corresponding to the eight cells in the two-way cross-classification of GPA by TIER. After fitting the model, we can get an observed number of success (admit) and an expected number of success (admit) for each profile. There are two well-known statistics for comparing the observed number with the expected number: the deviance and Pearson's chi-square.

The widely used chi-square statistic was introduced by Pearson in 1900 [20] and its theory and applications were subsequently expended by Fisher, Yates and others (Agresti, 1990[21]). The following calculation of a statistic X^2 called a "chi-square statistic" is used as a measure of how far observed sample data deviate from a theoretical model.

Suppose the observed data are classified into G exhaustive and unique bins, and the counts whithin each bin are tabulated, the chi-square statistic has the form:

$$X^2 = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g} \tag{1.10}$$

Here, O_g is the frequency of observed cases in g^{th} bin, E_g is the frequency of expected cases in g^{th} bin. If the theoretical model is correct and the summation is performed over all G bins of the data, under suitable regularity conditions and under the null hypothesis of no lack of fit, X^2 is asymptotically distributed as central chi-square with $(G - k - 1)$ degrees of freedom, where k is the number of predictors in the model (not counting the intercept). The more discrepant the proposed model is from the truth (i.e. lack of fit of the proposed

model), the larger the absolute difference $(O_g - E_g)$, and the larger the value of X^2.

The deviance statistic stems from an examination of the likelihood ratio test. The likelihood function summarizes the information that the data provide about the unknown parameters in a model of interest. Under the assumption of a particular form of the underlying distribution, the deviance of the model is a measure of the difference between the log-likelihood of the fitted model (LL_f) and the log likelihood of the saturated model that fits the data perfectly (LL_s), The deviance is defined to be

$$D = 2 * [LL_s - LL_f] = 2 \sum_{g=1}^{G} O_g log\left(\frac{O_g}{E_g}\right). \tag{1.11}$$

The deviance statistic has the same form as the likelihood ratio statistic from the logistic regression model. Therefore, the null distribution of the deviance follows directly from a result about likelihood ratio test. According to this result, under suitable regularity conditions, the deviance has approximately a chi-square distribution when the model holds. The degrees of freedom are $J-k-1$, where J is the number of distinct covariate patterns (a term used to describe a single set of values for the covariates in a model) and k is the number of parameters (not counting the intercept) in the model. Smaller deviance indicates better fit.

Pearson's chi-square and the deviance statistics are appropriate for grouped data. Although the deviance and Pearson's chi-square statistics are routinely provided in most statistical packages, one has to be aware that their chi-square limiting null distribution is only valid when the number of observations in each covariate pattern is large. This condition is often unrealistic when there are a large number of categorical covariates or when the continuous covariates are present in the model.

Hosmer and Lemeshow develop a series of goodness-of-fit test statistics to overcome the issue when continuous covariates are present in the model. The Hosmer-Lemeshow statistics (Hosmer and Lemeshow, 1980 [22]; Hosmer and Lemeshow, 1989 [5]) are practical goodness-of-fit measures for general situations, including those with continuous predictors.

The procedure is performed as follows. First,the probability of the event for each subject is calculated based on the fitted model and the estimated probabilities are sorted in ascending order. Second, the predictions are grouped into G bins according to the percentiles of the estimated probabilities. Finally, the Hosmer-Lemeshow goodness-of-fit statistic is calculated using the model's average predicted value in each bin. Either fixed cut points or data-driven cut points (to achieve equally-sized bins) can be used, and the statistic is denoted by $\hat{H}$, $\hat{C}$ respectively.

Through intensive simulation studies, Hosmer and Lemeshow [22] and Hosmer et al. [23] showed that when the logistic regression model is correct and the estimated expected values are "large" in all bins, the distribution of both $\hat{H}$ and $\hat{C}$ with G groups or bins is well approximated by a chi-square distribution with $G-2$ degrees of freedom.

The Hosmer-Lemeshow test has become the standard test for assessing goodness-of-fit in logistic regression and is implementated in all majoy statistical packages. Its popularity is due to its properties: (1) it is intuitively appearling and easy to compute; (2) it has sound support from simulation studies; (3) it is widely available in computer packages. Additionally lack of a better approach also contributes to its popularity. In spite of being used widely for goodness-of-fit measure in binary data situations, however the Hosmer-Lemeshow statistic has substantial deficiencies. It has been shown by many researchers that the test statistic is sensitive to the choice of grouping results. A sufficient demonstration could be found in Allison (2014) [16].

The RSS (residual sum of squares) test introduced by Copas (1989)[24] considers only the numerator of the Pearson's Statistics, where the summation is again over the individual observations as follows:

$$RSS = \sum_{i=1}^{M} (y_i - \hat{\pi}_i)^2. \tag{1.12}$$

Hosmer et al. (1997) [23] show how to calculate asymptotic moments of RSS and to perform a statistical test. Copas's RSS test is also a special case of the class of statistics considered by le Cessie and van Houwelingen [25].

Osius and Rojek [26] derived asymptotic moments for a general class of goodness-of-fit statistics in logistic regression.This class, the so-called "power-divergence family" of Cressie and Read [27], incorporates X^2 and D, however, moments in closed form can only be calculated for X^2, A statistical test can be performed by standardizing X^2 with tese moments and computing the resulting test statistic (X_o^2) to the standard normal distribution.

Royston (1992 [28], 1993 [29]) proposed two procedures designed to detect departure from linearity in the logit, that is the tests are designed to be sensitive to departures in monotonicity in the logit or to detect a quadratic logit, that use partial sums of residuals. Royston did not specifically advocate the use of these tests for overall assessment of goodness-of-fit. In the case of a single covariate the Royston monotone test is identical to a test proposed by Su and Wei (1991) [30]. Su and Wei proposed using a computationally intensive simulation to calculate the p-value. The computations for model based on a sample size n containing p main effect terms for continuous covariates are of order n^pR, where R is the number of simulations performed. The accuracy of the estimated p-values is a function of R. For example, 500 simulations are needed to estimate significance at the 5% level to within 2% with 95% confidence. In preliminary simulations the performance of Su and Wei method of obtaining the p-value was superior to Royston's analytic approach for models containing a single covariate.

Note how the prescribed tests differ in their constructing principles. To our knowledge, up to now there have only been three major simulation studies designed to investigate the global goodness-of-fit tests in logistic regression when continuous covariates are presented, namely Hosmer et al. (1997) [23], Hosmer and Hjort (2002) [31] and Kuss (2002) [32], and they shared same limitation in two aspects, i.e. small sample size (N =100 and/or 500) and simple design matrix.

In the present study we will include the Hosmer-Lemeshow, Copas's RSS and Su and Wei and Lin's cumulative sum of residuals methods and investigate their performance under varied sample size with expanded number of parameters in the design matrix. Further, we

propose a new partitioning method to group the observations based on their fitted values for logistic regression model when continuous covariates are presented. The proposed methodology uses a straight forward partition strategy in the dependent variable space to avoid small of estimated expected values within each bin. Specifically we use data-driven method to partition groups with large enough sample size under the rule that every bin should contain estimated expected values exceeding the benchmark value of 5. The new proposal attempts to overcome some of the deficiencies demonstrated in Hosmer-Lemeshow test, yet still maintain its applicability for a wide range of covariate configurations (discrete, continuous or a mixture of them).

This thesis is organized as follows: In chapter two, an extensive literature review is presented. Chapter three describes the proposal of a new partitioning method for goodness-of-fit test in logistic regression modeling. Chapter four presents comparisons of empirical size and power of lack-of-fit of the proposed test and three alternative tests under similar settings of three published studies. In chapter five, we further compare test size and power of four tests under a more generalizable set of simulation experiments. Detailed simulation plans are deployed and simulation results are analyzed. Chapter six addresses our considerations about computation through out this study. Chapter seven offers overall summary and discussion, future work directions are included as well.

Chapter 2

Literature Review

When all the predictions in a logistic regression model are categorical and the number of covariate patterns is small relative to the sample size, the Pearson chi-square statistic and deviance statistic are appropriate to use for assessing goodness-of-fit of a logistic regression model. However, when the number of covariate patterns is large, as in the case when continuous predictors are present, difficulties arise. Extensive studies have been done on how to assess the logistic regression model fit when sparse data are present. This chapter reviews the literature on this topic. We include some important and popular goodness-of-fit tests for logistic regression and the generalized linear model here, these goodness-of-fit tests are proposed mainly for the overall measure of fit.

2.1 Introduction

In general, there are two different approaches to assessing goodness-of-fit in logistic regression models. The first, known as residual analysis, investigates the model on the level of individuals and looks for those observations which are not adequately described by the model or which are highly influential on the model fit [33]. The second approach seeks to combine the information on the amount of lack-of-fit in a single number. Statistical tests, so-called goodness-of-fit tests, are then calculated to judge if this lack-of-fit is significant or due to random chance.

We can distinguish two types of goodness-of-fit tests: specific and global, or individual and collective. Specific tests embed the logistic model in a wider class of models, say, with

a more general link function, and check if the data at hand can be better described by the enhanced model. If not, we stay with our fitted model. Opposed to this, global tests do not evaluate specific alternatives, rather test unspecific hypotheses of the form "the model fits" versus the alternative "the model does not fit" as we addressed in Section 1.2.

On the one hand, a global test in the case of a bad model fit does not offer any insights on how to improve the model. On the other hand, it is dangerous to expect this from specific tests. In general, a specific test is derived under the assumption of a single isolated mis-specification (for example, mis-specified link function) that is checked in the alternative hypothesis, but is only valid when all other aspects of the model specification are correct [34]. Only in this special case will a rejection of the null hypothesis lead to an indication of how to improve the model. A second disadvantage of specific goodness-of-fit tests is that they require, at least if we consider likelihood ratio or Wald tests, the estimation of the parameters from the enlarged model which in most cases is unfeasible with standard software [5].

Suppose we have N independent observations of the pair $(y_i, \boldsymbol{X_i})$, where y_i is the dichotomous outcome and $\boldsymbol{X_i} = (1, x_{1i}, \ldots, x_{pi})^T$ is the $(p+1) \times 1$ covariate vector, $i = 1, \ldots, N$. Under the logistic regession model we assume that $P(Y_i = 1|\boldsymbol{X_i}) = \pi(\boldsymbol{X_i})$,where $\pi(\boldsymbol{X_i}) = e^{g(\boldsymbol{X_i})}/(1 + e^{g(\boldsymbol{X_i})})$, and $g(\boldsymbol{X_i}) = \boldsymbol{X_i}^T\boldsymbol{\beta}$. Parameter estimates are usually obtained by maximum likelihood and are donated by $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \ldots, \hat{\beta}_p)$. We denote the fitted values as $\hat{\pi}_i = \pi(\boldsymbol{X_i}, \hat{\beta})$, which is the estimated probability of positive response $(y_i = 1)$ for the i^{th} observation. We conclude that a model fits if

- summary measures of the distance between y_i and $\hat{\pi}_i$ is small
- the contribution of each pair $(y_i, \hat{\pi}_i)$ to these summary measures is unsystematic and is small relative to the error structure of the model.

Hosmer et al. (1997) [23] discussed the portfolio of goodness-of-fit process in the way forming three assumptions. They argued that the task as in examining a model's goodness-of-fit are to determine whether the fitted model's residual variation is small, whether the

model displays no systemic tendency and whether it follows the variability postulated by the mode. Evidence of lack-of-fit may come from a violation of one or more of these three characteristics. They further formed three essential components of fit in the context of a logistic regression model as follows:

1. the logit transformation is the correct function linking the covariates with the conditional mean, $logit[\pi(\boldsymbol{X})] = \boldsymbol{X\beta}$

2. the linear predictor,$\boldsymbol{X\beta}$, is correct (we do not need to include or exclude additional variables, transformations of variables, or interaction of variables).

3. the variance is Bernoulli, $var(Y_i|\boldsymbol{X_i}) = \pi(\boldsymbol{X_i})[1 - \pi(\boldsymbol{X_i})]$

The practical problem is that assumptions 1 through 3 are not mutually exclusive, which means they may be confounded with each other. In the case of a logistic regression model, a global goodness-of-fit test is actually checking three assumptions simultaneously.

2.2 Pearson's chi-square and deviance tests revisited

Now we consider the general case of g independent random variable $Y_1, Y_2, \ldots, Y_g$ corresponding to the number of success in g different subgroups or strata, i.e. aggregated data as shown in Table 2.1 and $Y_i \sim Bin(n_i, \pi_i)$, $i = 1, \ldots, g$. Then, by equation (1.7), the log-likelihood function is given by

$$l(\pi_1, \ldots, \pi_g; y_1, \ldots, y_g) = \sum_{i=1}^{g} \left[y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i log(1-\pi_i) + log\binom{n_i}{y_i} \right] \tag{2.1}$$

Under this framework of a $2 \times g$ table, it can be shown that, when X^2 is evaluated at

Table 2.1: Frequencies for g Binomial Distributions

	Subgroups			
	1	2	...	g
Success	Y_1	Y_2	...	Y_g
Failure	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_g - Y_g$
Total	n_1	n_2	...	n_g

the estimated expected frequencies, the Pearson statistic in equation (1.10) becomes

$$\begin{aligned} X^2 &= \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i} \ \ (where \ \ G = 2g) \\ &= \sum_{i=1}^{g} \sum_{j=1}^{2} \frac{(O_i - E_i)^2}{E_i} \ \ (here \ j \ is \ for \ success \ and \ failure \ groups) \\ &= \sum_{i=1}^{g} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \sum_{i=1}^{g} \frac{((n_i - y_i) - n_i(1-\hat{\pi}_i))^2}{n_i(1-\hat{\pi}_i)} \\ &= \sum_{i=1}^{g} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}(1 - \hat{\pi}_i + \hat{\pi}_i) \\ &= \sum_{i=1}^{g} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}. \end{aligned} \tag{2.2}$$

which can be viewed as the weighted residual sum of squares (WRSS) and the (residual) deviance in equation (1.11) becomes

$$D = 2\sum_{i=1}^{g} \left(y_i log\Big(\frac{y_i}{n_i\hat{\pi}_i}\Big) + (n_i - y_i) log\Big(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\Big)\right). \tag{2.3}$$

One can refer to lots of textbooks for the derivative of the above formula, among those say Dobson & Barnett (2008) [3] outlined the steps well. It can be shown these two statistics are asymptotically equivalent (See proof in Appendix A).

More generally, we can change the g subgroups to J covariate patterns to form a $J \times 2$ table as illustrated below without changing the conclusion.

Table 2.2: Data classification by covariate pattern for J Binomial Distributions

Covariate pattern	y=0	y=1	Total
$\boldsymbol{x}_1$	O_{10}	O_{11}	n_1
$\boldsymbol{x}_2$	O_{20}	O_{21}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_J$	O_{J0}	O_{J1}	n_J

Both Pearson statistic and residual deviance rely on the principle of comparing observed Y_i to predicted $n_i\hat{\pi}_i$ values and should be large if the model does not fit the data well. To judge statistical significance they are usually compared to a χ^2 distribution with $g-p-1$ degrees of freedom. The validity of this distribution, however, relies on the assumption of large $n_i\hat{\pi}_i$ and $n_i(1-\hat{\pi}_i)$, both tests show unsatisfactory behavior with sparse data, here sparse data can be sparse group data or continuous covariates. McCullagh et al. (1989) [2] showed that D degenerates to

$$D = 2\sum_{i=1}^{g}\left(\hat{\pi}_i log\Big(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\Big) + log(1-\hat{\pi}_i)\right) \tag{2.4}$$

in the extreme case when every individual observation has its own covariate pattern ($n_i = 1$), Then D is completely independent of the observations and contains absolutely no information about the model fit. The Pearson statistic performs not that much better in this situation, for it can also be shown that $X^2 \approx N$, and the sample size is not a sensible measure of fit [2].

From the view of the contingency table, we know that both of the two tests require that the number of columns in Table 2.1 must be fixed and the sample size should be large enough such that the counts (observed and expected) in each cell all exceed some minimum number, such as five, for the p value to be a valid measure of model fit. As a result, these two test statistics may not be appropriate for continuous covariates. Kuss (2002) [32] confirmed existing knowledge that X^2 and D are not valid goodness-of-fit tests in logistic regression with sparse data through simulation.

2.3 The Hosmer and Lemeshow tests

Hosmer and Lemeshow (1980, 1989 [5], [22]) developed two goodness-of-fit test statistics for binary logistic regression model when the model contains continuous covariates. These statistics are similar to X^2 statistic, by grouping observations. One of two grouping methods for binning group boundary cut points was applied when calculating the statistics. The group boundary cut points are either random as driven by data or fixed as pre-specified. Specifically they proposed two grouping strategies based on the values of the estimated probabilities: (1) collapse the data table based on percentile of the estimated probabilities into a pre-determined number of bins, and (2) collapse the data table based on pre-determined fixed values of the estimated probability. After binning data, the Hosmer-Lemeshow statistic is defined as:

$$\begin{aligned} H &= \sum_{k=1}^{g}\sum_{y=0}^{1}\frac{(O_{yk}-E_{yk})^2}{E_{yk}} \\ &= \sum_{k=1}^{g}\left(\frac{(O_{1k}-E_{1k})^2}{E_{1k}}+\frac{(O_{0k}-E_{0k})^2}{E_{0k}}\right) \\ &= \sum_{k=1}^{g}\left(\frac{(O_{1k}-E_{1k})^2}{n_k^{'}\pi_k}+\frac{(N_k-O_{1k}-(N_k-E_{1k}))^2}{n_k^{'}(1-\pi_k)}\right) \\ &= \sum_{k=1}^{g}\frac{(O_{1k}-E_{1k})^2}{n_k^{'}\pi_k(1-\pi_k)} \end{aligned} \tag{2.5}$$

Here O_{1k}, E_{1k}, O_{0k}, E_{0k}, N_k, and π_k denote the observed $y=1$ events, expected $y=1$ events, observed $y=0$ events, expected $y=0$ events, total observations, predicted risk for the k^{th} risk decile group. And g is the total number of groups, k is the number of collapsed groups or bins, $n_k^{'}$ is the total number of subjects in the k^{th} group, and y is the binary responses.

With the first grouping method, the Hosmer-Lemeshow procedure is as follows:

1. Use the derived model to estimate the probability of the event for each subject.

2. Sort the estimated probabilities in ascending order.

3. Group the data into g bins based on percentiles of the estimated probabilities. For example, if $g = 10$, the k^{th} group would contain the subjects whose estimated probabilities were between the $(k-1)^{th}$ and the k^{th} deciles of the whole estimated probabilities. The Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$, is obtained by calculating the Pearson chi-square statistic from the $g \times 2$ table of observed and estimated expected frequencies.

4. Construct the Hosmer-Lemeshow statistic $\hat{C}$. Let $n_k^{'}$ be the total number of observations in the k^{th} group, $O_k = \sum_{j=1}^{c_k} y_j$ be the total number of observed positive responses among all subjects who fall within the c_k covariate patterns represented in the k^{th} bin, and $\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k^{'}}$ be the average estimated probability of all subjects falling within the k^{th} bin, c_k denotes the number of covariate pattterns in the k^{th} bin and m_j denotes the number of subjetcs of covariate pattern j in the k^{th} bin. Then with similar algebra for H, the Hosmer-Lemeshow statistic is computed as $\hat{C}$ by using the following formula:

$$\hat{C} = \sum_{k=1}^{g} \left(\frac{(O_{1k} - \hat{E}_{1k})^2}{\hat{E}_{1k}} + \frac{(O_{0k} - \hat{E}_{0k})^2}{\hat{E}_{0k}} \right) = \sum_{k=1}^{g} \frac{(O_k - n_k^{'}\bar{\pi}_k)^2}{n_k^{'}\bar{\pi}_k(1 - \bar{\pi}_k)} \tag{2.6}$$

With the second method, use of g groups results in cutpoints defined as the values h/g, where $h = 1, \ldots, g-1$, and the groups contain all subjects with the estimated probabilities between adjacent cutpoints. For example, the first group contains all subjects whose estimated probability is less than or equal to $1/g$, while the g^{th} group contains those subjects whose estimated probability is greater than $(g-1)/g$. The Hosmer-Lemeshow statistic using this second approach is then calculated the same way as is with the first approach and denoted as $\hat{H}$. It is obvious that the second approach does not guarantee balanced numbers of observations in each of the g bins.

The distribution of Hosmer-Lemeshow test statistic is affected by two conditions. The first is that the estimates of the regression parameters are determined using likelihood functions for un-grouped data. The second is that the boundaries for any group are dependent on

the estimated parameters, and thus the groups are random (Hosmer, et al. 1980 [22]). Applying the work of Moore and Spruill (1975, [35]), Hosmer et al. showed that the asymptotic distribution of Hosmer-Lemeshow test statistic is

$$\chi^2(2G - G - (k+1)) + \sum_{k=1}^{K+1} \lambda_k \chi_k^2(1)$$

where K is the number of covariates, and λ_k is the k^{th} non-zero or 1 eigenvalue of the k^{th} covariance matrix of the HL statistic, $0 < \lambda_k < 1$, and $k = 1, \ldots, K$.

If the hypothesized model is correct, in large samples, both $\hat{C}$ and $\hat{H}$ have approximately a chi-square distribution with $g - 2$ degrees of freedom. Hosmer and Lemeshow's extensive simulation results have shown to support this statement. The test is carried out at approximate size α by rejecting the assumed model if $\hat{C}$ or $\hat{H}$ exceeds the $100(1 - \alpha)$ quantile of a chi-square distribution, $\chi^2(g - 2)$.

It's worth noting that some authors (for example, Yu et al. (2017, [36])) address that the $g-2$ degrees of freedom of the Hosmer-Lemeshow test could be proved just according to Theorem 5.1 in Moore and Spruill (1975) [35]. We justify that this may not be a must result as addressed by the authors. We will investigate the degrees of freedom problem related to the Pearson chi-square type statistics with details in chapter three.

Further research by Hosmer, Lemeshow, and Klar (1988) [38] has indicated that the grouping method based on percentiles of the estimated probabilities is preferable to the one based on fixed cut points in the sense of better adherence to the $\chi^2(g - 2)$ distribution, especially when many of the estimated probabilities are small. In fact if the deciles-of-risk partitioning method is used, and the predicted probabilities in a group are either all near 0 or all near 1, the expected frequency of an event (i.e. the sum of the predicted probabilities in that group) or of a non-event may be less than 1. This would invalidate the chi-square approximation for the distribution of HL. Based on these results, they have subsequently recommended grouping method one over method two. In this study we utilize method one,

i.e. the $\hat{C}$ statistic.

An alternative to the denominator shown in equation (2.6) is obtained if we consider O_k to the sum of independent nonidentically distributed random variables. This suggests that we should standardize the squared difference between the observed and estimated expected frequency by

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j).$$

In a series of simulations Xu (1996) [39] showed that use of this alternative results in a trivial increase in the value of the test statistic. We proved (see Appendix A.2. for the proof) mathematically

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j) < n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k). \tag{2.7}$$

The establish of Hosmer-Lemeshow test is a milestone in logistic regression modeling process and it has served as a standard method in model validation ever since its introduce. Unfortunately, Hosmer-Lemeshow test for the goodness-of-fit in logistic regression model have obtained criticism since its emerge on its deficiencies. Hosmer-Lemeshow's $\hat{C}$ test clearly has structural problems with highly granular data, the opposite of sparsity, which contains only a relatively small number of distinct covariates. Bertolini et al. (2000) [40] pointed out that in such cases software packages can report different p values for the same data set because of using different conventions in forming the groups. Hosmer et al. (1997) illustrated this variation in results using low birth weight data that resulted in six p values from 0.02 to 0.16 using six different statistical software packages. An even more extreme example was reported by Pigeon and Heyse (1999) [41] who obtained p values ranging from 0.02 to 0.45 for a single data set. Bertolini et al. [40] also pointed out that Hosmer-Lemeshow's test results may be inaccurate when the number of covariate patterns is much less than number of subjects. The power of Hosmer-Lemeshow's $\hat{H}$ test is also limited in cases when the estimated probabilities only span a small sub-interval of $(0, 1)$. The degrees of freedom would be 0, if, for example, the range of estimated probabilities is small (< 0.2).

Hosmer and Hjort (2002) [31] and Kuss(2002) [32] have summarized further the deficiencies of the Hosmer-Leleshow test: (1) its limiting distributions has not been rigorously derived; (2) it is a conservative test and has low power to detect specific types of lack of fit (such as nonlinearity in an explanatory variable); (3) it is highly dependent on how the observations are grouped; (4) if too few groups are used to calculate the statistic (for example, five or less groups), it will almost always indicate that the model fits the data.

Although these deficiencies have been reported, the Hosmer-Lemeshow statistic is the most widely used goodness-of-fit test for logistic regression modeling in practice (SAS Institute Inc. 1995, SPSS Inc. 2003, and a good number of R packages). The availability in statistical software packages may be one reason why the Hosmer-Lemeshow is popular. Its popularity may also due to its properties such as: (1) it is intuitively appealing and easy to compute; (2) it has sound support from simulation studies. Additionally, lack of a better approach also contributes to its popularity since no other methods have been put forward that do not also have difficulties. In consequence, the Hosmer-Lemeshow test statistic remains the standard goodness-of-fit test when evaluating the fit of a logistic regression model with continuous covariates.

2.4 Unweighted residual sum of squares test

In 1989, Copas [24] proposed an unweighted residual sum of squares (RSS) test for testing proportions. Here, we simplify Copas's original test statistic and provide its asymptotic distribution under the strict binary situation.

The test statistic is:

$$S = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2 \tag{2.8}$$

where $\hat{\pi}_i$ is the estimated probability of the positive response.

Hosmer et al. (1997) [23] show that the asymptotic moments of $\hat{S}$ can be calculated and used to perform a statistical test. In detail, under the null hypothesis that the fitted logistic

regression is correct in all aspects, the first two asymptotic moments are:

$$E\left[\hat{S} - trace(V)\right] = E\left[\hat{S} - \sum_{i=1}^{n} \pi_i(1-\pi_i)\right] \approxeq 0, \tag{2.9}$$

and

$$\begin{aligned} Var\left\{\hat{S} - trace(V)\right\} &= Var\left\{\hat{S} - \sum_{i=1}^{n} \pi_i(1-\pi_i)\right\} \\ &\approxeq \boldsymbol{d}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}\boldsymbol{d} \\ &= (1-2\boldsymbol{\pi})^T(\boldsymbol{W}-\boldsymbol{W}\boldsymbol{Q}\boldsymbol{W})(1-2\boldsymbol{\pi}), \end{aligned} \tag{2.10}$$

where $\boldsymbol{d}$ is the vector with general element $d_i = (1-2\pi_i)$ and $\boldsymbol{V} = diag(\pi_i(1-\pi_i))$ is the $n \times n$ covariance matrix, $\boldsymbol{M} = \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the logistic regression version of the hat matrix, where $\boldsymbol{X}$ is the design matrix. Then $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$, and $\boldsymbol{W}$ is the diagonal matrix with diagonal elements as $\pi_i(1-\pi_i)$. Therefore, after substituting π_i with its estimate, i.e. $\hat{\pi}_i$, one can formulate a standardized statistic

$$Z_{Copas} = \frac{\hat{S} - trace(\hat{V})}{\sqrt{\hat{Var}\left\{\hat{S} - trace(\hat{V})\right\}}} \tag{2.11}$$

to assess the significance using the standard normal distribution.

If y_i is the number of positive response within the i^{th} covariate pattern rather than at the individual level, then the test statistics S becomes $\sum_{i=1}^{n}(y_j - m_i\hat{\pi}_i)^2$, and accordingly $\boldsymbol{V}$ can be replaced with $diag(m_i\pi_i(1-\pi_i))$.

The RSS statistic is also a special case of the class of statistics considered by le Cessie and van Houwelingen [25].

Simulation results of Kuss (2002) [32] suggested that this approach may have comparable power to the Hosmer-Lemeshow test, but further studies of this method on a variety of model scenarios are needed.

2.5 Cumulative sums of residuals test

In 1991 Su and Wei [30] proposed a goodness-of-fit test for the generalized linear model based on the cumulative sums of residuals (CUSUM). The cumulative sums of residuals is built on partitions of "some space", which means the ordering of y_i may be determined by that of the fitted $\hat{y}_i$ or of the values of a covariate is required. The central idea is that under the null hypothesis that the fitted model is correct in all aspects, then the process of summing the residuals, $y_i - \hat{y}_i$, should yield a function that varies, over the partition defining the cumulative sums of residuals in an unsystematic manner about zero. If at any point the sum is large in absolute value then we may have evidence of the lack-of-fit.

First the residuals under the GLM framework is defined as $R_i = y_i - g^{-1}(\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}})$. With logistic regression, $g^{-1}(\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}) = \hat{\pi}_i = \frac{exp\{\boldsymbol{X_i}^T\hat{\boldsymbol{\beta}}\}}{1+exp\{\boldsymbol{X_i}^T\hat{\boldsymbol{\beta}}\}}$. The cumulative sums of residuals may be defined by:

$$d_{(j)} = \sum_{i=1}^{(j)}(y_i - \hat{\pi}_i) \tag{2.12}$$

where $i = 1, \ldots, j$, and $j = 1, \ldots, n$. Note that $d_{(n)} = 0$ since $\sum_{i=1}^{(n)} y_i = \sum_{i=1}^{(n)} \hat{\pi}_i = m_n$.

Under the assumption there is no association between the residuals R_i and the fitted values, $\hat{\pi}_i$, or a given covariate, the $d_{(j)}$ will fluctuate around zero. A cumulative measure of the distance between the observed data and the model assumed under the null hypothesis is the discrete Kolmogorov-Smirnov statistic (recommended by Horn (1997) [42]), namely,

$$d = \max_{1\leq j\leq n}\left|\sum_{k=1}^{j} R_{\sigma_k}\right| \tag{2.13}$$

where the ordering σ is a permutation of the integers $1, \ldots, n$.

Su and Wei consider the following statistics as an alternative version of residuals process:

$$W_n(\boldsymbol{t}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} R_i I(\boldsymbol{X}_i \leq \boldsymbol{t}), \tag{2.14}$$

where $\boldsymbol{t} = (1, t_1, \ldots, t_p)^T$, and $I(.)$ is the indicator function. And $W_n(\boldsymbol{t})$ is a function of $\boldsymbol{t}$ and a multi-parameter stochastic process. Under the null hypothesis, we would expect that this cumulative-sum process, based on residuals, fluctuates about 0 in a non-systematic manner. Thus, a large value of the Kolmogorov-Smirnov type test statistic, $G_n = \sup_{\boldsymbol{t} \in 1 \times \Re^p} \left| W_n(\boldsymbol{t}) \right|$, leads to the conclusion of model misspecification.

Under mild conditions, Su and Wei proved (Section 8 Appendix) [30] that G_n has the same asymptotic distribution as

$$\widetilde{G}_n = \sup_{\boldsymbol{t} \in 1 \times \Re^p} \left| \widetilde{W}_n(\boldsymbol{t}; \boldsymbol{\beta}) \right|, \tag{2.15}$$

where

$$\widetilde{W}_n(\boldsymbol{t}; \boldsymbol{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i R_i \left\{ I(\boldsymbol{X}_i \leq \boldsymbol{t}) + \eta^T(\boldsymbol{t}; \boldsymbol{\beta}) \mathfrak{I}^{-1}(\boldsymbol{\beta}) X_i u(\boldsymbol{X}_i^T \boldsymbol{\beta}) \right\}, \tag{2.16}$$

and

$$\eta(\boldsymbol{t}; \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g^{-1}(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \boldsymbol{X}_i I(\boldsymbol{X}_i \leq \boldsymbol{t}), \tag{2.17}$$

$$\mathfrak{I}(\boldsymbol{\beta}) = -\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}, \tag{2.18}$$

$$u(\boldsymbol{X}_i^T \boldsymbol{\beta}) = \frac{\partial g^{-1}(\boldsymbol{X}_i^T \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{2.19}$$

and Z_i is a random sample from $N(0, 1)$, independent of $(y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, and $U(\boldsymbol{\beta})$ is the likelihood score function as defined as follows

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} u(\boldsymbol{X}_i^T \boldsymbol{\beta}) \boldsymbol{X}_i (y_i - g^{-1}(\boldsymbol{X}_i^T \boldsymbol{\beta})). \tag{2.20}$$

The proposed test based on G_n is consistent against the alternative hypothesis [30]. One advantage of the cumulative sums of residuals test is that this procedure is asymptotically distribution free. That is, the asymptotic null distribution of our test statistic is independent of the underlying error distribution function. But this goodness-of-fit test is computationally

intensive so it might not be feasible when there are many covariates.

Since G_n has the same asymptotic distribution as $\widetilde{G}_n$, the p-value of G_n, should be the same as the p-value of $\widetilde{G}_n$, i.e.

$$Pr(G_n \geq g_n) = Pr(\widetilde{G}_n \geq g_n).$$

Su and Wei presented a simulation method for computing the p-value, $Pr(\widetilde{G}_n \geq g_n)$, through replacing the $\boldsymbol{\beta}$, and $\mathfrak{I}$ in $\widetilde{W}_n$ with $\hat{\beta}$ and $\hat{\mathfrak{I}}$, which are obtained from the observed data respectively, then generating random samples $Z_1, \ldots, Z_n$ from $N(0,1)$ independently and computing $\widetilde{W}_n$ repeatedly to estimate $Pr(\widetilde{G}_n \geq g_n)$. Based on their simulation studies, Su and Wei claim that this large sample approximation to the null distribution of G_n is fairly satisfactory for moderate sample sizes.

The computation of goodness-of-fit tests based on the cumulative sums of residuals is time consuming. The reason is that, to compute $I(\boldsymbol{X_i} \leq \boldsymbol{t})$, the indicator function in $W_n(\boldsymbol{t})$, we need to consider all the possible combinations of $\boldsymbol{t} = (t_1, \ldots, t_p)^T$ in multidimensional space. Therefore, one efficient way to compute the p-value of $W_n(\boldsymbol{t})$ is using Monte Carlo optimization. The p-value is defined as the proportion of Monte-Carlo simulations for which Monte Carlo optimization is a computational algorithm that relies on repeated random sampling to compute the results. It maybe due to the reason that the algorithm of calculating the p value for the CUSUM test is not as straight forward as alternative tests, the CUSUM test is not applied and investigated as widely as other goodness-of-fit tests.

Pan and Lin (2002) [43] proposed two new cumulative sums of residuals goodness-of-fit tests based on Su and Wei's work [30]. We introduce them as follows.

Let $x_k = \infty$ for all $k \neq j$, then $W_n(\boldsymbol{t})$ can be simplified as

$$W_j(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i I(x_{ij} \leq t) \tag{2.21}$$

where x_{ij} is the j^{th} component of X_i, $t \in \mathfrak{R}$. The corresponding test statistic is

$$G_j = \sup_{t \in \mathfrak{R}} |W_j(t)|. \tag{2.22}$$

Letting the indicator function $I(.)$ be $I((X_i^T \hat{\beta}_i) \leq t)$, $t \in \mathfrak{R}$. Then, we have

$$W_g(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i I((X_i^T \hat{\beta}_i) \leq t) \tag{2.23}$$

and the corresponding test statistic is

$$G_g = \sup_{t \in \mathfrak{R}} |W_g(\boldsymbol{t})|. \tag{2.24}$$

The p-values of G_j and G_g are still computed based on $\widetilde{G}_n$ in (2.15), but with the original indicator function replaced by $I(x_{ij} \leq t)$ and $I((X_i \hat{\beta}_i) \leq t)$, respectively. The first $W_j(t)$ is designed to test the functional form of one particular covariate x_i, $i = 1, \ldots, p$. The second $W_g(t)$ is more informative about the link function, but in fact this procedure is sensitive to any alternative that leads to incorrect specification of the marginal mean, including the link function, the functional form of the response variable, the linear predictor in GLM and the conditional linear predictor in GLMM (generalized linear mixed models) [43] [44].

One advantage of the cumulative sums of residuals test is that this procedure is asymptotically distribution free, that is, the asymptotic null distribution of the test statistic is independent of the underlying error distribution function, but this goodness-of-fit test is computationally intensive, so it might take very long time or even not be feasible to run the test when there are many covariates.

In our work, we will focus primarily on the performance of the supremum test statistic G_g since it can accesses the overall lack of fit of the fitted model. Simulation studies suggest that G_g has reasonable power against model misspecification such as the functional forms of covariates and omitted interaction terms [44].

2.6 Other goodness-of-fit tests

2.6.1 Stukel's generalized logistic regression method

A class of generalized logistic regression models proposed by Stukel (1988) [45] provides a convenient model amenable to testing the adequacy of the fitted logistic model. The Stukel model uses a logit function with two additional parameters α_1 and α_2, thereby allowing either for asymmetry in the curve or for a different rate of approach to the (0,1) bounds. The usual linear logistic model results when $\alpha_1 = 0$ and $\alpha_2 = 0$. The logistic model can be tested against this more general model by a simple procedure.

Let $\hat{\eta}_i$ be the linear predictor from the fitted model with the link function $g()$ as shown in equation (1.4), that is, $\hat{\eta}_i = g(\hat{\pi}_i|\mathbf{x}_i) = \mathbf{x}_i{}^T\hat{\boldsymbol{\beta}}$ where $\mathbf{x}_i$ is the vector of covariate values for individual i and $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients. Then define two new variables z_1 and z_2 based on $\hat{\eta}$ as follows:

$$z_1 = \frac{1}{2}\hat{\eta}^2 \quad if \quad \hat{\eta} \geq 0, \quad otherwise \quad z_1 = 0 \tag{2.25}$$

$$z_2 = -\frac{1}{2}\hat{\eta}^2 \quad if \quad \hat{\eta} \leq 0, \quad otherwise \quad z_2 = 0 \tag{2.26}$$

Add these two variables to the logistic regression model and test the null hypothesis that both of their corresponding coefficients α_1 and α_2 are equal to 0.

It is noticeable that the proposed generalization separates the standard logistic curve, in sigmoid shape, at the point of $\pi = 0.5$, which is equivalent to $\eta = 0$. Two shape parameters α_1 and α_2 govern the behavior of two tails of the curve respectively. When $\alpha_1 = \alpha_2$, the corresponding probability curve $\pi(\eta)$ is symmetric, when $\alpha_1 \neq \alpha_2$, the two tails are asymmetrical. The larger absolute value of α_1 and α_2 indicates the larger deviation from the standard sigmoid curve as illustrated in Figure 1b of Stukel [45].

Under the proposed generalization, the ordinary logistic model has $\alpha_1 = \alpha_2 = 0$. Stukel [45] further noted that $\alpha_1 = 0.62$, $\alpha_2 = -0.037$ gives the log-log and complementary log-log

model; $\alpha_1 = \alpha_2 = 0.165$ gives the probit model. Therefore, these two parameters allow the generalized logistic model to be either symmetric or asymmetric with tails either lighter or heavier than the case with the ordinary logistic model.

To investigate the characteristic of the test statistic, let $\boldsymbol{\beta}$ be the vector of coefficients of the original logistic regression model and $l(\boldsymbol{\beta}, \alpha_1, \alpha_2)$ be the log-likelihood function from n observations for the newly generalized logistic regression model. Then $s^T = (s1, s2) = (\partial l/\partial \alpha_1, \partial l/\partial \alpha_2)$ is the score function evaluated at $(\hat{\boldsymbol{\beta}}, 0, 0)$. Under the null hypothesis, the test statistics $s^T Var(s)^{-1} s$ has an asymptotic $\chi^2(2)$ distribution. Therefore Stukel proposed a two-degree-of-freedom score test that assesses the tails of the logistic regression model.

Hosmer and Lemeshow (1997) [23] argued that it's not a real goodness-of-fit since the test statistics are not based on residuals, However, they do agree that this method in general is more powerful in detecting lack-of-fit than the known methods in many situations. However, the performance of Stukel's score test with more complicated models is not clear.

Hosmer and Lemeshow (2000) [5] recommended the partial likelihood ratio test statistic can be used instead of the score test. The test statistic is in the following form:

$$ST = -2l(\boldsymbol{X}) - (-2l(\boldsymbol{X}, \boldsymbol{Z})),$$

where $l(\boldsymbol{X})$ and $l(\boldsymbol{X}, \boldsymbol{Z})$ are the maximum log-likelihoods from the assumed model and generalized model, respectively. ST has an asymptotic chi-square distribution with two degrees of freedom under null hypothesis.

It's worth noting that in 1982, Brown [46] developed a different two-parameter score test, which is so-called Brown's score test, based on an extended logistic regression model proposed by Prentice (1976) [47]. A comparison of the Prentice model to the Stukel model [45] showed that both offer the same level of flexibility in terms of generating alternative models, but Stukel's generalized logistic model is analytically easier to use. The latter does not need the integration work, whereas it's required with the Prentice model. Stukel's approach also

provides the expressions for the variables that are needed to carry out Brown's score test.

2.6.2 Smoothed residual based test

The idea of smoothed residuals is to compare a 'smoothed' value of the outcome variable for each subject (which is a weighted average of the y values for other subjects 'near' the subject) to a similarly smoothed estimate of the logistic probabilities. The idea of 'nearness' is defined as a distance measure in the $\boldsymbol{x}$ space as suggested by le Cessie and van Houwelingen (1995) [25], or it can be in the y space, and the bandwidth determines the size of the region over which the residuals are averaged. The weight function is defined using the uniform kernel for the $\boldsymbol{x}$ space by le Cessie and van Houwelingen (1991) [48], another common weight is a cubic weight in the y space introduced by Hosmer and Lemeshow (1997) [23].

1. The $\boldsymbol{x}$ space weight:

   - $w_{ij} = the\ distance\ between\ subject\ i\ and\ j\ = \Pi_{k=1}^{p} u(x_{ik}, x_{jk})$, where $u(x_{ik}, x_{jk}) = 1$ if $|x_{ik} - x_{jk}|/s_k \leq c_u$ or equals to zero otherwise. s_k is the sample standard deviation of x_{ik}, $i = 1, ..., n$, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the i^{th} observed covariate vector.
   - le Cessie and van Houwelingen recommend setting c_u so that about $\sqrt{n}$ subjects have non-zero weights. One recommended value of c_u by the authors is given by $c_u = \frac{1}{2}(4/n^{1/(2p)})$.

2. The cubic weights in the y space:

   - $w_{ij} = 1 - (|\hat{\pi}_i - \hat{\pi}_j|/c_{ci})^3$ if $|\hat{\pi}_i - \hat{\pi}_j| \leq c_{ci}$ and $w_{ij} = 0$ otherwise.
   - The constant c_{ci} depends on i and is chosen such that $\sqrt{n}$ weights are non-zero for each subject.

Then the smoothed standardized residuals are given by $\hat{r}_{si} = \sum_{j=1}^{n} w_{ij}\hat{r}_j$, here $\hat{r}_j = (y_i - \hat{\pi}_j)/\sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}$, the w_{ij}'s are defined as previously. The test statistic is given by:

$$\hat{T}_r = \sum_{i=1}^{n} \frac{\hat{r}_{si}^2}{\hat{var}(\hat{r}_{si}^2)} \tag{2.27}$$

Then the p-value of the statistic can be evaluated using either a normal approximation or an easily computed scaled chi-square distribution.

Although the method, based on sums of squares of smoothed residuals, avoids the problems of the various grouping methods, it does have the disadvantage that results can depend on the choice of the bandwidth. Its performance was found by Hosmer, et al. (1997) [23] to be similar to that of the Hosmer-Lemeshow test at detecting departures from the true model. However, its Type I error rate was higher than expected in some settings.

2.6.3 Scaled Pearson's chi-square test

Osius and Rojek's (1992) [26] derived a large-sample normal approximation to the Pearson chi-square test statistic, which is usually referred to the scaled Pearson chi-square. They derived asymptotic moments for a general class of goodness-of-fit statistics under sparseness assumptions. A statistical test can be performed by standardizing X^2 with these moments and comparing the resulting test statistic to standard normal. This class incorporates both Pearson's chi-square X^2 and the residual deviance D, but moments in closed form can only be calculated for Pearson's chi-square X^2. Osius and Rojek describe the major steps to construct the scaled Pearson's chi-square statistic as follows:

1. Let J denote the number of possible distinct values of covariate vector X. Denote the number of subjects with the j^{th} covariate pattern $X = x_j$ by m_j, $j = 1, \ldots, J$.

2. Denote the fitted values from the model as $\hat{\pi}_j$, create a random variable w with its j^{th} value as $w_j = m_j\hat{\pi}_j(1 - \hat{\pi}_j)$.

3. create a random variable u with its j^{th} value as $u_j = \frac{1-2\hat{\pi}_j}{w_j}$.

4. Compute the Pearson chi-square statistic X^2 as in equation (2.2), i.e., $X^2 = \sum_{j=1}^{J} \frac{y_j - m_j\hat{\pi}_j}{w_j}$.

5. Conduct a weighted linear regression of u on X, the model covariates, with weight w. Let $WRSS$ denote the weighted residual sum-of-square from this regression in J dimensions.

6. Let A denote the correction factor for the variance, and $A = 2(J - \sum_{j=1}^{J} \frac{1}{m_j})$, let p denote the number of unknown parameters. Construct the standardized test statistic as

$$z = \frac{X^2 - (J - p - 1)}{\sqrt{A + WRSS}}. \tag{2.28}$$

This finishes the construction of the scaled Pearson's chi-square statistic, and the test significance can be obtained against the standard normal distribution.

Two more goodness-of-fit tests are derived similarly by finding the asymptotic moments for Pearson's chi-square statistic X^2. One is proposed by McCullagh (1985, [49]), the other is proposed by Farrington (1996, [50]). We introduce them briefly below.

McCullagh [49] followed the same idea as Osius and Rojek considered to derive asymptotic moments for X^2, but he argued for using conditional asymptotic moments for X^2 given the parameter estimates $\hat{\boldsymbol{\beta}}$. This approach conditioning on a sufficient statistic of the parameter estimates removes the dependency of X^2 from $\hat{\boldsymbol{\beta}}$, and accounts for the fact that the parameters from the logistic regression model have been estimated and were not fixed in advance. McCullagh also formed a standardized statistic and the p value can be obtained by comparing the standardized X^2 to the standard normal distribution.

Farrington [50] also based his approach on the conditioning principle as McCullagh proposed, but he further investigated a family of generalized Pearson statistics which extend X^2 by an additive constant. The modified X^2 statistic is as follows:

$$X_F^2 = \sum_{i=1}^{N} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1 - \hat{\pi}_i)} + \sum_{i=1}^{N} \frac{-(1 - 2\hat{\pi}_i)}{m_i\hat{\pi}_i(1 - \hat{\pi}_i)}(y_i - m_i\hat{\pi}_i) \tag{2.29}$$

Farrington claimed that it has minimal variance in this family and has the property of

local orthogonality to $\hat{\boldsymbol{\beta}}$. Since the Farrington statistic removes the dependence of the distribution of the test statistic on the bias of the parameter estimates and thus can be considered as an improvement of the McCullagh method. Farrington addressed the advantage of the modified Pearson statistic X^2 is the simplicity of its first three moments after incorporating a degrees-of-freedom correction, and all moment tend to those of the χ^2 distribution on $n-p$ degrees of freedom in the limit as $\mu_i \to \infty$ or $m_i \to \infty$. Therefore the approximate moments for X_F^2 can be calculated in closed form and the standardized statistic can be compared to the standard normal distribution. However, the Farrington test has its structural deficiency in the case of extreme sparseness. For example, in the case of continuous covariate presented in the model, when $m_i = 1$ for all i, $X_F^2 = N$, and the test will never reject the null hypothesis of a good fit. Thus, under the situation with continuous covariates included in logistic regression models, Farrington's test won't be an appropriate choice for the lack of fit test in general.

2.6.4 Two-stage Hosmer and Lemeshow tests

Pulkstenis and Robinson (2002) [51] proposed a two-stage modification of the Hosmer–Lemeshow test. At the first stage the individual observations are grouped according to a cross-classification of all categorical covariates in the model, and at second stage they are split according to the median estimated probability of the within the newly defined groups. Analogous to the Hosmer-Lemeshow test, an ordinary Pearson test or the deviance is then calculated to compare expected and observed counts in the resulting contingency table. This model requires sorting all responses by model-based fitted probabilities within each unique covariate pattern, and then creating two subcategories within each covariate pattern, by splitting the category to two based on the median of fitted probabilities. Based on such defined sub-grouping, the proposed test statistics, X^{*2} and D^*, are given by

$$X^{*2} = \sum_{i=1}^{G} \sum_{h=1}^{2} \sum_{j=1}^{2} \frac{(O_{ihj} - E_{ihj})^2}{E_{ihj}} \tag{2.30}$$

$$D^* = 2 \sum_{i=1}^{G} \sum_{h=1}^{2} \sum_{j=1}^{2} O_{ihj} log \frac{O_{ihj}}{E_{ihj}} \tag{2.31}$$

where i indexes covariate patterns, h indexes the substratification due to ordering by fitted probabilities, and j indexes columns. The degrees of freedom for these statistics are obtained by modifying the degrees of freedom for the regular Pearson and deviance chi-square statistics given by $G - k - 1$, where G is the number of rows in the cross-classification of Table 2.2, where the notation is J instead, and k is the number of covariates in the model. The degrees of freedom for X^{*2} and D^* are given by $2G - k - 2$, where $2G$ refers to the number of rows in the new stratification splitting each row of Table 2.1 and k is the number of categorical variables in the model. The degrees of freedom are also analagous to the $G - 2$ suggested by the Hosmer-Lemeshow test, but subtract k additional degrees of freedom due to the modeled covariates defining the groups.

This approach is proposed to detect omitted interaction between a continuous variable and categorical variable, but also to incorporate the full design structure into the process. The authors showed by simulation that their tests are superior to the standard Hosmer–Lemeshow test. However, the requirement of both categorical and continuous covariates in the model, due to the construction principle of the proposed test statistics, is considered a weakness of this method.

2.6.5 Different partition methods

2.6.5.1 Tsiatis's score test

Tsiatis (1980) [52] introduced a score test statistic that can be used to evaluate the fit of a binary logistic regression model with continuous covariates by partitioning the covariate space for grouping. First, the covariate space is partitioned into G distinct regions without

reference to the estimated parameters or observed data, then an augmented logistic model is introduced that gives the conditional probability of a successful outcome, given the observed values of the covariates, as

$$\boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{I}) = \frac{exp\left\{\boldsymbol{x\beta} + \sum_{g=1}^{G} \gamma_g I^{(g)}\right\}}{1 + exp\left\{\boldsymbol{x\beta} + \sum_{g=1}^{G} \gamma_g I^{(g)}\right\}} \tag{2.32}$$

where $\left\{I^{(1)}, \ldots, I^{(g)}\right\}$ are a set of indicator functions that are defined as $I^{(g)} = 1$ when covariates lie in it the g^{th} region, and $I^{(g)} = 0$ otherwise, and $\left\{\gamma_1, \ldots, \gamma_G\right\}$ is the set of additional coefficients associated with each of the G indicator functions. Tsiatis's goodness-of-fit statistic tests the null hypothesis that $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_G) = \mathbf{0}$, which means the null model is

$$\boldsymbol{\pi}(\boldsymbol{x}) = \frac{exp\{\boldsymbol{x\beta}\}}{1 + exp\{\boldsymbol{x\beta}\}} \tag{2.33}$$

is the best fit to the data out of all of the possible instances of the augmented model. Here, the $\boldsymbol{\gamma}$ are considered the parameters of interest and the $\boldsymbol{\beta}$ are considered nuisance parameters.

The Tsiatis statistic is $T = \boldsymbol{S}^T \boldsymbol{V}^- \boldsymbol{S}$, where $\boldsymbol{S}$ is a G-dimensional column vector, with general elements $S_g = \frac{\partial l}{\partial \gamma_g}$, $g = 1, \ldots, G$. l notates for the log likelihood, and $\boldsymbol{V}^-$ is any generalized inverse of the $G \times G$ covariance matrix, here $\boldsymbol{V}$ is not full rank (Tsiatis 1980) [52].

Although partitioning in the covariate space overcomes some of the shortcomings of grouping methods that only look for discrepancies in the direction of the logit, it also suffers from some deficiencies. Tsiatis did not indicate how the partitioning of the covariate space should be accomplished. No methods are specified for determining what number of partitions to use, nor how they should be chosen. Su,et al. (1991) [30] point out that tests that partition the covariate space, including the Tsiatis test, can draw different conclusions when different partitions are applied. They give several specific examples, one of which shows the Tsiatis statistic, given two choices of partitioning, resulting in p values of 0.04 and 0.38. Lin et al.

(2002) [43] also point out that "the partition of the covariate space is arbitrary and different partitions may result in conflicting conclusions".

An alternative goodness-of-fit test statistic developed by Pigeon and Heyse [41] by combining characteristics of the Hosmer-Lemeshow and Tsiatis test statistics is a chi-square test denoted by J^2. To account for the heterogeneity of the predicted probabilities within partitioned groups they multiply the Hosmer-Lemeshow $\hat{C}$ by a "correction term". Canary et al. (2017) [53] compare it with the Hosmer-Lemeshow test and the Tsiatis's score test, and found it did not outperform over the Hosmer-Lemeshow test.

2.6.5.2 Clustering-based partition methods

Xie et al. (2008) [54] apply a partitioning strategy based on clustering in the covariate space to both a Pearson chi-square type statistic and a score statistic. The clustering method identifies regions within the covariate space where observations are close, as defined using a criterion such as Euclidean or Mahalanobis distance. They point out that this method has the advantage that observations within these groups will have similar covariate profiles. They state that both of their statistics should have asymptotic distributions that are between $\chi^2(G-k-1)$ and $\chi^2(G-1)$, where k is the number of covariate values. They use the rubric $G = 10$ if $k < 5$, and $G = k + 5$ if $k \geq 5$, applied to $df = g - k/2 - 1$ for the Pearson chi-square type statistic, and $df = g - 1$, which is the rank of the conditional covariance matrix of the scores. They compare the performance of their statistics to that of the original HL , which uses the deciles-of-risk grouping method. Both HL and their Tsiatis-like score statistic maintained the test size more consistently, while their Pearson chi-square type statistic was conservative. Both of their test statistics had more power than HL to detect departures from a true underlying model.

Dreiseitl, et al. (2012) [55] propose a strategy to overcome the problem of detecting lack of fit in a region in the covariate space by using the Pigeon-Heyse statistic (Pigeon, et al. [41]), which is reported to have an asymptotic chi-square with $G - 1$ degrees of freedom,

and applied a grouping method based on clustering. Three strategies they tested were based on (1) clustering with self-organizing maps; (2) clustering with a k Hmeans algorithm; and (3) random assignment of data points to groups. In their simulations, they varied the dimensionality of the data. The simulation study was small with data limited to 20 data sets. They report that their approach does aid in locating regions of poor calibration in the data space, although with such small samples this result is not very strong.

White (1982) [56] develop a test from a very different perspective to address the problem of lack of fit test. The test statistic involves the information matrix, thus it's called the information matrix test. This method is proposed as a general approach to testing for model misspecification by comparing two different estimates of the covariance matrix of the parameter estimates (the negative inverse of the information matrix), one constructed from the hessian of the log likelihood function, the other constructed using the mean (first derivatives of the log-likelihood function) of the contributions to the outer products of the gradient (second derivatives) of the log likelihood function, but the limiting distribution of the test statistic presents departure from the asymptotic χ^2 approximation to it. Kennan and Neumann (1988) and others ([57], [58]) suggest that the χ^2 approximation can be poor even in what are normally thought of as quite large samples.

We list out a great number of the existing goodness-of-fit tests designed to be suit for assessing the logistic regression model when sparse data are presented in the assumed model from the literature. This review presents a dynamic research field where various strategies have been proposed to detect the lack of fit in logistic regression, an important modeling approach in medical research. We can see those tests address the model validation from different angles and they all have their own merits and demerits. It seems no one method is phenomenal or outperformed over others in all scenarios. In consequence, when considering model validation strategy after fitting a proposed model, practitioners are often to some extend being confused by the various choices.

Chapter 3

Asymptotic Theorem Guided New Partitioning Method for Goodness-of-fit Test

3.1 Introduction

In this research project, we propose a well-acceptable strategy to partition the observations based on expected frequencies and to form a goodness-of-fit test based on this partition strategy. We proposed to use the Pearson chi-square for the test statistic. This method is very general and allows for the assessment of the goodness-of-fit for logistic regression models with all kinds of covariate configurations. Since ordinary Pearson chi-square tests work well when all the covariates are categorical and when the number of cross-classifications of categorical covariates is not too large relative to the sample size. Our research project will only focus on the situation when continuous predictor variables are presented.

Our grouping strategy is motivated by at least two considerations. One is the preferable rule about grouped frequency data, i.e. the expected frequencies should be greater than 5 to make the null chi-square distribution valid, the other is related to the drawbacks of the Hosmer-Lemeshow test, which can possibly break the rule, especially when there are rare of either events or non-events.

3.2 Asymptotic theorem: Pearson chi-square goodness of fit test

When we conduct Pearson chi-square goodness-of-fit test using X^2 similar as shown in (1.10) for mutually exclusive G grouped frequency data, i.e. frequency data in one categorical vari-

able with G levels, we actually form a series of G dependent Z^2-type statistics. Dependency here means that once the frequencies of the first $G-1$ groups are specified, the frequency of the last G group is fixed due to the sum of all frequencies is N.

The term "goodness-of-fit" is used to determine whether sample data are consistent with a hypothesized distribution, in other words, to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. X^2 is formulated to the goodness-of-fit test statistic, intuitively the limiting distribution of X^2 can be approximated by $\chi^2(G-1)$. We form the following proposition to address the asymptotic distribution of X^2 mathematically.

Proposition: Let X^2 be the random variable as $\sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$ for k mutually exclusive classes, we have that X^2 converges in distribution to χ^2 distribution with $k-1$ degrees of freedom. i.e.

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} \xrightarrow{d} \chi^2(k-1) \ \ as \ \ n \to \infty. \tag{3.1}$$

where n_i is a random variable which denotes the number of sample Y's falling in to class i, $i = 1, \cdots, k$, and π_i is a constant parameter which denotes the probability of a sample Y_l falling in to class i, $l = 1, \cdots, n$, such that $\pi_1 + \cdots + \pi_k = 1$.

Proof: Let's define random variables $I(Y_1 \in class \ \ i), \ldots, I(Y_n \in class \ \ i)$ that indicates whether each sample Y_l is in class i or not be coded as 1 or 0 respectively, and the random variable has Bernoulli distribution $B(\pi_i)$ with probability of success (i.e., in class i). Then

$$E[I(Y_l \in i)] = \mathbb{P}(Y_l \in i) = \pi_i$$

and variance

$$Var(I(Y_l \in i)) = \pi_i(1 - \pi_i).$$

By the Central Limit Theorem, the random variable

$$\begin{aligned}\frac{n_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}} &= \frac{\sum_{l=1}^{n} I(Y_l \in i) - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}} \\ &= \frac{\sum_{l=1}^{n} I(Y_l \in i) - nE[I(Y_l \in i)]}{\sqrt{nVar(I(Y_l \in i))}} \\ &\xrightarrow{d} N(0,1) \ \ as \ \ n \to \infty. \end{aligned} \tag{3.2}$$

i.e. converges in distribution to $N(0,1)$. Therefore the random variable

$$\frac{n_i - n\pi_i}{\sqrt{n\pi_i}} \xrightarrow{d} \sqrt{1-\pi_i}N(0,1) = N(0, 1-\pi_i) \ \ as \ \ n \to \infty. \tag{3.3}$$

Let Z_i stand for the random variable, i.e. $Z_i = \frac{n_i - n\pi_i}{\sqrt{n\pi_i}} = \sqrt{n}\frac{(n_i/n) - \pi_i}{\sqrt{\pi_i}}$, the Central Limit Theorem states that the vector $\boldsymbol{Z}$ converges in distribution to $N(\mathbf{0}, \mathbf{\Omega})$, a multivariate normal distribution. We will find the covariance matrix $\mathbf{\Omega}$ next.

To compute the covariance between Z_i and Z_j is equivalent to compute the covariance between $\frac{n_i - n\pi_i}{\sqrt{n\pi_i}}$ and $\frac{n_j - n\pi_j}{\sqrt{n\pi_j}}$, which is

$$\begin{aligned} Cov(Z_i, Z_j) &= E[(Z_i - E(Z_i))(Z_j - E(Z_j))] \\ &= E[Z_i Z_j] \quad (since \quad E(Z_i) = E(Z_j) = 0) \\ &= E\left[\frac{n_i - n\pi_i}{\sqrt{n\pi_i}} \times \frac{n_j - n\pi_j}{\sqrt{n\pi_j}}\right] \\ &= \frac{1}{n\sqrt{\pi_i\pi_j}}\left\{E[n_i n_j] - E[n_i n\pi_j] - E[n_j n\pi_i] + n^2\pi_i\pi_j\right\} \\ &= \frac{1}{n\sqrt{\pi_i\pi_j}}\left\{E[n_i n_j] - n\pi_j E[n_i] - n\pi_i E[n_j] + n^2\pi_i\pi_j\right\} \\ &= \frac{1}{n\sqrt{\pi_i\pi_j}}\left\{E[n_i n_j] - n\pi_j n\pi_i - n\pi_i n\pi_j + n^2\pi_i\pi_j\right\} \\ &= \frac{1}{n\sqrt{\pi_i\pi_j}}\left\{E[n_i n_j] - n^2\pi_i\pi_j\right\}. \end{aligned} \tag{3.4}$$

Since one sample can only be included in one class, we have

$$I(Y_l \in i)I(Y_l \in j) = 0 \tag{3.5}$$

Thus,

$$\begin{aligned}
E[n_i n_j] &= E\Big[(\sum_{l=1}^{n} I(Y_l \in i))(\sum_{l'=1}^{n} I(Y_{l'} \in j))\Big] \\
&= E\Big[\sum_{l,l'}^{n} I(Y_l \in i)I(Y_{l'} \in j)\Big] \\
&= E\Big[\sum_{l=l'} I(Y_l \in i)I(Y_{l'} \in j)\Big] + E\Big[\sum_{l\neq l'} I(Y_l \in i)I(Y_{l'} \in j)\Big] \\
&= 0 + E\Big[\sum_{l\neq l'} I(Y_l \in i)I(Y_{l'} \in j)\Big] \ \ (by \ \ (3.5)) \\
&= n(n-1)E[I(Y_l \in i)]E[I(Y_{l'} \in j)] \\
&= n(n-1)\pi_i\pi_j. \qquad (3.6)
\end{aligned}$$

Plug (3.6) to (3.4), we have

$$\begin{aligned}
Cov(Z_i, Z_j) &= \frac{1}{n\sqrt{\pi_i\pi_j}}\Big[n(n-1)\pi_i\pi_j - n^2\pi_i\pi_j\Big] \\
&= \frac{1}{n\sqrt{\pi_i\pi_j}}(-n\pi_i\pi_j) \\
&= -\sqrt{\pi_i\pi_j}. \qquad (3.7)
\end{aligned}$$

Therefore, the covariance matrix of the vector $\boldsymbol{Z}$ is given by

$$\boldsymbol{\Omega} = Cov(\boldsymbol{Z}) = \begin{pmatrix} 1-\pi_1 & -\sqrt{\pi_1\pi_2} & \dots & -\sqrt{\pi_1\pi_k} \\ -\sqrt{\pi_1\pi_2} & 1-\pi_2 & \dots & -\sqrt{\pi_2\pi_k} \\ \vdots & \vdots & \dots & \vdots \\ -\sqrt{\pi_1\pi_k} & -\sqrt{\pi_2\pi_k} & \dots & 1-\pi_k \end{pmatrix} \tag{3.8}$$

Let $\boldsymbol{\pi} = (\sqrt{\pi_1}, \dots, \sqrt{\pi_k})^T$ be the k dimensional probability (column) vector, then $\boldsymbol{\Omega}$ can be rewritten as:

$$\boldsymbol{\Omega} = \boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T. \tag{3.9}$$

where $\boldsymbol{I}$ is the k dimensional identity matrix.

From equation (3.8) we can easily find the trace of $\boldsymbol{\Omega}$, i.e.

$$trace(\boldsymbol{\Omega}) = 1 - \pi_1 + \dots + 1 - \pi_k = k - \sum_{i=1}^{k} \pi_i = k - 1. \tag{3.10}$$

We can also get it by using the linearity and the commutativity property of the trace as follows.

$$trace(\boldsymbol{\Omega}) = trace(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T) = trace(\boldsymbol{I}_k) - trace(\boldsymbol{\pi}\boldsymbol{\pi}^T) = k - trace(\boldsymbol{\pi}\boldsymbol{\pi}^T).$$

Noticing that $trace(\boldsymbol{\pi}\boldsymbol{\pi}^T) = \sum_{i=1}^{k} \pi_i = 1$, then we have $trace(\boldsymbol{\Omega}) = k - 1$.

It can also be shown that $\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T$ is an idempotent matrix.

$$\begin{aligned} (\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T)^2 &= (\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T)(\boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T) \\ &= \boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T - \boldsymbol{\pi}\boldsymbol{\pi}^T + \boldsymbol{\pi}\boldsymbol{\pi}^T\boldsymbol{\pi}\boldsymbol{\pi}^T \\ &= \boldsymbol{I}_k - 2\boldsymbol{\pi}\boldsymbol{\pi}^T + \boldsymbol{\pi}\boldsymbol{\pi}^T \ \ (since \ \ \boldsymbol{\pi}^T\boldsymbol{\pi} = \sum_{i=1}^{k} \pi_i = 1) \\ &= \boldsymbol{I}_k - \boldsymbol{\pi}\boldsymbol{\pi}^T. \end{aligned} \tag{3.11}$$

Therefore, $\boldsymbol{\Omega}$ is a projection matrix of rank equal to its trace $k - 1$. Immediately we get

that $\boldsymbol{\Omega}$ has $k-1$ eigenvalues equal to 1, one eigenvalue equals to 0, which also implies that the vector $\boldsymbol{Z}$ converges in distribution to $N(\mathbf{0}, \boldsymbol{\Omega}^*)$, a $k-1$ dimensional multivariate normal distribution.

Furthermore, there exists a rotation matrix $\boldsymbol{A}$ that makes

$$\boldsymbol{A\Omega A}^T = \begin{pmatrix} 1 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 \\ \vdots & \vdots & ... & \vdots & \vdots \\ 0 & 0 & ... & 1 & 0 \\ 0 & 0 & ... & 0 & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_{k-1} & \mathbf{0}_{(k-1)\times 1} \\ \mathbf{0}_{1\times(k-1)} & 0_{1\times 1} \end{pmatrix} \tag{3.12}$$

where $\mathbf{0}_{i\times j}$ is the i rows by j columns matrix filled with 0, particularly $0_{1\times 1} = 0$.

Denote $\boldsymbol{W} = \boldsymbol{AZ} \sim N_k(\mathbf{0}, \boldsymbol{A\Omega A}^T)$. Then $\boldsymbol{W}$ is a vector $(W_1, \ldots, W_{k-1}, 0)$ of i.i.d. $N(0,1)$ Gaussians with only $k-1$ non null coordinates (the first $k-1$ coordinates). The function $f(\boldsymbol{Z}) = Z_1^2 + Z_2^2 + \ldots + Z_k^2$ is the norm square, i.e. $||\boldsymbol{Z}||^2$, and it is invariant if we rotate its argument. This means $f(\boldsymbol{Z}) = f(\boldsymbol{AZ}) = f(\boldsymbol{W}) = W_1^2 + W_2^2 + \ldots + W_{k-1}^2$ is chi-square distributed with $k-1$ degrees of freedom. This completes the proof.

In the above proposition, $n \to \infty$ is a required condition by applying the Central Limit Theorem appropriately (refer to (3.2), (3.15)), but it is hard to justify in reality. A more practical and general rule of thumb is that the sample size n is "sufficiently large" if

$$n\pi_i \geq 5, \ \textit{and} \ \ n(1-\pi_i) \geq 5. \tag{3.13}$$

If the requirements are satisfied, then when the null hypothesis is true, the CDF of X^2 is closely approximated by $\chi^2(G-1)$.

It is noticeable that this rule of thumb is the same as that for the approximation of the binomial distribution by the normal distribution [59], which is often suggested as:

$$F_{n,\pi}(k) \approxeq \Phi\left(\frac{k+0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right), \tag{3.14}$$

where $F_{n,\pi}(\cdot)$ denotes the distribution function of a binomial distribution with parameters n and π, and $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution, $k \in \{0, 1, \ldots, n\}$ denotes the number of event or success out of n Bernoulli trials, the value of 0.5 is a continuity correction term. The rule of thumb shown in (3.13) for the domain of application of the approximation are given by many text books, for example, Agresti's book [4].

We illustrate the above rule of thumb for the general chi-square goodness-of-fit test statistic. Suppose that the possible outcomes of an experiment are $1, \ldots, G$, with probabilities $\pi_1, \ldots, \pi_G$, respectively. The experiment is carried out n times independently. Let $O_1, \ldots, O_G$ be the number of observations in group $1, \ldots,$ group G respectively in the n outcomes. Note that $\sum_i^G O_i = n$ and $\sum_i^G \pi_i = 1$. The chi-square statistic is defined as

$$X^2 = \sum_{i=1}^{G} \frac{(O_i - n\pi_i)^2}{n\pi_i}. \tag{3.15}$$

here O_i is called the observed frequency of cell i and $n\pi_i$ is the expected frequency. When the null hypothesis is true and all expected frequencies $n\pi_i$ are greater than 5, the CDF of X^2 is closely approximated by the chi-square distribution of $G-1$ degrees of freedom, i.e. $\chi^2(G-1)$.

Suppose we get $G = 6$ groups. Figure 3.1 shows the true CDF of X^2 when all $n\pi_i$'s are 5 (the staircases) and $\chi^2(5)$ (the smooth curve), so the sample size is $n = 30$. The two CDF's are close to each other.

Figure 3.2 shows the staircases and the curve when all $n\pi_i$'s are 2, that is the sample size is $n = 12$ only. In this graph, the curve deviates noticeably from the staircases. The jumps are quite bigger than that in case of all $n\pi_i$'s are 5.

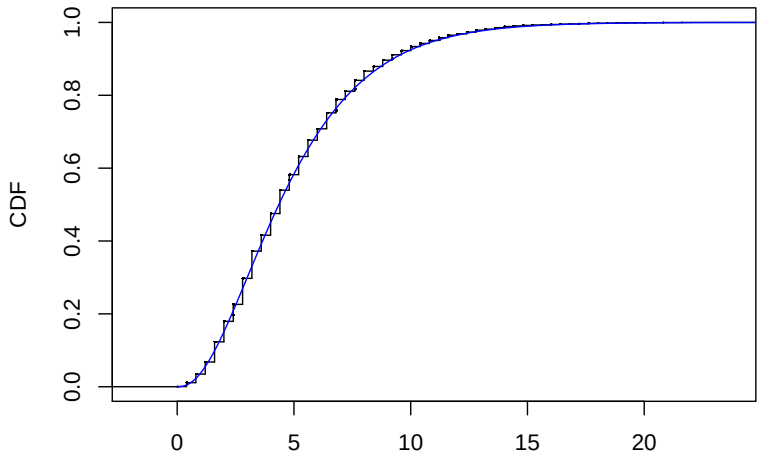

Figure 3.1: The CDFs of X^2 and their approximations, $\chi^2(5)$
(G=6, n=30, all $\pi_i = 1/6$)

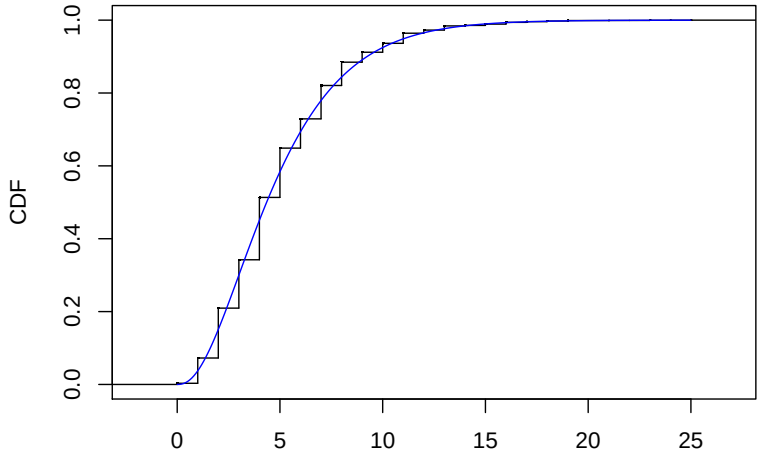

Figure 3.2: The CDFs of X^2 and their approximations, $\chi^2(5)$
(G=6, n=12, all $\pi_i = 1/6$)

The two empirical cumulative distribution functions of X^2 under different situations are both based on $10,000$ simulations. These visualizations illustrate that to ensure the CDF

can be closely approximated by the chi-square distribution, the chi-square goodness-of-fit test requires all expected frequencies be at least 5.

Cochran (1952) [60] suggested to use enough cells to keep the expectations down to the levels recommended by Williams (1950) [61], for example, let the expected cell number be 12 per cell for $n = 200$, 20 per cell for $n = 400$, 30 per cell for $n = 1,000$. but his recommendations are not explicit and requires more detailed study.

Based on our experience, it is not rare that when performing the Hosmer-Lemeshow test, the expected cell number in one or even more bins/groups among $G = 10$ groups is less than 5. It is possible to illustrate that, for example, Hosmer and Hjort (2002) [31] reported that the Hosmer-Lemeshow goodness-of-fit test will result in an uninterpretable p value when applied to relatively small data sets or when applied to data with low expected decile cell frequencies.

Our proposal of the new partitioning method is guided by the asymptotic theorem and attempted to obey the general rule of thumb in term of non-small expected frequencies for the chi-square type of goodness-of-fit test. We describe our proposal with details nextly.

3.3 New partitioning method for chi-square goodness-of-fit test

The procedure to build the proposed test statistic is as follows:

1. Use the derived model to estimate the probability of the event for each of N subjects.

2. Sort the estimated probabilities in ascending order, we get $\hat{\pi}_{(1)}, \ldots, \hat{\pi}_{(N)}$.

3. Form N intervals based on these ordered estimated probabilities. The N intervals can be demonstrated in Table 3.1.

4. Group the data into G intervals/bins based on the preferable rule that each estimated frequency in a interval in both outcome classes (i.e., success/failure, yes/no, 1/0) should be at least greater than 5. The estimated frequency is actually the cumulative sum

Table 3.1: Form N intervals with frequency of 1 for each interval

i	interval i	observed y	fitted $\hat{y}$	Number of obs. in interval i
1	$(0, \hat{\pi}_{(1)}]$	0	$\hat{\pi}_{(1)}$	1
2	$(\hat{\pi}_{(1)}, \hat{\pi}_{(2)}]$	0	$\hat{\pi}_{(2)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$(\hat{\pi}_{(N-1)}, \hat{\pi}_{(N)}]$	1	$\hat{\pi}_{(N)}$	1

of the fitted values of subjects falling in to this specific interval. The resulting $G \times 2$ frequency table is formed as demonstrated in Table 3.2.

Table 3.2: Form G intervals with observed and estimated expected frequencies

k	interval k	Number of observed		Number of expected	
		$y = 0$	$y = 1$	$y = 0$	$y = 1$
1	$[\hat{\pi}_{(1)}, \hat{\pi}_{(n_1')}]$	O_{01}	O_{11}	$\hat{E}_{01}$	$\hat{E}_{11}$
2	$(\hat{\pi}_{(n_1')}, \hat{\pi}_{(n_1'+n_2')}]$	O_{02}	O_{12}	$\hat{E}_{02}$	$\hat{E}_{12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	$(\hat{\pi}_{(N-n_G')}, \hat{\pi}_{(N)}]$	O_{0G}	O_{1G}	$\hat{E}_{0G}$	$\hat{E}_{1G}$

The goodness of fit test statistic, T_G, is obtained by calculating the Pearson chi-square statistic from the resulting $G \times 2$ table of observed and estimated expected frequencies.

5. Construct the test statistic T_G. Let $n_k^{'}$ be the total number of observations in the k^{th} group, $O_k = \sum_{j=1}^{c_k} y_j$ be the total number of observed positive responses among all subjects who fall within the c_k covariate patterns represented in the k^{th} bin, and $\bar{\pi}_k$ be the average estimated probability of all subjects falling within the k^{th} bin, c_k denotes the number of covariate pattterns in the k^{th} bin and m_j denotes the number of subjetcs of covariate pattern j in the k^{th} bin. Then after partitioning data yields G distinct bins, we can readily construct a Pearson chi-square statistic (similar to Hosmer and Lemeshow, 1989) as follows:

$$T_G = \sum_{k=1}^{G} \left(\frac{(O_{1k} - \hat{E}_{1k})^2}{\hat{E}_{1k}} + \frac{(O_{0k} - \hat{E}_{0k})^2}{\hat{E}_{0k}} \right) = \sum_{k=1}^{G} \frac{(O_k - n_k^{'}\bar{\pi}_k)^2}{n_k^{'}\bar{\pi}_k(1 - \bar{\pi}_k)} \tag{3.16}$$

where $n_k^{'}$ is the total number of subjects in the k^{th} bin, i.e., $n_k^{'} = O_{0k} + O_{1k}$. O_k is the number of positive responses in the k^{th} bin, i.e., $O_k = O_{1k}$. And

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k^{'}} \tag{3.17}$$

is the average estimated probability in the k^{th} bin and c_k denotes the number of covariate patterns in the k^{th} bin, m_j denotes the number of subject of covariate pattern j in the k^{th} bin.

3.4 More theoretical considerations of the new partitioning method

We proposed to apply the rule for valid asymptotic null distribution to partition the observations into G groups, we will examine the statistic T_G computed based on these groups.

3.4.1 Pearson chi-square type statistic

Pearson chi-square type statistics are usually defined in terms of cells which are fixed prior to taking observations. The standard theorem on the asymptotic distribution of the chi-square test is given by Cramer (1946, [62]) and Rao (1973, [63]). The asymptotic chi-square distribution rests on several assumptions about the true distribution of the observations and the estimated any unknown parameters. These assumptions may not be satisfied in practice due to the following reasons:

1. Estimators are obtained from the ungrouped data rather than from the grouped observations,

2. The groupings/bins may be driven by the data. For example, percentile splits,

3. The method of estimation is not efficient.

Chemroff and Lehmann (1954, [64]) and Watson (1958 [65], 1959 [37]) investigate the

above cases and conclude that when the estimates are not based on grouped data, the asymptotic χ^2_{G-p-1} distribution does not hold anymore.

Let $T_{G^\star}$ be this chi-square type statistic. In fact, it would have the following limiting distribution (with the condition of $G > p$, here p is the number of parameters involved in model fitting and outcome estimation) by Theorem 1 of Chemroff and Lehmann:

$$T_{G^\star} \xrightarrow{d} \sum_{j=1}^{G-p-1} y_j^2 + \sum_{j=G-p}^{G-1} \lambda_j y_j^2, \tag{3.18}$$

where $y_j \overset{i.i.d.}{\sim} N(0,1)$, $0 \leq \lambda_j \leq 1$, $\lambda_{G-p}, \ldots, \lambda_{G-1}$ are the roots of the characteristic equation:

$$|\widetilde{\mathfrak{I}} - (1-\lambda)\hat{\mathfrak{I}}| = 0, \tag{3.19}$$

where $\widetilde{\mathfrak{I}}$ is the information matrix estimated from the frequency data, and $\hat{\mathfrak{I}}$ is the information matrix estimated from the original data. As a consequence of their Lemma 1, the values of the p roots $\lambda_{G-p}, \ldots, \lambda_{G-1}$ are between 0 and 1 (Watson, 1958 [65]).

Moore and Spruill's (1975, [35]) Theorem 4.2 facilitates a unified derivation of the limiting distribution of three version of statistics. They denote Chemroff-Lehmann's version of statistic T_n as T_{2n} (i.e., they used different notations for the same statistic). By Theorem 5.1 in Moore and Spruill's paper, when some regularity assumptions hold, T_{2n} has limiting distribution (under (θ_0, η_0))

$$T_{2n} = ||V_n(\hat{\theta}_n, \phi_n)||^2 \xrightarrow{d} \chi^2_{M-m-1} + \sum_{j=M-m}^{M-1} \lambda_j \chi^2_{1j}. \tag{3.20}$$

By the remark after Lemma 1, Chemroff and Lehmann confirmed that the roots would determine the distribution of the test statistic $T_{G^\star}$. Without known weights λ_j, the null distribution of $T_{G^\star}$ is not well specified.

Since the model parameters in logistic regression with continuous covariates are estimated using maximum likelihood (as addressed in chapter 1) based on non-grouped data, our

proposed test statistic T_G is a Chemroff-Lehmann version statistic and would be labeled as T_{2n} by using Moore and Spruill's notation, its limiting distribution follows (3.18). Specifically under the null hypothesis, T_G in (3.16) obeys a central chi-square distribution χ^2, while under the alternative hypothesis, it obeys a non-central chi-square distribution χ^2_λ with λ being a non-central parameter. We will justify its degrees of freedom next.

3.4.2 Degrees of freedom

As we mentioned previously, the Pearson chi-square type statistics are defined in terms of cells which are fixed prior to taking observations. Moreover if parameters are to be estimated from the data, they must be estimated by asymptotically good estimators based on the observed cell frequencies, for example, the maximum likelihood estimators. Both Chemroff and Lehmann, and Moore and Spruill showed that if MLE's based on full samples are used, the asymptotic distribution of the statistic need to no longer be chi-square. Instead, stochastically the distribution of the test statistic is bounded by known chi-square distributions,

$$\chi^2_{G-p-1} \leq T_{G^\star} \leq \chi^2_{G-1}, \tag{3.21}$$

here the range is based on the roots of the characteristic equation (3.19). We justify the situations when these boundaries hold.

Let's look into the asymptotic distribution of (3.18) (same as (3.20)). The first term has a χ^2 distribution with d.f. of $(G-p-1)$ since y_i's are independent standard normal. The degrees of freedom of the second term would depend on those p λ's. Molinari (1977, [66]) again points out that those λ's are the eigenvalues of the determine equation (3.19) involving unknown information matrix and are possibly different from 0 and 1. We consider two extreme cases.

- Suppose all those p eigenvalues are close to 1, then the second term will have degrees of freedom being close to p. Put two terms together, $T_{G^\star}$ will have degrees of freedom

of $(G-p-1)+p=G-1$.

- Suppose all those p eigenvalues are close to 0, then the second term could be ignored. As a result, $T_{G^\star}$ will have degrees of freedom of $G-p-1$.

Molinari (1977, [66]) suggests that it would be safe (conservative) to consider $T_{G^\star}$ as having a χ^2 distribution with $G-1$ degrees of freedom. Based on their rather extensive simulation studies, Hosmer and Lemeshow suggest using $G-2$ as the degrees of freedom for their two Pearson type statistics, $\hat{C}$ and $\hat{H}$. Based on our previous justification, $G-2$ falls to the wide range and is thus a valid number to approximate the true distribution, but not a necessary result from Theorem 5.1 in Moore and Spruill (1975, [35]) as claimed by Yu et al. (2017) [36]. Watson (1959) [37] mentioned that further work would be needed to develop a procedure to find estimates of the λ_j weights in the characteristic equation (3.19) directly, which would lead to an improved approximation to the null distribution.

Note that the weights are all between 0 and 1, and the larger the value p, the number of coefficients in the $\boldsymbol{\beta}$ matrix, the more terms in the weighted sum. For the chi-square type of test statistic resulting in the number of partitioned groups G and with an aim of finding a representative approximation to the null distribution, given the bounded limiting distribution of $T_{G^\star}$ in (3.21), we proposed the following degrees of freedom for its corresponding null, i.e., the central chi-square distribution, with which we compare test statistic $T_{G^\star}$ to get $p-$values.

$$d.f. = \begin{cases} G-1 & \text{if } G < p \\ G-[\frac{p}{2}]-1 & \text{if } p \leq G < p+10 \\ G-p-1 & \text{otherwise} \end{cases} \tag{3.22}$$

where p is the number of parameters involved in estimating the outcome variable, $[p/2]$ denotes the flooring value, i.e. the integer part, of $p/2$.

Note that it is in lack of theoretical guidance on how to choose an accurate degrees of freedom for chi-square type of goodness-of-fit tests, our choice of these values are ruled purely

by the bounded limiting distribution of $T_{G^\star}$ in (3.21) with an aim of finding representative degrees of freedom. Our simulation and real data analysis result using this method show at least not as worse as the original Hosmer-Lemeshow test under many scenarios as shown in later chapters.

3.5 Example study

A disadvantage in the use of the chi-square like goodness-of-fit tests for the logistic regression model with continuous predictor variables proposed by Hosmer and Lemeshow that use fixed groups of the estimated probabilities has shown in recent work. It is possible to demonstrate situations where one set of fixed groups shows the model fits while the test rejects fit using another set of fixed group. We concern that if the estimated cell frequency is not large enough, say at least 5, which would make the required conditions unsatisfied and return misleading p values. The proposed new partition method leads to a chi-square type statistic similarly as the Hosmer-Lemeshow test, however our approach is different from the Hosmer-Lemeshow test in three aspects:

1. The proposed partitioning strategy is driven by asymptotic theorem and the number of partitions is determined by expected cell counts, whereas the Hosmer-Lemeshow test uses arbitrary and fixed decile bins.

2. We obey strictly the general rule, that is non-small expected frequency is allowed in each and every cell, the Hosmer-Lemeshow test can not avoid the issue of small expected cell frequency.

3. The degrees of freedom in our method is varied and involved with the number of unknown parameters, which is different from the fixed $G-2$ employed by Hosmer and Lemeshow.

We compare the proposed method to the Hosmer-Lemeshow test by analyzing the ICU data set next.

3.5.1 Analysis of ICU data set

The intensive care unit (ICU) study is included in Hosmer and Lemeshow's text book [5]. It is part of a large study on survival of patients following admission to an adult intensive care unit. The aim of this study is to develop a logistic regression model to describe the probability of survival to hospital discharge of these patients. The data set is available from R package **aplore3**.

In total 200 patients were included in the study, of which 160 patients lived and were discharged from the hospital, the rest 40 patients died. The outcome variable "sta" is a two-level factor for the patient status (lived/died). The independent variables are: age of the patient at ICU admission (range from 16 to 92, "age"), gender at two levels (male/female), systolic blood pressure at ICU admission (integers between 36 and 256, "sys"), heart rate at ICU admission (integers between 39 and 192, "hra"), cancer part of present problem (yes/no with coding 1/0, "cpr"), previous admission to an ICU (yes/no with coding 1/0, "pre"), type of admission (Emergency/Elective with coding 1/0, "type"), PH from initial blood gases (two levels with $1 \leq 7.25$ and $0 \geq 7.25$, "ph"), PCO2 from initial blood gases (two levels with $1 \geq 45$ and $0 \leq 45$, "pco"), level of consciousness at ICU admission (three levels with 0 = No coma/Deep Stupor, 1 = Deep stupor and 2 = Coma, "loc").

The study has been analyzed by many researchers, for example, by Lemeshow et al. (1993, [67]) and by Lemeshow and Le Gall (1994, [68]). We remain "age" be the continuous predictor variable in our analyses and convert other continuous variables to categorical variables based on some clinical thresholds. For example, "hra" is grouped into greater than 150 beats/min or less (1/0), and "sys" is grouped into less than 90 mmHg or greater (1/0). We combine level 2 and level 3 to one level in "loc", then "loc" becomes a two-level factor, i.e. 1 is for "no coma or deep stupor" and 2 for "coma or deep stupor", due to small number of patients in the original levels 2 and 3.

Based on the literature references and our modeling results, we presented two final models here. The first model includes seven predictor variables and the second one includes six. Both

models perform well in terms of goodness of fit. Table 3.3 shows the fitting results of Model 1, and Table 3.4 shows its corresponding goodness of fit result.

Table 3.3: Estimation result for final model 1

var	Estimate	Standard Error	Z Value	p-value
Intercept	-5.963	1.235	-4.829	< 0.001
age	0.037	0.013	2.771	0.006
sys1	2.239	0.819	2.735	0.006
cpr1	1.178	0.829	1.421	0.155
Type1	2.038	0.843	2.419	0.016
ph(< 7.25)	1.644	0.880	1.868	0.062
pco(> 45)	-2.556	1.022	-2.50	0.012
loc1	3.749	0.931	4.026	< 0.001

Table 3.4: Goodness-of-fit testing result for final model 1

Method	Statistic	p-value
HL	$\hat{C}$	0.540
NP	T_G	0.843
RSS	Z	0.857
CUSUM	$\hat{W}$	0.934

All the results from four goodness-of-fit tests, namely the Hosmer-Lemeshow (HL), our new partitioning method(NP), the unweighted residual sum of squares (RSS), and the cumulative sums of residuals (CUSUM), support that the fitted model 1 fits adequately, even it's noticeable that variable "cpr" doesn't have a significant effect on the outcome, the patients' binary outcome. And the p-value of the proposed new partition method is close to the p-values of RSS and CUSUM methods. The interest here is to compare our new partitioning method with the Hosmer-Lemeshow's fixed grouping method. Let's look into the partitioned tables which are generated based on these two methods.

Table 3.5 presents the grouping results from the Hosmer-Lemeshow and Table 3.6 presents the grouping results from the proposed partitioning method after fitting Model 1 to the ISU data set.

Both Tables 3.5 and 3.6 show clearly that 80% of the observation has the estimated values

Table 3.5: Partitioning result from the HL test for final model 1

	Observed		Estimated	
	$y = 0$	$y = 1$	$\hat{y} = 0$	$\hat{y} = 1$
[0.00232,0.0195]	22	0	21.732812	0.2671876
(0.0195,0.0344]	18	0	17.468205	0.5317948
(0.0344,0.0401]	18	2	19.250780	0.7492203
(0.0401,0.0557]	18	2	19.103605	0.8963954
(0.0557,0.102]	21	0	19.310318	1.6896822
(0.102,0.164]	16	3	16.378229	2.6217712
(0.164,0.219]	17	3	16.180418	3.8195820
(0.219,0.28]	14	6	14.991926	5.0080737
(0.28,0.63]	12	8	12.256788	7.7432116
(0.63,0.997]	4	16	3.326919	16.6730810

Table 3.6: Partitioning result from the proposed test for final model 1

	Observed		Estimated	
	$y = 0$	$y = 1$	$\hat{y} = 0$	$\hat{y} = 1$
[0.00232,0.135]	104	6	104.78736	5.212638
(0.135,0.22]	26	4	24.63700	5.362995
(0.22,0.28]	14	6	14.99193	5.008074
(0.28,1]	16	24	15.58371	24.416293

below 0.28. The pattern of the cumulative sums of residuals can be shown as below in Figure 3.3.

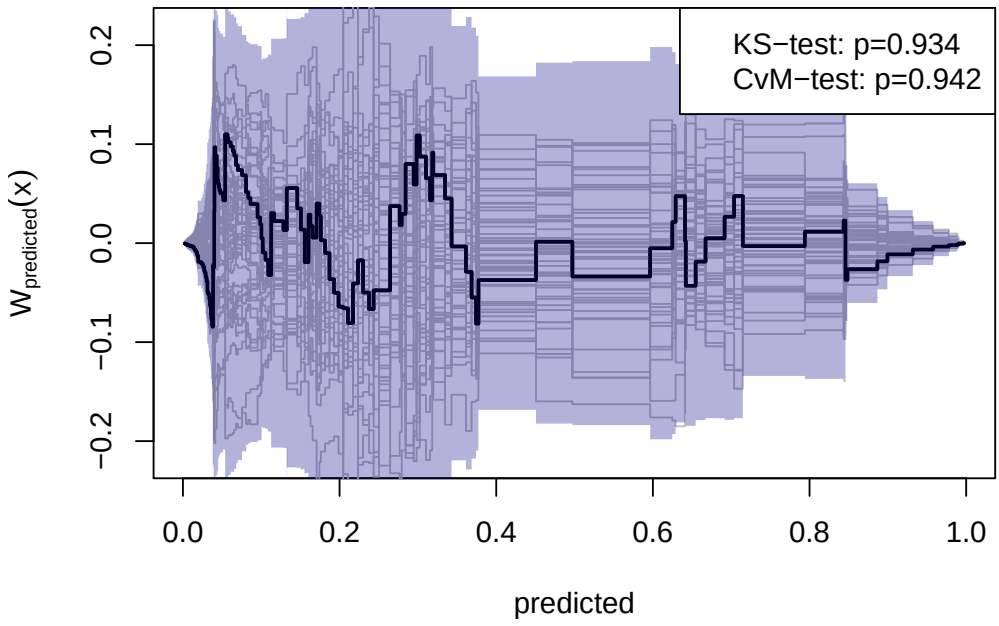

Figure 3.3: The cumulative sums of residuals process: final model 1

Final model 2 removes the predictor variable "cpr" for the purpose to pursuing a simpler model, so model 2 remains five variables. Its estimating results are shown in Table 3.7.

Table 3.7: Estimation result for final model 2

var	Estimate	Standard Error	Z Value	p-value
Intercept	-5.878	1.217	-4.828	< 0.001
age	0.036	0.013	2.727	0.006
sys1	2.067	0.796	2.598	0.009
Type1	2.141	0.844	2.536	0.011
ph(< 7.25)	1.597	0.871	1.833	0.067
pco(> 45)	-2.291	0.981	-2.335	0.020
loc1	3.868	0.928	4.167	< 0.001

Table 3.8 shows its corresponding goodness of fit result.

Table 3.8: Goodness-of-fit testing result for final model 2

Method	Statistic	p-value
HL	$\hat{C}$	0.906
NP	T_G	0.798
RSS	Z	0.730
CUSUM	$\hat{W}$	0.924

Let's look into the grouping results of the HL test and the proposed test for model 2. Table 3.9 presents the grouping results from the Hosmer-Lemeshow and Table 3.10 presents the grouping results from the proposed partitioning method after fitting Model 1 to the ISU data set.

Table 3.9: Partitioning result from the HL test for final model 2

	Observed		Estimated	
	$y=0$	$y=1$	$\hat{y}=0$	$\hat{y}=1$
[0.00297,0.0211]	20	0	19.752157	0.2478426
(0.0211,0.0338]	20	0	19.433410	0.5665905
(0.0338,0.0443]	19	1	19.205330	0.7946697
(0.0443,0.0599]	18	2	19.023688	0.9763122
(0.0599,0.114]	19	1	18.270899	1.7291012
(0.114,0.168]	20	3	19.751302	3.2486984
(0.168,0.218]	16	3	15.253959	3.7460413
(0.218,0.281]	12	6	13.460896	4.5391040
(0.281,0.62]	12	8	12.524025	7.4759746
(0.62,0.991]	4	16	3.324335	16.6756655

Table 3.10: Partitioning result from the proposed test for final model 2

	Observed		Estimated	
	$y=0$	$y=1$	$\hat{y}=0$	$\hat{y}=1$
[0.00297,0.125]	101	6	101.852767	5.147233
(0.125,0.215]	27	3	24.928123	5.071877
(0.215,0.28]	16	7	17.370750	5.629250
(0.28,0.39]	11	5	10.630391	5.369609
(0.39,1]	5	19	5.217969	18.782031

Similarly as the partitioning results from model 1, Tables 3.9 and 3.10 show almost the same that 80% of the observation has the estimated values below 0.28 from model 2. But

the pattern of the cumulative sums of residuals in Figure is not exactly the same as shown in Figure 3.4.

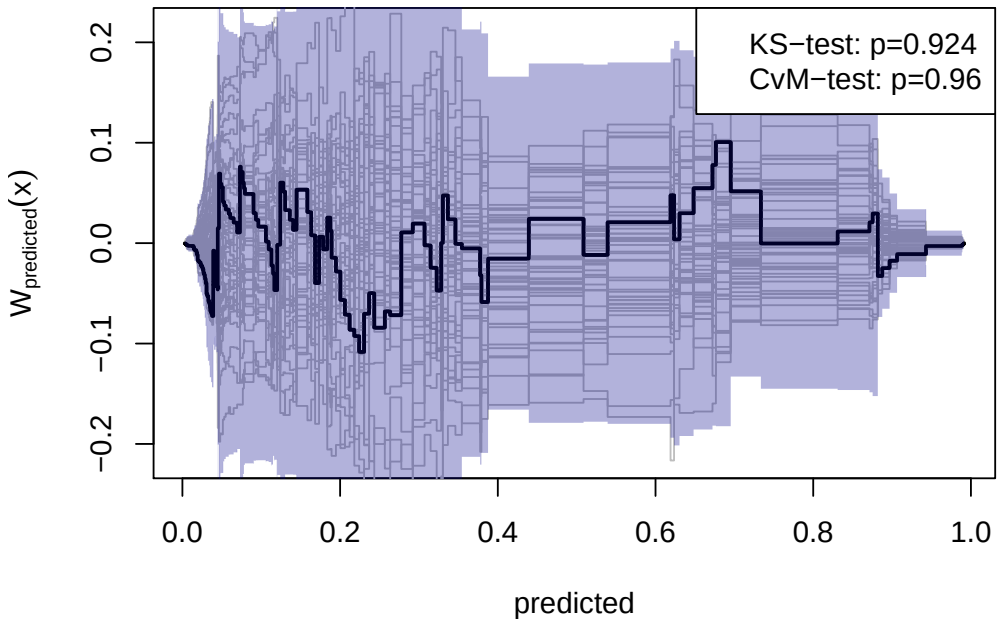

Figure 3.4: The cumulative sums of residuals process: final model 2

We list two final models here just to remind the fact that a model without detecting the lack of fit does not imply it is the best one, goodness of fit tests just provide information on the adequacy of the fitted model, or on how well the theoretical distribution fits the empirical distribution of the observed data, it is not aimed for model comparison or model selection. However, a model with merits passing some specific criteria does not guarantee that the model fit the data well. That's why it is important for practitioners to conduct model validation after model selection. Specifically, back to the fitting results of two final models, their AIC values are very close to each other, and model 2 holds a bit smaller AIC ($\delta = 0.02$), so model 2 could be recommended due to its parsimony, i.e. with one less predictor variables than model 1. However, just as Homer and Lemeshow point out, the choice of model should always depend on biological or clinical considerations in addition to

statistical results.

To further assess the performance of these models, for example the empirical size and power of each test, it's hard to get insights from real data. We conduct our simulation study in the followed chapters .

Chapter 4

Comparison of Empirical Size and Power of Goodness-of-fit Tests

4.1 Introduction

When sparse data, say continuous covariates or small number of observations within some covariate patterns, are presented in the logistic regression framework, the Pearson's chi-square and deviance tests for goodness-of-fit purpose are not appropriate anymore. Alternative test statistics have been developed to be suit for the situations. In 1980's Hosmer and Lemeshow develop two test statistics, namely $\hat{C}$ and $\hat{H}$, to meet the challenging situation, and their methods serve as a standard to follow since they're introduced. Other tests have been introduced as well to tackle the goodness-of-fit through various aspects of model validation processing, but it seems no one test is dominant enough. As discussed in chapter 2, each test has its own merits and demerits. It's hard for users to choose.

Even though the Hosmer-Lemeshow tests have been criticized in many aspects in terms of model validation as discussed in the previous chapters, they are still standing out as one of the best known methods with some great properties. For example, it is concordant with many situations in medical research that the population are divided into equal risk deciles (e.g. 0% to 10%, 11% to 20%, etc.), the grouping method is pre-defined and fixed, it is easy to implemented in software packages, a lot more features to list here.

Due to its serious deficiencies, Hosmer and Lemeshow suggest to make statistical decision by using other model evaluation merits together with their method, and also to perform

goodness-of-fit test with other available methods. More importantly, clinical factors and practical considerations should be taken into account when building a good model.

Empirical size and power are two major measures of the performance of goodness-of-fit tests. The empirical rejection rate of the null hypothesis under the null hypothesis is the size of a test, the empirical rejection rate of the null hypothesis under an alternative hypothesis is the power of a test. Therefore size of a test is the type I error rate of a test when the specified model is correct, power of a test is the probability of a test to detect deviation from the specified model when the model is incorrect in simulation studies. The pair of hypotheses for goodness-of-fit tests are stated as in section 3 of chapter 1.

As we introduced in chapter 1, three major simulation studies have been published to assess the power of goodness-of-fit tests for logistic regression with sparse data, which are performed by Hosmer et al. (1997, [23]), Hosmer and Hjort (2002, [31]) and Kuss (2002, [32]) respectively. Recently Liu et al. (2012, [69]) performed a power comparison as well when introducing her omnibus test method. All the first three studies use only sample size of 100 and 500, the latter one uses 100, 500 and 1000. And the test statistics under consideration are different. Table 4.1 lists these studies with goodness-of-fit tests considered individually.

It is worth noting that recently two more simulation studies extended the power comparison to large data under the frame of Hosmer-Lemeshow test. Paul et al. (2013) [58] increase the sample size of their simulated study up to 25,000. Yu et al. (2017) [36] increase the sample size of their simulated study beyond 25,000 and up to 50, 000. They share the same goal, that is to force the power of the Hosmer-Lemeshow test as stable as possible in case of large data is under model fitting. Those two studies focus on optimizing the power of the Hosmer-Lemeshow test under the situation with large sample size rather than comparing power of different tests.

As we discuss in the previous chapter, it's hard to know which test(s) should perform better in a specific situation since we do not know the features of each test. Through simulation, we can empirically assess the performance of our proposed new partition methods

Table 4.1: Simulation study designed to compare the power of various goodness-of-fit tests

Study	Sample size	Statistics for power comparison
Hosmer et al (1997)	100, 500	Hosmer-Lemeshow $\hat{C}$ Hosmer-Lemeshow $\hat{H}$ Pearson X^2 RSS smoothed residuals Royston Stukel's score
Hosmer & Hjort (2002)	100, 500	Hosmer-Lemeshow $\hat{C}$ Hosmer-Lemeshow $\hat{H}$ Pearson X^2 RSS Partial sums of residuals
Kuss (2002)	100, 500	Hosmer-Lemeshow $\hat{C}$ Pearson X^2 Deviance RSS Information Matrix Farrington's X^2
Liu et al. (2012)	100, 500, 1000	Hosmer-Lemeshow $\hat{C}$ Hosmer-Lemeshow $\hat{H}$ Stukel's score Xie

with other three tests under some common situations we could encounter every day, the performance of a test includes evaluating the type I error rate (size) and the power. In this chapter, we present our simulation study with similar setup as described in the published studies. We compare three tests, namely the Hosmer-Lemeshow, the unweighted residual sum of squares and the cumulative sums of residuals. The first two tests have been investigated in many studies, the third one is introduced not specifically for logistic regression models, instead it's for all classes of generalized linear models, in that it is worth investigating, it seems like a more general tool to us. To the best of our knowledge, no studies have been presented in the literature that compare these statistics with either small or large simulated samples. We assess both the empirical size and power of these goodness-of-fit tests.

4.2 Simulation setup

4.2.1 Independent variables

The simulation setting is similar to the existing comparative studies. We consider two aspects of the model, i.e., the distribution of the covariates and the coefficient values of the parameters in the model. One of the most commonly studied model is: one continuous variable with normal distribution, one dummy-coded categorical variable, i.e. a dichotomous term, and the interaction term between the two. We denote $\boldsymbol{x}$ as the continuous variables from normal distributions (and other distributions for a continuous random variable are considered as well later), denote $\boldsymbol{z}$ as the dummy variables at two levels, i.e., a dichotomous variable, and let $\boldsymbol{X}$ be the design matrix, i.e. it includes all independent variables plus a constant intercept term.

4.2.2 Dependent variable

We consider a binary outcome y_i here. $\pi_i = \pi(\boldsymbol{X}_i)$ is denoted as the probability of a positive response y_i, say, $y_i = 1$, and $\eta(\boldsymbol{X}_i) = g(\pi(\boldsymbol{X}_i)) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ as the linear predictor, where $g()$ is the link function, $\boldsymbol{\beta}$ is the vector of unknown parameters. $\hat{\pi}(\boldsymbol{X}_i)$ is denoted as the estimated probability of a positive response y_i, obtained from the estimated parameters $\hat{\boldsymbol{\beta}}$.

The mechanism of generating the outcome variable y is based on the definition of the Bernoulli variable. Let $\pi(\boldsymbol{X}) = \frac{exp\{\eta(\boldsymbol{X})\}}{1+exp\{\eta(\boldsymbol{X})\}}$, we generate $Bern(p)$ random variable $y = 1$ with $p = \pi(\boldsymbol{X})$.

4.2.3 Model setting

The simulation study contains two scenarios. Scenario 1 is designed to investigate the power when non-linear terms, i.e. the quadratic term of a continuous variable, the interaction term between a continuous covariate and a dichotomous variable, are omitted, scenario 2 is used

to investigate the power when a quadratic term is added to the fitted model or when an interaction term is added to the model.

For scenario 1, the settings of the null models and the fitted models are presented in Table 4.2.

Table 4.2: Scenario 1: Settings for null models and fitted models

Setting	Null Model	Fitted Model
1	$\eta = g(\pi) = -2 + x + 0.2x^2 + z - 2xz$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz$
2	$\eta = g(\pi) = -2 + x + \beta_2 x^2 + z - 2xz$ $\beta_2 = \begin{cases} 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \\ 0.8 \\ 1.0 \end{cases}$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$
3	$\eta = g(\pi) = -2 + x + 0.2x^2 + z + \beta_4 xz$ $\beta_4 = \begin{cases} -2 \\ -1.5 \\ -1.2 \\ -1.0 \\ -0.5 \\ -0.2 \end{cases}$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z$

Table 4.3 presents the settings of the null models and the fitted models for scenario 2.

Table 4.3: Scenario 2: Settings for null models and fitted models

Setting	Null Model	Fitted Model
1	$\eta = g(\pi) = -2 + x + z$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 z$
2	$\eta = g(\pi) = -2 + x + z$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x^2$
3	$\eta = g(\pi) = -2 + x + z$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$
4	$\eta = g(\pi) = -2 + x + z$	$\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 u$ $u \sim \begin{cases} N(0,1) \\ Unif(-3,3) \\ Beta(2,4) \end{cases}$

In the above tables, the null model is the model used to generate data sets, the fitted

model is the specified model assumed to fit the data sets. Thus the fitted model in setting 1 of two scenarios are correct since they specify exactly the same covariate forms as in the null models, i.e. the underlying models. When we apply different goodness of fit tests to the data, the p value should be small which would lead to a conclusion of accepting the null hypothesis, that is the specified model is correct. The p value in this setting is referred as the probability of rejecting the null hypothesis under null, that is the type I error or size of test. Whereas in other settings the p values are expected to be big, which would lead to a conclusion of rejecting the null hypothesis, that is the specified model is incorrect. In these cases the p values are referred as the probability of rejecting the null hypothesis under alternative hypothesis, that is the power of the tests. We exam the size and power of the four goodness of fit tests through these two scenarios in this chapter.

Specifically, setting 2 and 3 of scenario 1 are assessing the performance of those tests when the quadratic term and interaction term are omitted, setting 2 and 3 of scenario 2 are assessing the performance of those tests when adding a quadratic form of the continuous variable or an interaction term to the specified model. These settings are the the mostly discussed situations in literature.

In Table 4.2 two coefficient values in null model are varied, i.e. β_4 in scenario 4 and β_2 in scenario 5 are not fixed, instead they vary from -2 to -0.2, specifically we choose β_2 and β_4 equal $-2, -1.5, -1.2, -1, -0.5, -0.2$ respectively. These settings differentiate scenario 4 to scenario 2, and scenario 5 to scenario 3, and they are expected to examine how the weights of parameter coefficients affect the performance of those goodness of fit tests by strictly holding other coefficient values fixed.

4.2.4 Sample size

The sample size of simulated data are set as $N = 200$, 500, 1000, 2000, 5000, 10000, 15000, 25000 to represent a fixed but with a big range of coverage. Just as any hypothesis tests with a p value as a major result, it's well-known that p value decreases as the sample size

increase, consequently the power will be affected. And it has been reported from all the previous simulation studies that sample size matters for all goodness of fit tests in logistic regression models when continuous covariates are presented.

In all simulation scenarios we run 200 replications to compute the proportion of p values less than a conventional significant level $\alpha = 0.05$ for all hypothesis tests. This gives us the empirical size and power of those tests under each scenario as described previously. The reason we chose 200 replications is because it ensures the maximum margin of error of the rejection rate below 7% (the maximum standard error of the rejection rate is 0.5), and it is also reasonably enough to retain stable results through some practical trials with an aim in minimizing the computing time.

4.3 Test size: rejection rate under the null hypothesis

As discussed previously, setting 1 under two scenarios are for assessing the size of a test since they reflect the probability of rejecting the null hypothesis given the null is true. We refer this as the empirical size comparison, in fact it tells us how likely a goodness-of-fit test rejects a correct model, the smaller value is desired, which means the test can maintain the type I error well, say at the conventional level of 5%.

Table 4.4 shows the proportion of times each of the goodness-of-fit test rejects the null hypothesis given the null hypothesis is true at the significance level $\alpha = 5\%$, based on 200 rounds of model fitting trials, under scenario 1. That is, given the specified model is true and based on the nominal p value of each test on the same single simulated data by fitting the same model, the rejection rate (rejection of the null hypothesis if the nominal p value is less than the significance level $\alpha = 5\%$, non-rejection otherwise) would be calculated as the empirical size of a test.

Table 4.4: Setting 1 under scenario 1: Rejection rate of four goodness-of-fit tests

N	HL[1]	USS [2]	CUSUM [3]	NP [4]
200	0.015	0.055	0.05	0.085
500	0.02	0.02	0.08	0.065
1000	0.03	0.055	0.06	0.06
2000	0.035	0.02	0.06	0.075
5000	0.03	0.065	0.06	0.065
10000	0.035	0.065	0.055	0.04
15000	0.03	0.065	0.055	0.04
25000	0.02	0.045	0.065	0.035

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

Overall, the Hosmer-Lemeshow test maintains the empirical rejection rate below 5% across all sample size settings, that is the HL test controls type I error better than other three tests. USS test controls type I error a bit better than the CUSUM and the proposed NP test when sample size is below 2000, the proposed NP test present a higher rejection rate than all other three existing test when sample size is below 5000. In contrast, when sample size is greater than 10000, the NP controls the type I error better than the USS and CUSUM tests, it maintains type I error below 5%.

We conduct an inferential statistical test to compare these rejection rates by treating the number of rejection/non-rejection as aggregated binary outcome variable, treating "sample size" as a categorical variable with eight levels, and letting the proposed NP test be the base level of a categorical variable named "test", the analysis of deviance table in Table 4.5 shows that factor "test" has significant effect on the outcome rejection rate.

Table 4.5: Analysis of deviance for individual variables in setting 1 under scenario 1

	DF	Chisq	Pr(>Chisq)
sample size	7	2.026	0.958
test	3	23.427	<0.001

Then we look into the fitted model summary table related to the "test" variable, Table

4.6 shows that the NP test is more likely to reject the null hypothesis than the HL test and its rejection rate is significantly different than that of the HL test. The rejection rate of the NP test is not significantly different than that of other two tests, namely the Copas's residual sum of squares test and the cumulative sums of residual test.

Table 4.6: Logistic regression model fitting result for predictor variable "test" with the "NP" test as baseline level after adjustment of "sample size" effect

	Estimate	S.E.	Z	Pr(>|z|)
test: CUSUM	0.045	0.15	0.299	0.765
test: HL	-0.804	0.188	-4.279	< 0.001
test: RSS	-0.186	0.158	-1.178	0.239

Table 4.7 shows the proportion of times each of the goodness of fit test rejects the null hypothesis given the null hypothesis is true at the significance level $\alpha = 0.05$, based on 200 rounds of model fitting trials, under scenario 2.

Table 4.7: Setting 1 under scenario 2: Empirical size of four goodness-of-fit tests

N	HL [1]	USS [2]	CUSUM [3]	NP [4]
200	0.04	0.065	0.04	0.025
500	0.035	0.055	0.045	0.075
1000	0.035	0.065	0.05	0.045
2000	0.035	0.05	0.035	0.065
5000	0.04	0.055	0.065	0.045
10000	0.055	0.03	0.075	0.04
15000	0.05	0.035	0.05	0.025
25000	0.05	0.035	0.065	0.035

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

From Table 4.7 we can see that the HL test maintains the empirical rejection rate below 5% almost for all sample size settings, but it is clear that the HL can control type I error rate better when sample size is below 5000 than that when the sample size is above 10000. All other three testes do not control type I error as well as the HL test at the significant level at $\alpha = 5\%$. The USS test and the NP test perform very similar, and they control type

I error better when sample size is large, say greater than 5000. The CUSUM test relatively control type I error rate worse than other three tests at this scenario.

Similarly, we conduct an inferential statistical test to compare those rejection rates by fitting a logistic regression model to binary rejection outcomes, the analysis of deviance table in Table 4.8 shows that factor "test" has no significant effect on the outcome rejection rate.

Table 4.8: Analysis of deviance for individual variables in setting 1 under scenario 2

	DF	Chisq	Pr(>Chisq)
sample size	7	3.573	0.828
test	3	3.437	0.329

4.4 Test power: rejection rate under the alternative hypothesis

We examine the power of these test under two scenarios. Under scenario 1, setting 2 is to investigate the rejection rate of the null hypothesis when the specified model does not include the quadratic term of a continuous covariate. Setting 3 is to investigate the rejection rate of the null hypothesis when the specified model does not include the interaction term of a continuous covariate. Under scenario 2, setting 2 is to investigate the rejection rate of the null hypothesis when the specified model includes an additional quadratic term for the continuous variable x. Setting 3 is to investigate the rejection rate of the null hypothesis when the specified model includes an additional interaction term. Setting 4 is to investigate the rejection rate of the null hypothesis when the specified model includes an unrelated continuous variable with normal, uniform and beta distributions respectively.

4.4.1 Omission of the quadratic form of continuous variable, x^2

Table 4.9 shows the proportion of times each of the goodness of fit test rejects the null hypothesis based on 200 rounds of model fitting trials, in case of setting 2 under scenario 1 (See Table 4.2).

In this case, the correct model includes quadratic form of continuous variable x^2 with coefficient of β_2, whereas the wrong model is the one that drops the quadratic term x^2. Thus the rejection rate reflects the empirical power of each test under specific settings. We can see that the empirical power goes higher as the sample size increases for each test, also the parameter coefficient affects the power. Under a fixed sample size, as β_2 increased from 0.2 to 1.0, the power of each test increases. This trend is true across different sample size settings.

It is noticeable that the proposed new partition test is relatively conservative to detect the missing quadratic term across different sample sizes. Table 4.9 shows it has smaller power/rejection rate than other three tests with a specific parameter coefficient across all sample size.

Table 4.9: Setting 2 under scenario 1: Rejection rate of the omission of a quadratic term

N	β_2	HL [1]	RSS [2]	CUSUM [3]	NP [4]
200	0.2	0.085	0.225	0.09	0.09
	0.3	0.175	0.37	0.15	0.14
	0.4	0.23	0.545	0.21	0.16
	0.5	0.33	0.735	0.31	0.31
	0.8	0.715	0.925	0.72	0.575
	1.0	0.8	0.99	0.86	0.72
500	0.2	0.125	0.37	0.2	0.145
	0.3	0.33	0.715	0.28	0.29
	0.4	0.535	0.89	0.545	0.49
	0.5	0.795	0.98	0.79	0.65
	0.8	1	1	0.99	0.98
	1.0	1	1	1	1
1000	0.2	0.225	0.655	0.25	0.145
	0.3	0.575	0.935	0.59	0.325
	0.4	0.945	0.99	0.885	0.575
	0.5	0.98	1	0.985	0.875
	0.8	1	1	1	1
	1.0	1	1	1	1
2000	0.2	0.51	0.925	0.45	0.175
	0.3	0.915	1	0.88	0.58
	0.4	1	1	1	0.85
	0.5	1	1	1	0.975
	0.8	1	1	1	1
	1.0	1	1	1	1
5000	0.2	0.915	1	0.91	0.31
	0.3	1	1	1	0.83
	0.4	1	1	1	0.995
	0.5	1	1	1	1
	0.8	1	1	1	1
	1.0	1	1	1	1
10000	0.2	1	1	1	0.495
	0.3	1	1	1	0.985
	0.4	1	1	1	1
	0.5	1	1	1	1
	0.8	1	1	1	1
	1.0	1	1	1	1

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

4.4.2 Omission of the interaction term, xz

Table 4.10 shows the proportion of times each of the goodness of fit test rejects the null hypothesis based on 200 rounds of model fitting trials, in case of setting 3 under scenario 1 (See Table 4.2).

In this case, the correct model includes an interaction term xz with coefficient of β_4, whereas the wrong model is the one excluding the interaction term xz. We can see that the empirical power goes higher as the sample size increases for each test, also the parameter coefficient affects the power. For sample size of 2000, 5000, and 10000, when β_4 varies from -2.0 to -1.2, the power of each test increases; when β_4 varies from -1.0 to -0.2, the power of each test decreases. For sample size less than or at 1000, this pattern does not hold for the varied β_4.

The proposed NP test achieves higher power than other three alternative tests, that is the NP test would reject the wrong model (with missing interaction term) more often than other three test for sample size less than or at 1000. And most of the tests achieve its highest power when β_4 equals -2 or -1.2.

We expect all the tests achieve its highest power when β_4 equals -2, which is at the largest magnitude of coefficient value for the interaction term, however the result table shows that almost all tests achieve their highest power at -1.2 or -1.0 rather than -2.0. We will discuss more on why this happens in chapter 5.

Table 4.10: Setting 3 under scenario 1: Rejection rate of the omission of an interaction term

N	β_4	HL[1]	RSS[2]	CUSUM[3]	NP [4]
200	-2.0	0.23	0.405	0.335	0.345
	-1.5	0.2	0.385	0.23	0.19
	-1.2	0.12	0.415	0.19	0.21
	-1.0	0.115	0.355	0.175	0.225
	-0.5	0.075	0.155	0.06	0.095
	-0.2	0.07	0.055	0.06	0.075
500	-2.0	0.295	0.4	0.415	0.395
	-1.5	0.225	0.475	0.36	0.325
	-1.2	0.36	0.745	0.51	0.405
	-1.0	0.27	0.73	0.455	0.39
	-0.5	0.105	0.22	0.135	0.14
	-0.2	0.035	0.085	0.085	0.135
1000	-2.0	0.355	0.485	0.45	0.355
	-1.5	0.36	0.68	0.515	0.32
	-1.2	0.67	0.95	0.855	0.55
	-1.0	0.675	0.965	0.815	0.53
	-0.5	0.18	0.575	0.24	0.155
	-0.2	0.05	0.095	0.045	0.13
2000	-2.0	0.5	0.57	0.615	0.32
	-1.5	0.6	0.925	0.735	0.42
	-1.2	0.93	0.99	1	0.705
	-1.0	0.93	1	1	0.645
	-0.5	0.385	0.845	0.595	0.245
	-0.2	0.085	0.15	0.125	0.1
5000	-2.0	0.79	0.735	0.895	0.33
	-1.5	0.955	1	0.985	0.625
	-1.2	1	1	1	0.92
	-1.0	1	1	1	0.965
	-0.5	1	0.99	0.98	0.315
	-0.2	0.125	0.12	0.135	0.115
10000	-2.0	0.965	0.925	0.985	0.45
	-1.5	0.9955	1	1	0.865
	-1.2	1	1	1	0.995
	-1.0	1	1	1	1
	-0.5	1	0.99	1	0.505
	-0.2	0.28	0.42	0.335	0.115

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

Figure 4.1 through Figure 4.4 illustrates all test achieve their highest power not at the largest magnitude of the beta coefficient of the interaction term, i.e. not at $\beta_4 = -2$.

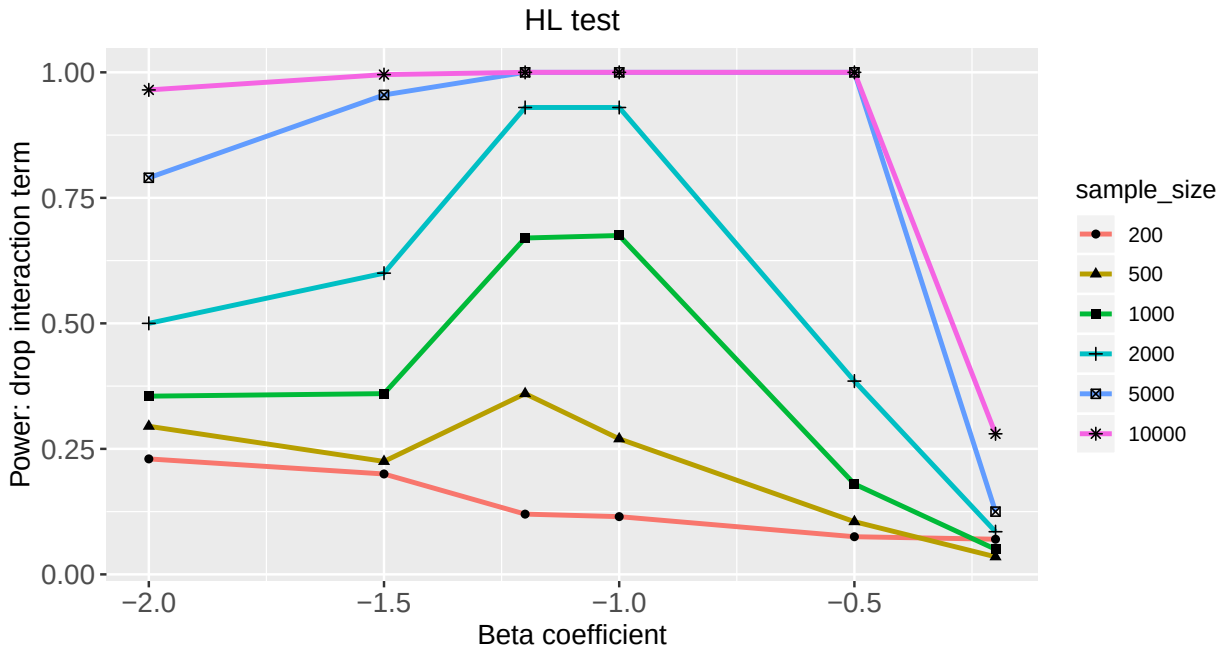

Figure 4.1: Power of HL test against beta coefficient under different sample size

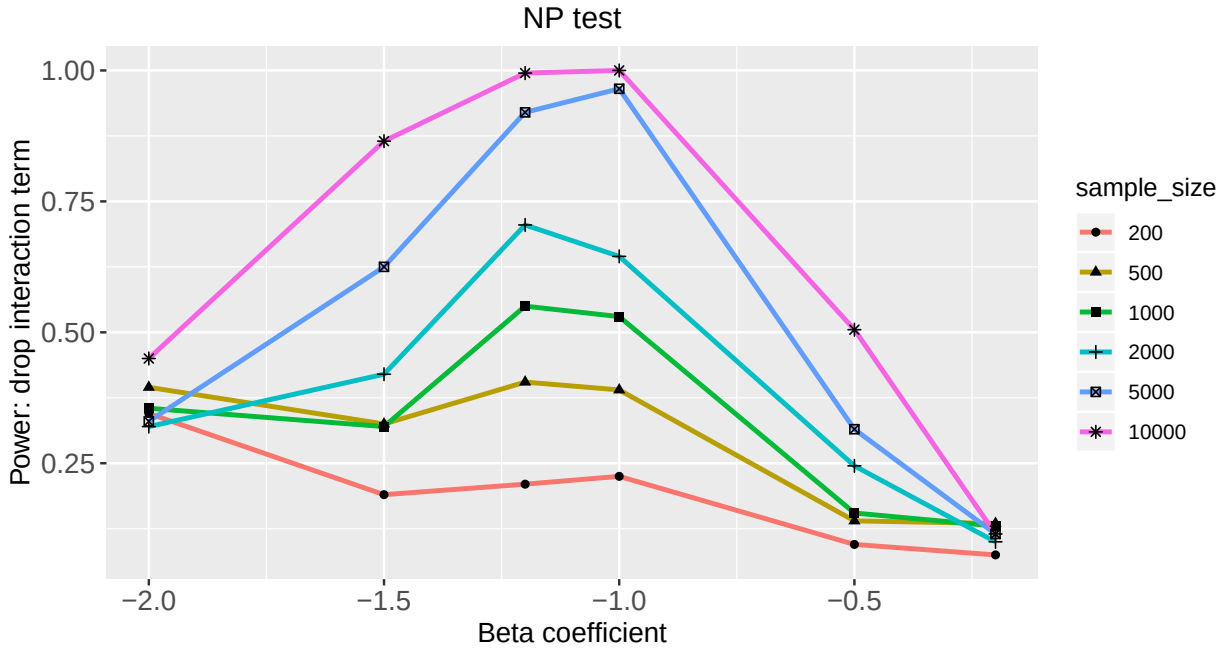

Figure 4.2: Power of NP test against beta coefficient under different sample size

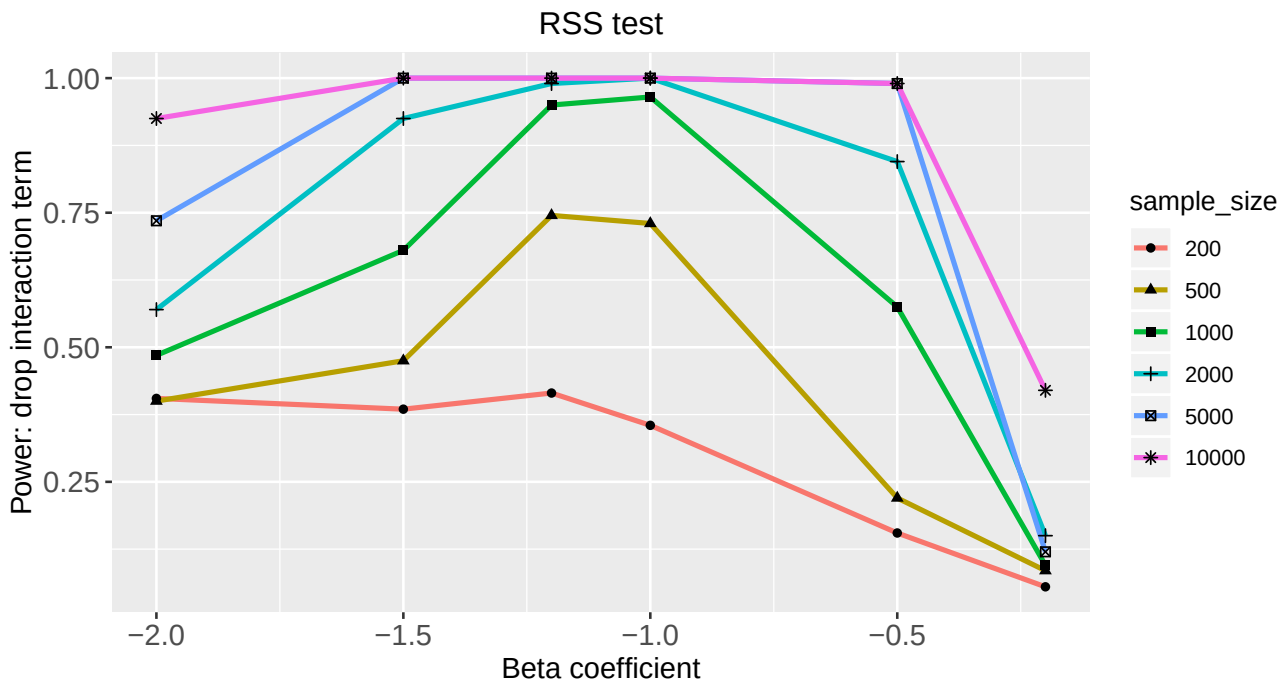

Figure 4.3: Power of RSS test against beta coefficient under different sample size

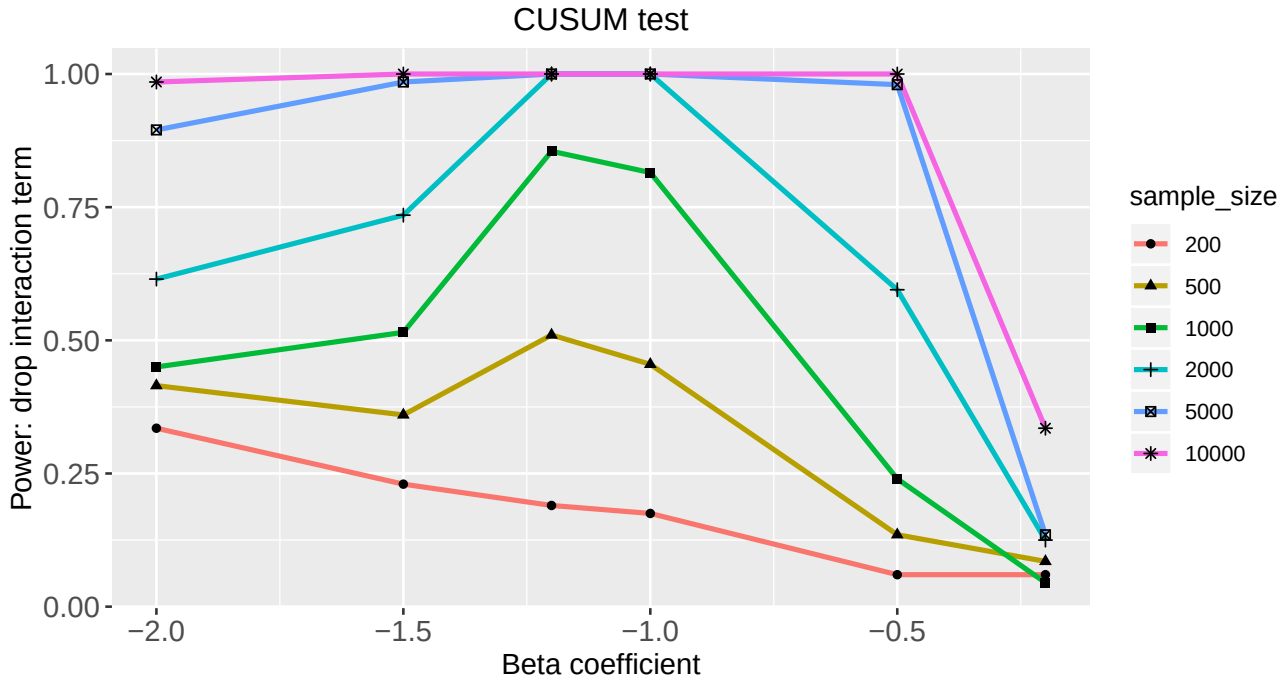

Figure 4.4: Power of CUSUM test against beta coefficient under different sample size

4.4.3 Addition of an interaction term, xz

Table 4.11 shows the proportion of times each of the goodness of fit test rejects the null hypothesis based on 200 rounds of model fitting trials, in the case of setting 3 under scenario 2, i.e. adding an interaction term to the fitted model. We get very similar results as the case that adding a quadratic term of a continuous variable to the specified model, that is, all these tests have quite low power to detect any lack of fit by adding an interaction term of a continuous covariate and a categorical variable (their main effect are included in the model), for the case in either large or small sample size.

Table 4.11: Setting 3 under scenario 2: Rejection rate of the addition of an interaction term

N	HL [1]	USS [2]	CUSUM [3]	NP [4]
200	0.035	0.085	0.1	0.105
500	0.04	0.095	0.05	0.095
1000	0.03	0.05	0.045	0.09
2000	0.05	0.035	0.05	0.055
5000	0.04	0.04	0.06	0.075
10000	0.065	0.05	0.06	0.065
15000	0.04	0.07	0.07	0.045
25000	0.06	0.075	0.065	0.07

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

4.4.4 Addition of a quadratic form of continuous variable, x^2

Table 4.12 shows the proportion of times each of the goodness of fit test rejects the null hypothesis based on 200 rounds of model fitting trials, in the case of setting 2 under scenario 2, i.e. adding a quadratic term of continuous predictor variable to the fitted model.

Unlike the situation of the omission of a quadratic term in the fitted model, it is noticeable that all these tests have low power to detect any lack of fit when adding a quadratic term of a continuous covariate, which exists (is included already) in the model, to the fitted model,

Table 4.12: Setting 2 under scenario 2: Rejection rate of the addition of a quadratic term

N	HL [1]	USS [2]	CUSUM [3]	NP [4]
200	0.065	0.055	0.03	0.06
500	0.025	0.035	0.07	0.13
1000	0.025	0.045	0.05	0.095
2000	0.045	0.06	0.065	0.1
5000	0.025	0.065	0.065	0.035
10000	0.06	0.055	0.055	0.055
15000	0.04	0.05	0.035	0.035
25000	0.025	0.07	0.055	0.035

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

no matter how large or small the sample size is.

4.4.5 Addition of an unrelated continuous predictor variable, u

Table 4.13 shows the proportion of times each of the goodness of fit test rejects the null hypothesis based on 200 rounds of model fitting trials, in the case of setting 4 under scenario 2, i.e. adding an unrelated continuous predictor variable, u, to the fitted model.

Similarly as in the previous two cases of the addition of terms to the systematic component of the model, when adding an unrelated continuous predictor variable to the systematic component of the specified model, all goodness-of-fit tests reject the wrong model at a very low rate across different distributions from which the additional continuous predictor variable is sampled, under different sample size. The highest power in this case is 12.5% achieved by the NP test.

The above results (Table 4.11 through Table 4.13) suggest that even when sample size increases up to 5,000, the test powers don't increase under the over-fitting (by adding more predictor variable to the fitted model) scenarios.

Table 4.13: Setting 4 under scenario 2: Rejection rate of the addition of an unrelated continuous predictor variable

Setting	N	HL [1]	USS [2]	CUSUM [3]	NP [4]
$u \sim N(0,1)$	200	0.055	0.035	0.06	0.075
	500	0.035	0.05	0.04	0.125
	1000	0.06	0.055	0.08	0.09
	2000	0.045	0.025	0.045	0.075
	5000	0.04	0.065	0.07	0.065
$u \sim Unif(-3,3)$	200	0.055	0.025	0.07	0.08
	500	0.045	0.05	0.065	0.11
	1000	0.04	0.055	0.035	0.085
	2000	0.06	0.05	0.055	0.085
	5000	0.035	0.025	0.055	0.085
$u \sim Beta(1,2)$	200	0.025	0.035	0.055	0.07
	500	0.045	0.02	0.05	0.105
	1000	0.055	0.05	0.045	0.08
	2000	0.04	0.055	0.045	0.05
	5000	0.045	0.025	0.05	0.06

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

4.5 Summary

We observed that when adding either a quadratic term of a continuous predictor variable or an interaction term, or an unrelated continuous predictor variable to the assumed model, which can be referred to the case of over-fitted model, the power is very low, almost the same as the low rate as the type I error rate. The result implies that all goodness-of-fit tests are not sensitive to models of over-fitting.

In cases of omission of a non-linear covariate term, i.e. an interaction term or a quadratic term of a continuous covariate, the power of all goodness-of-fit tests is affected by the magnitude of the associated parameter coefficient, the smaller the magnitude of the associated beta, the lower the test power to detect the missing of a non-linear covariate term. In these cases the test power increases along with the increasing sample size when holding all other

beta coefficients fixed.

We also observed that the performance of test power for detecting the omission of an interaction term is more complicated than that for detecting the omission of a quadratic term of a continuous covariate, one possible reason is that the former case has more variable(s) involved than the latter case. For instance, Table 4.10 shows a complicated pattern of test power. In this case, by holding other β coefficients fixed, when $\beta_4 = -1$ or $\beta_4 = -1.2$, all tests hold the highest power across all sample size settings. The test power does not achieve the highest lever when $\beta_4 = -2$, the case of the largest magnitude of β_4. On the other hand, even with sample size of 10000, the maximum power among all four tests is 42% when $\beta_4 = -0.2$, which is well below the typical desirable nominal power level, 80%.

In general we can see that the test power of detecting the omission of a quadratic term is higher than that of detecting the omission of an interaction term. Based on the simulation results, we can see that all goodness-of-fit tests can detect the missing of a quadratic term more robustly than the missing of an interaction term.

Chapter 5

Further Comparison of Empirical Size and Power of Goodness-of-fit Tests under Generalized Simulation Settings

5.1 Introduction

As we discussed in the previous chapters, even though so many tests have been published for many years to assess model fitting when sparse data is presented in logistic regression models, unfortunately it still lacks of a clear guidance for users to follow under various circumstances. At the early stage when we started our study in this field, we thought it maybe due to reasons like:

1. the published studies are too specific or relatively simple
2. under different or even slightly different scenarios, one test may perform quite differently, it is not easy to generalize the features of a specific test
3. different results from different tests is not convincing, and would confuse users
4. lack of systemic comparison studies under varied scenarios
5. It is hard to optimize a test as one-size-fits-all

We design and deploy a generalized comparative study to advance our knowledge in this field. In this chapter, we continue our comparative study through a more robust simulation framework. The difference between this study and the study in chapter 4 is the simulation

setting, i.e. we consider a relatively wide variety of situations in terms of sample size, design matrix or systematic component, parameter coefficient, under-fitting and over-fitting, and different combinations of them. Another difference lies in the strategy of simulation. In chapter 4 we randomly sample multiple data sets under one setting, in this chapter, we only draw one sample under a randomly selected situation. We describe it next.

5.2 Design of simulation study

We set up the goal of the new simulation study the same as that of chapter 4, that is to empirically assess the performance of three existing goodness-of-fit tests and the proposed chi-square test in terms of size and power. However the simulation settings of this chapter is proposed to overcome the limitations of those settings of chapter 4 and the published studies in many aspects, such as we increase the sample size of simulated data, we do not intend to fix the number of samples (e.g. $N = 100, 500, 1000, \ldots$) rather than to randomly simulate data set with sample size varying between 200 to $30,000$.

We do not fix the weight of each coefficient of the unknown parameters, in stead the design matrix is randomly sampled from continuous and categorical predictor variables.

Two correlated continuous covariates are also randomly sampled with different correlation coefficients. It is possible there is no correlated covariates in the null model as well (i.e. when the correlation coefficient equals zero).

Once the above aspects of the simulation situations are randomly selected and combined together, we are ready to simulate data set(s) under this randomly pre-determined setting. As we mentioned previously, only one data set would be randomly generated under one setting. We describe the settings in detail below.

5.3 Simulation setup

The setup for dependent variable are the same as in chapter 4, i.e., a binary outcome. The independent variables are set up as follows.

- continuous covariates: from distributions such as $N(0,1)$, $N(0,4)$, $Unif(-1,1)$, and $Beta(2,2)$
- dichotomous variables: 0 or 1 samples from Bernoulli distributions (with $\pi \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
- interaction terms: the interaction between existing continuous and dichotomous variables
- quadratic terms: quadratic form of existing continuous covariates
- two correlated continuous covariates: with correlation coefficient $\rho \in \{0.1, 0.2, \ldots, 0.9\}$
- parameter coefficients (β's): $(-4, -2, -1.5, -1, -0.75, -0.5, -0.2, -0.1, 0.1, 0.2, 0.5, 0.75, 1, 1.5, 2, 4)$
- intercept (β_0): 2 or -2
- sample size N: varied between 200 and 30,000.

A simulated data set was formed/combined by nc continuous covariates, here nc is randomly sampled from $\{3, 4, 5\}$, nd dichotomous variables and nd is randomly sampled from $\{2, 3, 4, 5\}$, ni interaction terms and $ni \in \{0, 1, 2\}$, if $ni = 0$, then there is no interaction term in the model used to simulate data, nq quadratic terms and $nq \in \{0, 1, 2\}$, similarly as ni, $nq = 0$ implies there is no quadratic from of any continuous covariate in the model, with a randomly sampled size N, here $N \in \{200, 201, \ldots, 30000\}$, either with or without two correlated covariates. For a given design matrix (the number of predictor variables are determined), the parameter coefficients are samples from β, the parameter coefficient for intercept will be a sample from $\{-2, 2\}$.

We arbitrarily divide the sample size N into five classes, specifically small class for sample size between 200 and 500, moderate for sample size between 500 and 2,000, large for sample size between 2,000 and 5,000, extra large for sample size between 5,000 and 15,000 and super large for sample size between 15,000 and 30,000. Around 2,000 data sets are simulated in each class (not exactly 2,000 is because in some cases the specified logistic regression model failed in converge and would be dropped in a batch simulation, this can be improved by writing a more effective program module in the future), therefore the total number of simulated data sets or models is over 10,000.

There were a number of different profile design matrices or systematic components available. For example, suppose the model contains 5 continuous covariates (then the number of different combination of continuous covariates is $4^5 = 1024$), 5 categorical variables (the number of different combinations of categorical variables is $5^5 = 3125$), even without considering interaction terms and quadratic terms, $3,200,000$ different combinations would be available. Therefore the data sets or models we simulated is just a very small sample from the possible models under our simulation setting.

5.4 Experiment: multiple scenarios

We assess the performance of goodness-of-fit tests just as the same as in chapter 4, i.e. investigating the tests on correct model and wrong models. By investigating the tests on correct model, we can get the estimated size of each test, by investigating the tests on wrongly specified models, we can get the estimated power of each test.

Table 5.1 lists the scenarios we investigate to assess the size and power of each of the four goodness-of-fit tests we compared in chapter 4 before. Specifically we will examine the power under six scenarios, i.e. scenario 1: omission of interaction term, scenario 2: omission of quadratic term, scenario 3: omission of a correlated covariate, scenario 4: omission of a main effect, scenario 5: addition of an interaction term, and scenario 6: addition of one unrelated continuous covariate, here "unrelated" means having nothing to do with the outcome variable.

Table 5.1: Scenarios for size and power comparison of four goodness-of-fit tests

	Scenario	Assumed/Specified model
Size (correctly specified model)	1	exactly the same as the underlying model
Power (wrongly specified model)	1	omission of interaction terms
	2	omission of quadratic terms
	3	omission of one correlated term
	4	omission of one main effect
	5	addition of one interaction term
	6	addition of one unrelated continuous covariate

5.5 Simulation Results

In this section we present all the simulation results under various scenarios.

5.5.1 Size

Table 5.2 shows the empirical size of each test under different classes of sample size of simulated data sets.

Table 5.2: Rejection rate of four goodness-of-fit tests on correctly specified models under different classes of sample size

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.075	0.072	0.078	0.088
moderate	0.071	0.098	0.07	0.076
large	0.074	0.099	0.065	0.069
extra large	0.131	0.162	0.114	0.147
super large	0.181	0.211	0.134	0.182

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

All tests cannot retain the type I error rate at the conventional desirable level of 5%. For small through large sample size (N is between 200 and 5,000), the HL test retains the type I error rate at 7.5%. The RSS test retains smaller error rate among four goodness-of-fit tests across different sample sizes except for the case of small sample size ($N < 500$), but at about 10% level on average. Similarly as we observed in chapter 4, the new partition test holds

higher type I error rate than other three tests. It is noticeable that all goodness-of-fit tests inflated type I error rates across all classes of sample size with over 10,000 different models.

What would the type I error rate look like if we go back to the traditional simulation way as introduced in chapter 4 by using the simulation method in this chapter? To make the simulation similar to the setting as introduced in the previous chapter, we modified the setting as follows:

- one continuous covariate: from distributions such as $N(0,1)$, $N(0,4)$, $Unif(-1,1)$, and $Beta(2,2)$
- one dichotomous variable: 0 or 1 samples from Bernoulli distributions
- one interaction term: interaction between continuous and dichotomous variables
- one quadratic term: quadratic form of the continuous covariate
- the pool of parameter coefficients (β's):
  $(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)$
- fixed intercept (β_0): -2
- sample size N: varied between 200 and 30,000.

Based on the above settings, the model we simulated was

$$\eta = g(\pi) = -2 + x + x^2 + z + xz. \tag{5.1}$$

We fit the correct model as $\eta = g(\pi) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz$. Table 5.3 shows the empirical size of each test under each class of sample size of $B = 500$ simulated data sets.

Table 5.3: Rejection rate of four goodness-of-fit tests on correctly specified model (5.1) under different classes of sample size

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.034	0.104	0.008	0.104
moderate	0.034	0.068	0.042	0.042
large	0.02	0.05	0.046	0.036
extra large	0.038	0.04	0.062	0.046
super large	0.04	0.054	0.052	0.058

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

We can see from Table 5.3 that in most cases the four tests retained type I error rate reasonably well just as what we have seen before in chapter 4. It is also noticeable that only the Hosmer-Lemeshow test controlled the type I error rate at 5% level across different classes of sample size. When sample size was between 200 and 500 (i.e. "small" class), the type I error rates of both the proposed new partition and the CUSUM tests are above 10%.

Due to the limitation of time, we did not explore more about the inflated type I error rate under the settings of the generalized simulation study, but it is worth of further investigation.

A possible way to investigate what factors, among the sample size N, the number of continuous covariates nc, the number of categorical variables nd, the number of interaction terms ni, the number of quadratic terms nq, correlation coefficient $ccoef$ of collinearity terms, drives the inflated type I error rate (probability of rejecting the null hypothesis when correct models are specified), we define a binary dependent variable as 1/0 for rejection/non-rejection (the rejection rule is determined by p-value of each test for every simulated data set, that is $y = 1$ if $p < 0.05$ and $y = 0$ otherwise) and fit a logistic regression model with those factors. All four goodness-of-fit tests show three factors, namely the sample size N, the number of categorical variables nd and correlation coefficient $ccoef$ of collinearity terms, are significantly associated with the rejection probability. The individual analytical results are shown in Table 5.4 through Table 5.7.

Table 5.4: logit model fitting result with HL test on rejection data (size)

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.544	0.2023	-12.576	< 0.0001
N	0.00004	0.000003	13.319	< 0.0001
nc	-0.0413	0.03997	-1.033	0.3014
nd	0.2504	0.02931	8.546	< 0.0001
ni	0.04791	0.03936	1.217	0.2236
nq	-0.07304	0.04033	-1.811	0.0701
$ccoef$	-2.364	0.1234	-19.163	< 0.0001

Table 5.5: logit model fitting result with NP test on rejection data (size)

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.769	0.1872	-14.791	< 0.0001
N	0.00004	0.000003	14.09	< 0.0001
nc	0.053	0.03672	1.444	0.149
nd	0.2307	0.02681	8.605	< 0.0001
ni	0.0086	0.03614	0.239	0.811
nq	-0.00084	0.03681	-0.023	0.982
$ccoef$	-1.836	0.1059	-17.334	< 0.0001

Table 5.6: logit model fitting result with RSS test on rejection data (size)

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.533	0.2092	-12.108	< 0.0001
N	0.00003	0.000004	9.547	< 0.0001
nc	-0.04475	0.04142	-1.08	0.28
nd	0.2228	0.03031	7.531	< 0.0001
ni	-0.02219	0.04089	-0.543	0.587
nq	-0.01801	0.04166	-0.432	0.666
$ccoef$	-1.961	0.1232	-15.919	< 0.0001

Table 5.7: logit model fitting result with CUSUM test on rejection data (size)

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.376	0.1963	-12.108	< 0.0001
N	0.00004	0.000003	12.643	< 0.0001
nc	-0.00086	0.03889	-0.022	0.982
nd	0.1975	0.02836	6.962	< 0.0001
ni	0.03066	0.03839	-0.799	0.425
nq	-0.01539	0.03907	-0.394	0.694
$ccoef$	-2.415	0.1205	-20.038	< 0.0001

5.5.2 Power

5.5.2.1 Scenario 1: omission of interaction terms

Table 5.8 shows the empirical power of each test to detect the missing of interaction terms under different classes of sample size. This pooled result table (by pooling over 10,000 model fitting results together) shows that each test does not hold desirable power at the nominal level 80% when interaction terms are omitted from the model (the null model, or the underlying model used to generate the data set includes interaction terms). For sample size varying between 500 and 15,000, the new partition method of goodness-of-fit test achieves slightly higher power than the rest of the three tests.

Table 5.8: Rejection rate of four goodness-of-fit tests: omission of interaction terms

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.093	0.065	0.099	0.092
moderate	0.1222	0.2104	0.1126	0.1156
large	0.1479	0.2516	0.1409	0.1386
extra large	0.2158	0.2559	0.2201	0.2315
super large	0.2892	0.2831	0.2816	0.2914

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

The reason why the empirical power is below the nominal power at 80% level might be due to the following possible reasons:

- the test power would depend on the magnitude of the associated parameter coefficient β's.
- the test power would also depend on what else remains in the model, especially when those remaining terms are correlated with that which is missing.
- when model changes at all aspects, in terms of number of covariates, collinearity, nonlinear parameter pattern, the associated beta coefficients, and sample size of data

sets, the test power will be affected (jointly) by many of these factors in the case of missing interaction terms.

For each of the pooled model fittings, the interaction term has its own characteristics based on its specific design matrix. As shown in Table 4.10, for some design matrices the test powers would be high, for other design matrices the test powers would be low, when different models are pooled together, the average power becomes low. Even in case of large sample size, the test power of each method would be diluted. That's the reason why the test powers are smaller than those in chapter 4, for example we keep the model (design matrix) fixed but just increase sample size, or under the same (fixed) sample size, we only change the beta coefficient of the interaction term for simulated data set. In those cases, for one specific model, the changing pattern of test power maybe easily caught, but for another model it maybe not easy to find the characteristics of the test power. We will illustrate with an example later.

Therefore if model changes at all aspects as mentioned above, when we collect evidence from pooled model fitting results (p-values), we need a much larger sample size to detect the missing of interaction term to achieve a desirable power level. In other words, the conventional power level of 80% is hard to achieve, instead we maybe interested in simply comparing the test power of different tests.

Similar as we did before, to investigate what factors, among the sample size N, the number of interaction terms ni, the number of categorical variables nd, affect the probability of rejecting the null hypothesis across all models, we define a binary dependent variable as 1/0 for rejection/non-rejection, then we fit logistic regression model with those factors, i.e. regressing on N, ni, and nd. All four goodness-of-fit tests show those factors are significantly associated with the rejection probability when interaction terms are missed. The individual analytical results are shown in Table 5.9 through Table 5.12 below.

Let's go back to chapter 4 for the test power of missing interaction term. In chapter 4, we found the test power for detecting the mission of interaction term did not change as expected

Table 5.9: Scenario 1: logit model fitting result with HL test

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-3.014	0.1521	-19.814	< 0.0001
N	0.00005	0.0000034	15.302	< 0.0001
nd	0.1052	0.0291	3.619	0.0003
ni	0.4129	0.0655	6.308	< 0.0001

Table 5.10: Scenario 1: logit model fitting result with NP test

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.332	0.1369	-17.033	< 0.0001
N	0.00004	0.0000033	10.977	< 0.0001
nd	0.1457	0.02673	5.451	< 0.0001
ni	0.1309	0.5977	2.19	0.0286

Table 5.11: Scenario 1: logit model fitting result with RSS test

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.93	0.152	-19.277	< 0.0001
N	0.00005	0.0000034	14.526	< 0.0001
nd	0.09363	0.02915	3.212	0.00132
ni	0.3877	0.06562	5.909	< 0.0001

Table 5.12: Scenario 1: logit model fitting result with CUSUM test

	Estimate	S.E.	Z	Pr(>|z|)
Intercept	-2.75	0.15	-18.334	< 0.0001
N	0.00005	0.0000034	15.87	< 0.0001
nd	0.09653	0.02902	3.326	0.0009
ni	0.2516	0.0651	3.864	0.0001

with varied beta coefficient (see Table 4.10). Here we conduct two more simulations to reveal what factors affect test power except for the beta coefficient for the interaction term. We follow the same simulation strategy as in chapter 4 for these two more simulations, i.e. 200 samples under one specified simulation setting.

Firstly, we set all the coefficients of parameter xz, the interaction term, as positive values rather than all are negative values as in setting 3 scenario 1 in chapter 4. Table 5.13 presents the new setting of β_4 and new rejection rates.

As we can see from Table 5.13, under the same sample size, when β_4 increases from 0.2 to

2.0 (specifically β_4 taking values as 0.2, 0.5, 1.0, 1.2, 1.5, 2.0), the power of each test increases and the highest power obtained when β_4 equals 2, the largest magnitude of beta coefficient, as expected. This trend holds under different settings of sample size. This changing pattern of rejection rate become different than that as shown in Table 4.10 in chapter 4. In chapter 4, β_4 decreases from -2.0 to -0.2 (i.e. β_4 taking values as -2, -1.5, -1.2, -1.0, -0.5, -0.2), and its sign is different than the sign of coefficients for both x and z. These two tables (Table 4.10 & Table 5.13) suggest that not only the magnitude but also the sign of parameter coefficient of the interaction term would affect the power of goodness-of-fit tests.

Table 5.13: Setting 3 under scenario 1 of chapter 4: Rejection rate of detecting the omission of interaction term with a set of different values of β_4

N	β_4	HL[1]	RSS [2]	CUSUM [3]	NP [4]
200	0.2	0.035	0.035	0.03	0.11
	0.5	0.065	0.165	0.075	0.09
	1.0	0.165	0.265	0.145	0.085
	1.2	0.175	0.355	0.14	0.06
	1.5	0.25	0.45	0.15	0.12
	2.0	0.45	0.58	0.215	0.19
500	0.2	0.07	0.06	0.08	0.1
	0.5	0.085	0.26	0.1	0.12
	1.0	0.285	0.525	0.23	0.22
	1.2	0.36	0.685	0.305	0.24
	1.5	0.595	0.78	0.41	0.345
	2.0	0.8	0.915	0.57	0.475
1000	0.2	0.035	0.135	0.075	0.095
	0.5	0.115	0.445	0.175	0.14
	1.0	0.55	0.875	0.37	0.335
	1.2	0.665	0.91	0.525	0.33
	1.5	0.845	0.995	0.66	0.42
	2.0	0.975	0.995	0.905	0.655
2000	0.2	0.05	0.185	0.09	0.08
	0.5	0.325	0.665	0.24	0.09
	1.0	0.83	0.975	0.68	0.37
	1.2	0.965	1	0.855	0.515
	1.5	0.99	1	0.97	0.76
	2.0	1	1	1	0.935
5000	0.2	0.09	0.335	0.14	0.095
	0.5	0.705	0.98	0.67	0.12
	1.0	1	1	0.985	0.665
	1.2	1	1	1	0.865
	1.5	1	1	1	0.98
	2.0	1	1	1	1
10000	0.2	0.225	0.645	0.2	0.045
	0.5	0.97	1	0.935	0.22
	1.0	1	1	1	0.9
	1.2	1	1	1	0.99
	1.5	1	1	1	1
	2.0	1	1	1	1

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] Copas's Unweighted Residual Sum of Squares Test
[3] Cumulative Sums of Residuals Test
[4] New Partition Chi-square Test

Secondly we conduct another simulation with β_4 increases from -1.5 to 1.5 by 0.5, here β_4 is the coefficient parameter for the interaction term as before, but we change other beta coefficients meanwhile, specifically we use three different models for this simulation study and we focused on the change of rejection rate under different sample size using the HL test.

The three models are under the same form of design matrix but with different beta coefficients. The design matrix is also the same as in the first additional simulation. Thus the systematic component is in the form as follows.

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{1} & \boldsymbol{x} & \boldsymbol{x}^2 & \boldsymbol{z} & \boldsymbol{xz} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

Then we define three different models by different settings of $\boldsymbol{\beta}$ respectively. For model 1, we set $\boldsymbol{\beta} = (-2, 1, 0.2, 1, \beta_4)^T$, for model 2, we set $\boldsymbol{\beta} = (1, 1, 1, 1, \beta_4)^T$, and for model 3, we set $\boldsymbol{\beta} = (1, -2, 1, -2, \beta_4)^T$, and for all three models we set six different values for β_4 as previously described $\beta_4 \in \{-1.5, -1, -0.5, 0.5, 1, 1.5\}$, i.e. β_4 takes three pairs of values with opposite signs.

Bearing in mind the purpose of this simulation is to investigate how different systematic component would change the rejection rate when an interaction term is missed under varied sample sizes. Table 5.14 shows the rejection rate of the HL test based on three different simulation models.

Table 5.14: Setting 3 under scenario 1: Rejection rate of detecting the omission of interaction term with a set of different values of β_4

N	β_4	Model 1 [1]	Model 2[2]	Model 3 [3]
200	-1.5	0.2	0.05	0.11
	-1.0	0.115	0.07	0.08
	-0.5	0.075	0.065	0.05
	0.5	0.065	0.03	0.01
	1.0	0.165	0.06	0.075
	1.5	0.25	0.045	0.05
500	-1.5	0.225	0.07	0.27
	-1.0	0.27	0.1	0.16
	-0.5	0.105	0.045	0.085
	0.5	0.085	0.045	0.05
	1.0	0.285	0.075	0.06
	1.5	0.595	0.08	0.08
1000	-1.5	0.36	0.14	0.58
	-1.0	0.675	0.13	0.275
	-0.5	0.18	0.015	0.115
	0.5	0.115	0.035	0.04
	1.0	0.55	0.03	0.14
	1.5	0.845	0.12	0.13
2000	-1.5	0.6	0.22	0.935
	-1.0	0.93	0.085	0.675
	-0.5	0.385	0.07	0.175
	0.5	0.325	0.01	0.105
	1.0	0.83	0.055	0.235
	1.5	0.99	0.235	0.295
5000	-1.5	0.955	0.575	0.995
	-1.0	1	0.21	0.985
	-0.5	1	0.06	0.455
	0.5	0.705	0.02	0.26
	1.0	1	0.205	0.71
	1.5	1	0.68	0.795
10000	-1.5	0.9955	0.955	1
	-1.0	1	0.49	1
	-0.5	1	0.09	0.79
	0.5	0.97	0.1	0.62
	1.0	1	0.435	0.965
	1.5	1	0.955	0.985

[1] In Model 1, $\boldsymbol{\beta} = (-2, 1, 0.2, 1, \beta_4)^T$
[2] In Model 2, $\boldsymbol{\beta} = (1, 1, 1, 1, \beta_4)^T$
[3] In Model 3, $\boldsymbol{\beta} = (1, -2, 1, -2, \beta_4)^T$

It is noticeable that smaller magnitude of beta coefficient (β_4) of the missing interaction term achieves lower test power for a same model under a fixed sample size. For the largest magnitude of β_4, Model 1 achieves its largest power only when $\beta_4 = 1.5$, Model 3 achieves its largest power only when $\beta_4 = -1.5$, whereas Model 2 presents a different pattern. For model 2 with sample size of 200 and 500 , the test power achieves its largest when β_4 equals -1.0, for the rest cases of sample size, Model 1 achieves its largest power when $\beta_4 = 1.5$ or $\beta_4 = -1.5$.

For a a specific β_4, all the test powers increase as sample size increases within a model, this is true for all three models. This result confirmed that sample size affects test power regardless the model components.

Another noticeable result is the test power of model 2 is well below that of other two models, it's almost always true across different sample size scenarios. Even when sample size increases to 10,000, the test power of model 2 is less than 10% whereas the test power for model 1 is 100% and model 3 is 79% when β_4 is -0.5, which suggests that systematic component affects the rejection rate more than sample size does. This simulation study helps us understand why the test powers in the studies of this chapter are much lower than those in chapter 4, that is we mix test powers from different systematic components and the test powers get diluted.

Thus far, we revealed that not only the beta coefficient (including both magnitude and sign) and sample size affect the rejection rate of the HL test for missing interaction term, but also different systematic component affects the rejection rate through simulations. In addition, our study suggests that the beta coefficients for other related components remained in the model affect the test power as well. All those aspects of data characteristics or underlying model profiles affect the performance of goodness-of-fit tests in terms of the power of detecting inadequate models.

5.5.2.2 Scenario 2: omission of quadratic terms

Table 5.15 shows the empirical power of each test to detect the missing quadratic terms under different classes of sample size. Similarly as the result shown in Table 5.8, the rejection rate shows that each test does not hold desirable power at the nominal level 80% when quadratic terms are omitted from the model from our simulation study. For sample size less than 5,000, the new partition method of goodness-of-fit test achieves higher power than other three tests. For sample size greater than 5,000, the cumulative sums of residual test achieves higher power than other tests.

Table 5.15: Rejection rate of four goodness-of-fit tests: omission of quadratic terms

	HL [1]	NP [2]	RSS[3]	CUSUM [4]
small	0.1376	0.166	0.1635	0.1585
moderate	0.2747	0.4025	0.2656	0.2995
large	0.4237	0.4687	0.3618	0.4164
extra large	0.5282	0.5453	0.4733	0.5475
super large	0.6081	0.5763	0.5318	0.6328

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

We also investigate what factors affect the test powers by conducting logistic regression model fitting as in scenario 1, all four goodness-of-fit tests suggest sample size N and the number of quadratic terms nq are the significant factors. To save pages we summarize the analytical results in the sign of estimated beta's (indicating the effect direction of predictor variables) and significant indication (the symbol $\star$ within parenthesis indicates statistical significance).

Table 5.16: Scenario 2: logit model fitting results

	N	nq
HL	+ ($\star$)	+ ($\star$)
NP	+ ($\star$)	+ ($\star$)
RSS	+ ($\star$)	+ ($\star$)
CUSUM	+ ($\star$)	+ ($\star$)

5.5.2.3 Scenario 3: omission of one correlated term

Table 5.17 shows the empirical power of each test to detect the missing correlated continuous covariate under different classes of sample size. For sample size less than 500, the CUSUM test has higher power than all other three tests, for sample size between 500 and 5000, the proposed NP test has high power than others, and the RSS test has higher power than other three test when sample size is larger than 5,000. Just as when dropping the interaction terms, all test powers in this scenario are lower than that when dropping quadratic terms.

Table 5.17: Rejection rate of four goodness-of-fit tests: omission of one correlated term

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.0476	0.0422	0.0529	0.0743
moderate	0.0393	0.195	0.0658	0.0646
large	0.0482	0.177	0.0923	0.0615
extra large	0.0765	0.1984	0.2201	0.1356
super large	0.1556	0.1729	0.3354	0.2376

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

We also investigate which factor(s) would affect the test power similarly as fitting logistic regression model as before. The analytical results show that the sample N and the beta coefficient of the correlated term are positively associated with rejection rate, and ρ, the correlation coefficient between two correlated predictor variables, is negatively associated with rejection rate, but there is an exception, the correlation coefficient is not significantly associated with rejection rate for the new partition test. The summary table for four tests is followed. Note that the results in Table 5.18 holds no matter missing which one among the two correlated covariates.

Table 5.18: Scenario 3: logit model fitting results

	N	ρ	β
HL	+ ($\star$)	- ($\star$)	+ ($\star$)
NP	+ ($\star$)	-	+ ($\star$)
RSS	+ ($\star$)	- ($\star$)	+ ($\star$)
CUSUM	+ ($\star$)	- ($\star$)	+ ($\star$)

5.5.2.4 Scenario 4: omission of one main effect

Table 5.19 shows the empirical power of each test to detect the missing main effect from the model under different classes of sample size. For sample size less than 15,000, the NP test has higher power than other three tests, for sample size greater than 15,000, the RSS and CUSUM test perform better to detect the missing term of the model.

Table 5.19: Rejection rate of four goodness-of-fit tests: omission of one main effect

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.0595	0.0864	0.06	0.0785
moderate	0.0653	0.1973	0.0828	0.0938
large	0.0918	0.2109	0.1349	0.1173
extra large	0.1695	0.2622	0.2492	0.2307
super large	0.2668	0.2723	0.3351	0.3223

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

In a similar approach as before, the analytical results show that the sample N and the beta coefficient of the main effect are positively associated with rejection rate for three test, namely the HL, NP and CUSUM test, but not the RSS test. The summary table for four tests is followed.

Table 5.20: Scenario 4: logit model fitting results

	N	β
HL	+ ($\star$)	+ ($\star$)
NP	+ ($\star$)	+ ($\star$)
RSS	+	+
CUSUM	+ ($\star$)	+ ($\star$)

5.5.2.5 Scenario 5: addition of one interaction term

Table 5.21 shows the empirical power of each test to detect the addition of an interaction term to the model under different classes of sample size. For sample size less than 500, all test powers are low. For sample size greater than 500, the NP test has higher power than other three tests, which suggest the NP test can detect the additional interaction term more often than other three tests.

Table 5.21: Rejection rate of four goodness-of-fit tests: addition of one interaction term

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.0672	0.0755	0.0826	0.087
moderate	0.0701	0.1745	0.0728	0.0797
large	0.0751	0.2646	0.0682	0.0682
extra large	0.1275	0.2399	0.1179	0.1418
super large	0.1764	0.2231	0.1415	0.1803

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

5.5.2.6 Scenario 6: addition of one unrelated continuous covariate

Table 5.22 shows the empirical power of each test to detect the addition of an unrelated continuous covariate to the model under different classes of sample size. Similarly as in scenario 5, for sample size greater than 500, the NP test has higher power than other three tests, which suggest the NP test can detect the additional continuous covariate more often than other three tests.

5.5.3 Summary

Scenarios 1 through 4 are under-fitting cases, which means the assumed model misses some terms while they are included in the model used to generate the simulated data set. Whereas scenarios 5 through 6 are over-fitting cases, which means the assumed model contains some

Table 5.22: Rejection rate of four goodness-of-fit tests: addition of one unrelated continuous covariate

	HL [1]	NP [2]	RSS [3]	CUSUM [4]
small	0.0704	0.0731	0.083	0.0747
moderate	0.06822	0.1818	0.0691	0.0751
large	0.0761	0.2683	0.065	0.0701
extra large	0.1321	0.2445	0.1156	0.1436
super large	0.1794	0.2314	0.1445	0.1857

[1] Hosmer-Lemeshow's $\hat{C}$ Test
[2] New Partition Chi-square Test
[3] Copas's Unweighted Residual Sum of Squares Test
[4] Cumulative Sums of Residuals Test

terms while they are not included in the model used for data simulation. Overall the powers of all four tests for under-fitting cases are higher than that for over-fitting cases. The proposed NP test has higher power than other three tests for over-fitting cases, which means the NP test can detect the additional terms to the model more often than the alternatives.

All four goodness-of-fit tests achieve the highest power or rejection rate in the case of missing quadratic terms (scenario 2). The highest power is over 63% by the CUSUM test. This suggests that the goodness-of-fit test tend to detect the missing of quadratic terms easier than the missing of other terms, for example, the missing of interaction terms, the missing of a main effect term, and/or the missing of a correlated term for a continuous covariate.

The sample size plays an important role in rejection rate, the reason is that the outcome from all goodness-of-fit tests are p-values. Just as the general issue with p-value outcome from a hypothesis test, a test is able to detect even very small variation from the null model when increasing sample size to large enough. It is notable that for the proposed new partition method, its rejection rate does not always strictly increase as sample size increases, the potential reason is that the degrees of freedom of this test are varied for different data sets, as a consequence, the p-value of this test is determined differently for different models. However, the trend of increasing rejection rate still remains if all situations are kept the same with sample size as the only one changing factor as shown in the chapter 4.

Both the magnitude and the sign of a parameter coefficient affect the rejection rate of the lack-of-fit test for detecting the omission of interaction terms. This feature makes the usefulness of goodness-of-fit test to detect the missing interaction terms not powerful in real data analysis.

Chapter 6

Computing Considerations

6.1 Introduction

It is always a good idea to simulate data to investigate some important characteristics of a new statistic. On the one hand, we can simulate data from some known distribution and make it close to the observed data set, on the other hand, simulation can be computationally intensive in case a model fitting or some specific test processing is needed. In the computing process of the present research, the model fitting process is time consuming for large sample size, for example, when the sample size is beyond 10,000. It is a challenging task to ensure software program can run efficiently and can shorten the computing time sufficiently. Nowadays, the parallel computing system can help us reduce the computer time. Many computers have multiple processors, making it possible to split a simulation task in many smaller, and hence faster, sub-simulations parallelly. Based on our experience in working on this research project, R offers a great computing environment to speed up our computing needs. Fortunately, R possesses the potentiality to speed up the program process by parallel computing.

6.2 An example: The need of speeding up computation

Let's look into the real computing time through the simplified model as shown in setting 2 under scenario 2 in chapter 4. We need to repeat the program 200 times, which means we simulate 200 different data sets from the same distribution specified within R program.

N	Elapsed Time
200	01s
500	01m20s
1000	02m35s
2000	05m52s
5000	20m14s
10000	57m52s
15000	01h51m27s
25000	05h12m45s

Obviously it is not a linear relationship between the elapsed time and the sample size. If we increase the number of replications or if the model becomes complicated, i.e. with more complicated design matrix or more predictor variables are involved, the computing time becomes much longer than the illustration here. It is possible that the computer would be out of memory when a complicated model fits to data with large sample size. Thus when doing simulation with large sample size, we need to consider speeding up the software package and writing efficiently fast programming code.

Not only is the long time computation due to large sample size, but also due to the algorithm here utilized to calculate the p value for the CUSUM test. As we discussed in chapter 2, the computation of goodness-of-fit tests based on the cumulative sums of residuals is time consuming.

As introduced in chapter 2, the cumulative sums of residuals is built on partitions of "some space", which means the ordering of y_i may be determined by that of the fitted $\hat{y}_i$ or of the values of a covariate is required. For example, if the partition space is based on covariate patterns, say $\boldsymbol{X}$, to compute $I(\boldsymbol{X_i} \leq \boldsymbol{t})$, the indicator function in $W_n(\boldsymbol{t})$, we need to consider all the possible combinations of $\boldsymbol{t} = (t_1, \ldots, t_p)^T$ in multidimensional space. For overall assessment of goodness-of-fit test, the ordering of observations with outcome values y would be based on that of the predicted values $\hat{y}$, in which case the space dimensionality of $\boldsymbol{t}$ is much higher when $n \gg p$, especially when n is large. An estimated p-value for the supremum test G_g in (2.24) can be estimated by generating a large number of realizations from $W_g(\boldsymbol{t})$ through Monte-Carlo simulation [30] [43] [44].

Next, we illustrate why the SUCUM test takes much longer time to calculate a p-value than alternatives with R software.

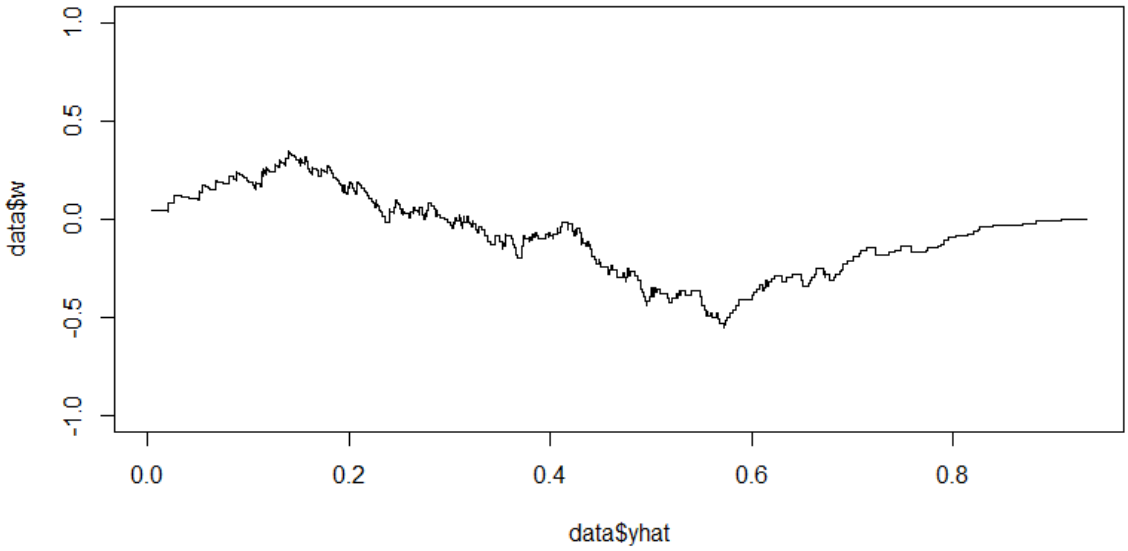

Figure 6.1: The SUCUM test process: observed W

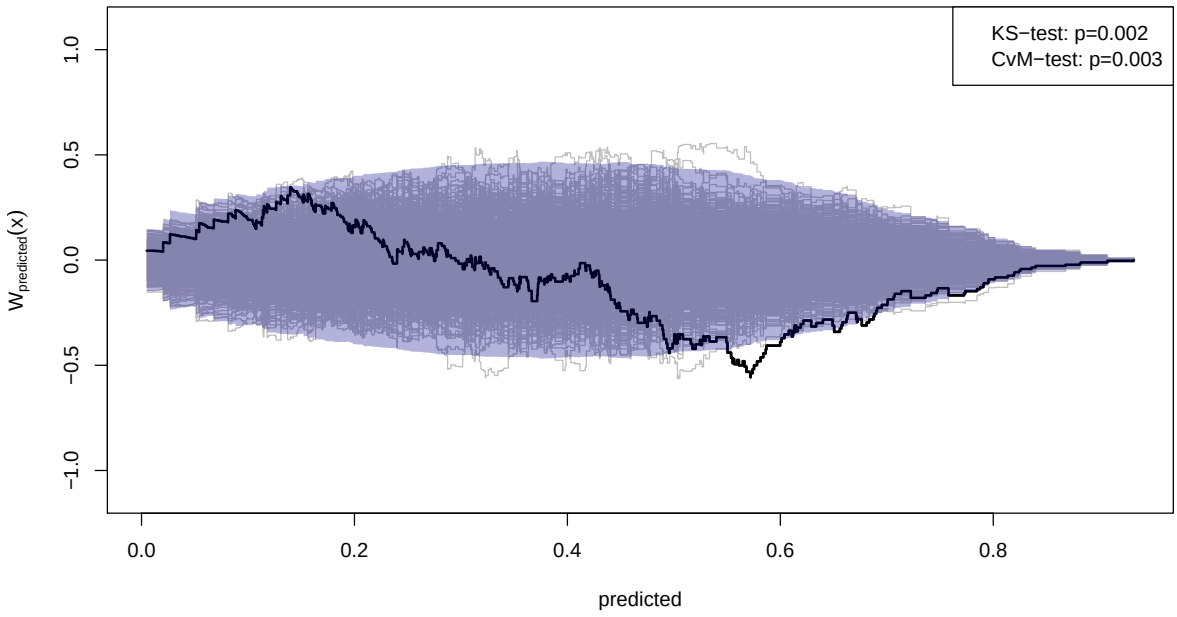

Figure 6.2: The SUCUM test process: observed W with simulated W

To investigate the algorithm of the CUSUM test to estimate an empirical p value, Figure 6.1 shows the observed distribution of W_n and the observed supremum of W_n, $g_n = 0.5569$, in one (simulation) case as shown in Table 4.9 with ($\beta_2 = 0.5$, $N = 500$) in setting 2 under scenario 1 in chapter 4. The goal of this setting is to assessing the performance of four goodness-of-fit tests to detect the omission of a quadratic form of a continuous covariate with a coefficient value of 0.5 in the null model.

```
> (g_n<-max(abs(cumsum(data$err[1:N])))/sqrt(N))
[1] 0.5569097
```

To compute $Pr(\widetilde{G}_n \geq g_n)$, let $\boldsymbol{\beta}$ and $\mathfrak{I}$ in (2.16) and (2.17) be replaced with their observed values, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathfrak{I}}$, respectively. then based on $B = 1000$ randomly repeated sets of samples $\{Z_1, \ldots, Z_n\}$ from $N(0, 1)$, the distribution of $\widetilde{G}_n$ can be estimated, further the p-value can be estimated by the proportion of the number of cases with $\widetilde{G}_n \geq g_n$ out of $B = 1000$ replications. After one simulation in this way, we obtain the estimated p-value from the Kolmogorov-Smirnov type of test as $p = 0.002$, i.e. two cases among these 1000 simulations have their $\widetilde{G}_n$'s greater than g_n. This is shown in Figure 6.2, where the 1000 sampling distributions of W_n are shown by grey curves, and the observed distribution of W_n is shown by black curve, the $p = 0$ from the Cramer-von Mises type test is also shown there.

Due to the indirect and simulation way to calculating p value, we would expect that the computing time for the CUSUM test should be longer than the alternatives, especially when the assumed model is more complicated than this simplified one. Based on our experience, it is not a rare issue that the Windows operation system would be crashed when we run R program for the above tests with complex design matrix in the model and with large sample size simulations.

In R, we can check which two simulations return the largest $\widetilde{G}_n$'s with the example code below.

```
> what.abs<-abs(what)
> G_n<-apply(what.abs, 2, max)
> sum(G_n>=g_n)/B
[1] 0.002
> which(G_n>=g_n)
[1] 298 992
> G_n[which(G_n>=g_n)]
[1] 0.5940841 0.5637542
```

The resulting p-value of 0.002 suggests an alternative model should be considered. By running this module repeatedly, we can estimate the power of the CUSUM test in this case, based on how likely in proportion the resulting p values are less than 0.05, the nominal significant level of a hypothesis test.

6.3 Monitoring the progress of computation

There are two modes to program in R. One mode is interactive programming, the other is batch processing. For batch processing, a series of programs or only one task on a computer environment is executed without manual intervention, whereas an interactive session accepts input from human. Interactive session are usually used so that we can test our program code before attempting a long production run as a batch job. It is straight forward to monitor the progress by adding a "cat()" function to track the replicates if an index of the replications are setup, or by displaying the progress bar as the screenshot shows below (Figure 6.3), or by adding "Sys.time()" function directly before and after some complex computing tasks, or use "system.time()" function outside of an appropriate chunk of R program to simply just check the processing time. We will show an example later.

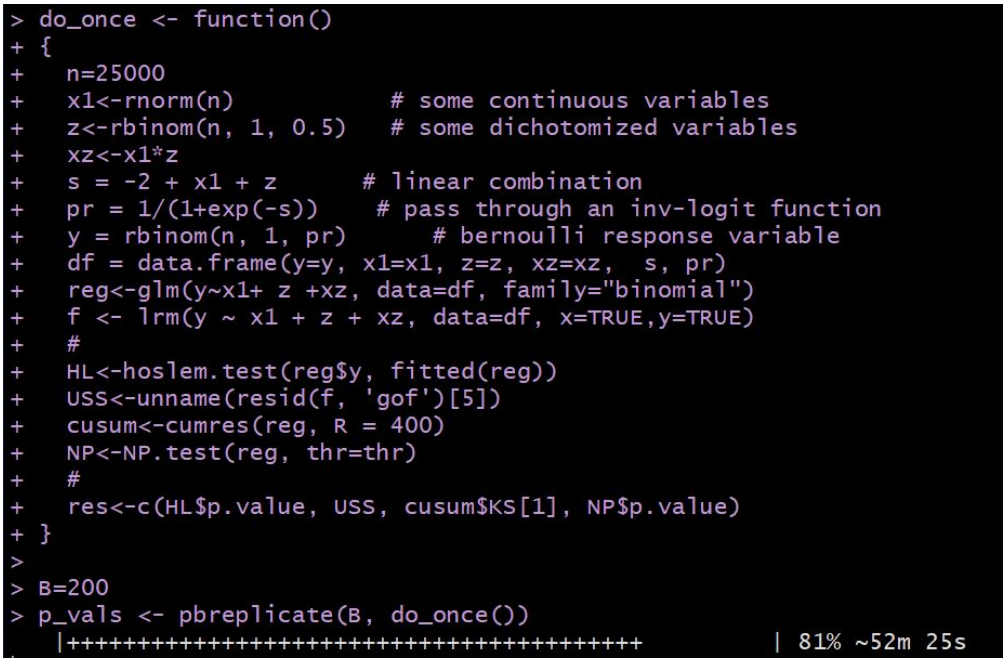

Figure 6.3: R computing progress shown in percentage completed and computing time left

The benefit of monitoring the program progress is that we can estimate the total computing time for some specific tasks. Then at the time when we need to submit a batch job, we can request a relatively precise and reasonable wall-clock time duration of a computer or high performance clusters.

6.4 High performance computing

In theory, it is not difficult to copy your R simulation code into as many files as necessary to run parallel processing manually, but as the number of repetitions becomes larger, this task becomes increasingly tedious. On the other hand, R is single-threaded program rather than parallel, which means no matter how many CPU cores are available, R can only use one of them by default. High performance computing clusters are usually equipped with a great amount of cores (Note: cluster=multiple nodes (or servers) $\times$ multiple cores per node). Fortunately we can run R program using multiple cores with the help of packages such as **multicore** and **snow** (now they are integrated into **parallel** package) to process parallel computing.

The present research project utilities the **parallel** package to perform parallel computing through "mclapply()" function. This package is designed to use multiple cores in an unix environment. The advantage of using **parallel** remains in at least two folds: (1) parallel computing rather than linearity or consequence computing, (2) overcoming the limitation of memory size, it usually exists in Windows operation systems. Users can use it in Windows systems by setting the option parameter "mc.cores" equals 1, but it loses its advantage of processing program with multiple cores simultaneously. However, the "mclapply()" function can not use cores on multiple nodes, instead it can only use multiple or all CPU cores on one node. To enhance th computing with R, the package **doParallel** is built on top of packages **parallel**, **foreach**, and **iterator**, and it can run one multiple nodes by functions "makeCluster() → registerDoParallel() → %dopar% → stopCluster()".

There are also two ways to submit batch jobs to a computer or a cluster. One way is to type in command line as below. The outcome will be saved in the user specified file. For example, "result.txt" is the output file when the job is done.

```
R CMD BATCH test.R result.txt
```

However this batch method is also a trial version to some extend. If the processing time is longer than the allowed grace period of time, or if the program requests a good amount of facility cores, the job will be stopped by the facility system automatically. A safer and better way to do batch submission is to submit a job request to the high performance computing cluster with a script. An example .slurm script is shown below.

```
#!/bin/bash
#SBATCH --job-name=test
#SBATCH --partition=yourgroup
#SBATCH --nodes=2
#SBATCH --tasks=1
#SBATCH --cpus-per-task=8
#SBATCH --mem=4G
#SBATCH --time=4-11:59:00
#SBATCH --output=test.%j.out
#SBATCH --mail-user=youraccount@kumc.edu
#SBATCH --mail-type=ALL

echo "== Submit dir. : ${SLURM_SUBMIT_DIR}"
echo "== Starting run at $(date)"
echo "== Job ID: ${SLURM_JOBID}"
echo "== Node list: ${SLURM_NODELIST}"
echo "== Running R script: test"
module load R/3.4
Rscript /panfs/pfs.local/scratch/yourgroup/youraccount/test.R
echo "== Calculation ended at:`date`"
```

As a conclusion, to secure simulation work as in our study cases, utilizing the high performance computing facility is necessary.

6.5 R package

It is a very common task to simulate binary outcomes based on some specified model forms in practice. We design our computing programs for very general purpose so that they are flexible enough for future reproducible research work. For instance, it is straight forward to simulate binary outcomes for a logistic regression with customized parameter coefficients and specific distributions for predict variables by using our simulation function as shown in Appendix A. Eventually we can extend our programs as a R package for public use in the future.

Chapter 7

Overall Summary and Discussion

7.1 Overview of chapters

We introduce the logistic regression under the framework of generalized linear regression models and the background of goodness-of-fit testing for logistic regression model in chapter 1.

In chapter 2, we conduct a literature review of existing goodness-of-fit tests developed for logistic regression model specifically when sparse data presented, we introduce different ideas to formulate different test statistics, and we also address the advantages and drawbacks of those tests.

Chapter 3 proposes a new partition based chi-square type goodness-of-fit test as an alternative to the Hosmer-Lemeshow test. It is motivated by the necessary conditions for applying asymptotic theorem to the limiting distribution of the grouped goodness-of-fit test. The proposed new partition test is intend to tackle the issue of small expected cell frequency (commonly the threshold value is 5) within some bins, which can occur in the Hosmer-Lemeshow test. The limiting distribution of the proposed chi-square type test statistic, including its degrees of freedom, is addressed in detail.

Chapter 4 compares the new partition test to the alternative three tests, namely the Hosmer-Lemeshow (HL) test, the unweighted residual sum of squares (RSS) test and the cumulative sums of residuals (CUSUM) test. The results of test size suggest that the HL retain the type I error rate better than three alternatives in general. The other three goodness-of-fit tests perform closely to each other. The proposed NP test holds slightly high type I error

rate when sample size is less than 2,000, however when sample size is greater than 5000, it controls type I error rate at the desired level of 5% level, and it holds slightly lower type I error rate than the RSS and the CUSUM tests. The test power for detecting the missing quadratic term is affected by both the magnitude and the sign of corresponding beta coefficient, but other related components in the model would affect the test power as well. The proposed NP test achieves higher power than other three tests for sample size less than 1000 under our study settings, however, the proposed new partition test is relatively conservative to detect the missing quadratic term(s). it has smaller power/rejection rate than other three tests with a specific parameter coefficient across all sample size. Comparing to the under-fitting situations, all tests are not sensitive to detect the over-fitting situations under simulation settings of our study.

Chapter 5 further compares the performance of four goodness-of-fit tests under generalized simulation settings. Due to the complicated combined form of systematic components in logistic regression models, both the type I error rate and power of four tests shifted to undesired direction, i.e. the type I error rates increased and the test powers decreased among four goodness-of-fit tests. This generalized simulation study shows under some scenarios the power of the HL test for detecting the missing interaction term is less than 10% even when sample size is up to 10,000. The study also shows that all test achieve higher power to detect the omission of quadratic terms than the omission of other terms, such as the omission of interaction terms, correlation term, and a main effect term even. The over-fitting situations, such as adding a main effect, adding a quadratic term, adding interaction term(s), are not detected as often as the under-fitting cases by four tests.

Chapter 6 addresses the computing consideration. Practically when conducting the assessment of the lack-of-fit by using the CUSUM test, we need to calculate p-values by simulation or sampling, this type of algorithm (using a proportion value as p-value) set up a challenging situation for R program. It is often the case a R session would be aborted due to using up the computer memory. This requires the use of a high performance cluster com-

puting facility to overcome this barrier, especially when data with large sample size needs to be simulated.

7.2 Discussion

Just as Royle et al. (2014) pointed out, conducting a goodness-of-fit test is not always easy to do. And, moreover, it is never really easy (or especially convenient) to decide if a goodness-of-fit test is worth anything [70]. Despite this we try to find some insights of this research field.

7.2.1 Contribution of this research

The Hosmer-Lemeshow goodness-of-fit statistic is widely used for evaluating the fit of logistic regression models when continuous covariates are presented. Usually, the HL test statistic is calculated using the deciles-of-risk grouping method by forming groups based on the predicted probabilities. Under this method, the group boundaries are determined by referencing the random outcome data; therefore these boundaries are random. Even though the HL test has been criticized by many others regarding its obvious disadvantages, it still serves as a standard test nowadays. Not many studies are designed for the investigation of the performance of the other two tests, namely the RSS and the CUSUM tests, as we focus on in this research.

Kuss (2002) addresses that the RSS may have potential higher power than the HL test, but more simulation study under different scenarios is needed, this research project is a part of our effort following the author's suggestion. We believe the CUSUM has a broad application in the goodness-of-fit tests in generalized linear models, and it deserves more investigation. Unfortunately due to its considerably long computing time than other two methods, very few studies consider applying the CUSUM test. This research study brought it to researchers' attention. In this sense, this research is the first one to assess the performance of those three

goodness-of-fit tests.

Th previous publications conduct simulation study with very few covariates included in the model. To make our knowledge of the characteristics of each test more applicable to real world data sets, we conduct our simulation in both the traditional way and a generalized way. This research is more complex than the published studies in two folds: more complicated systematic component (by expanding the model design matrix) and larger sample size (by increasing sample size up to 30,000).

We propose a new partition method for binning data differently than the HL test, which would present an issue of small expected frequency within a bin, to ensure the expected frequency within each bin/cell is at least five. To the best of our knowledge this research is the first one motivated by the conditions required by the asymptotic theorem for the limiting distribution of a chi-square type statistics.

7.2.2 Limitation of this research

The simulation results under varied scenarios show that the proposed NP test achieves higher power than three alternatives, but in many cases it is more conservative than others to reject the wrong model. The proposed NP test possesses some disadvantages:

1. The number of bins/groups increases when sample size increases, which is also a potential issue for the chi-square type tests statistic with random boundaries of bins.

2. We use three different degrees of freedom for different data/model(s). Even it represents a range of limiting distributions of the chi-square type statistic, it is still arbitrary and lacks of theoretical guidance on how to determine the appropriate one.

3. The proposed NP tests did adjust for number of estimated parameters in the logistic regression model, but not fully consider penalty for extreme over-fitting of the model.

4. It does not possess great power to detect missing quadratic terms. Even it achieves higher power than other studied tests with sample size less than 5,000 under the general

simulation setting as shown in chapter 5, it did show more conservative than other tests under the specific setting in chapter 4. These brought up some concerns of its consistent power performance.

The simulation settings has limitations too.

1. Even though beta coefficients including the intercept coefficient is randomly sampled, but the pool of beta coefficients were pre-specified and the values in the pool was quite arbitrary.

2. The number of replication was set up to 200, it can be a larger number so that the simulation results could be more stable and reliable.

3. The sample size class was quite arbitrary too, and the distance between each class is uneven, consequentially more models with smaller sample size were sampled.

7.2.3 Some guidelines of using goodness-of-fit test

We would like to offer some guidelines of how to use different goodness-of-fit tests in logistic regression when continuous covariates are presented based on our simulation study.

- We agree with the suggestion provided by Hosmer et al. that practitioners can consider using multiple goodness-of-fit tests together to ensure the specified model describe the data adequately.

- All goodness-of-fit tests perform better to detect under-fitted models than to detect over-fitted models, i.e. the goodness-of-fit tests reject models with missing terms more often than to reject models with additional terms. This suggests that goodness-of-fit tests are not sensitive to over-fitted models.

- It's easier for a test to detect omission of a quadratic term than to detect omission of an interaction term, i.e. all goodness-of-fit tests achieve higher power for the missing quadratic term than the missing interaction term.

- If keeping all other factors fixed and only with varied sample size and beta coefficient, The new partition test achieves higher power than the HL test to detect the omission of an interaction term for sample size smaller than 500 across varied beta settings. For sample size greater than 1,000, the NP test performs conservative than the rest three tests, i.e. not often to reject the wrong model, when sample size is less than 10,000 in most cases of varied beta settings to detect the omission of an interaction term.

- The NP test has higher power to detect the addition of an unrelated continuous covariate than three alternatives across varied sample size, and it is true no matter what distribution is the continuous covariate from (we investigate three distributions only, i.e. standard normal, $Uniform(-3, 3)$ and $Beta(1, 2)$).

- The NP test has higher power to detect the omission of a main effect than three alternatives when sample size is less than 15,000, when sample size is larger than 15,000, both the RSS and the CUSUM tests perform better than the HL and the NP tests.

- For some specific models, the test power could be very lower even when sample size is greater than 10,000.

- Sample size, the corresponding beta coefficient of the missing or additional term and the beta coefficients of related terms remaining in the model would affect the performance of all test powers. In general, larger sample size, larger magnitude of beta coefficient of the missing or additional term and smaller magnitude of beta coefficient of related terms remained in the model would increase test powers.

7.3 Future work

The directions we can consider for future work to advance this research exists in at least three areas:

1. Alternative simulation setting could be designed as increasing the complicity of the systematic components of the specified logistic regression model, that is we can grow the model by adding predictor variables step-by-step, to ensure more meaningful comparison, we can keep the existing terms fixed in the model when growing the systematic components. This way we may find more characteristics of the existing goodness-of-fit tests.

2. We may try different criteria, for example 10, for the expected cell frequency to partition data. We use 5 as the threshold in this research, we may search for optimal criteria for the combination of different systematic components and different scenarios.

3. Further more work is needed to see if there is potential to find an optimized degrees of freedom for the asymptotic distribution of the chi-square type of test statistic, $T_{G^\star}$, to replace the currently one proposed in (3.21).

References

[1] J. A. Nelder and R. W. M. Wedderburn. *Generalized Linear Models.* Journal of the Royal Statistical Society. Series A (General) Vol. 135, No. 3 (1972), pp. 370-384.

[2] P. McCullagh, John A. Nelder. *Generalized Linear Models, Second Edition.* CRC Press; Aug 1, 1989.

[3] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models.* CRC Press; 2008.

[4] Alan Agresti. *Categorical Data Analysis, 3rd Edition.* John Wiley & Sons.; December 3, 2012.

[5] David W. Hosmer, Stanley Lemeshow. *Applied Logistic Regression.* John Wiley & Sons.; 1989, 2000(2E), 2013(3E).

[6] Neter, J., W. Wasserman, and M. H. Kutner. *Applied Linear Statistical Models.* Irwin, Burr Ridge, Illinois. 1990.

[7] Glantz, S.A. and Slinker, B.K.. *Primer of Applied Regression and Analysis of Variance.* Health Professions Division, McGraw-Hill, New York. 1990.

[8] Douglas C. Montgomery, Elizabeth A. Peck. *Introduction to Linear Regression Analysis.* Wiley; Mar 4, 1992.

[9] Robert F. Woolson, William R. Clarke. *Statistical Methods for the Analysis of Biomedical Data, Second Edition.* John Wiley & Sons.; 2002.

[10] Tabachnick and Fidell. *Using Multivariate Statistics, 6th Edition.* Pearson; 2013.

[11] David W. Hosmer, Scott Taber, and Stanley Lemeshow. *The Importance of Assessing the Fit of Logistic Regression Models: A Case Study* . American Journal of Public Health 1991; Vol. 81, No. 12, 1630-1635.

[12] Peter J. Bickel, Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, 1st Edition.* Holden-Day; 1977.

[13] A. W. F. Edwards. *Likelihood, Expanded Edition.* Johns Hopkins University Press; October 1, 1992.

[14] D.R. Cox, E. J. Snell. *Analysis of Binary Data.* Chapman and Hall/CRC; May 15, 1989.

[15] Giselmar A. J. Hemmert, Laura M. Schons, Jan Wieseke, and Heiko Schimmelpfennig. *Log-likelihood-based Pseudo-R2 in Logistic Regression: Deriving Sample-sensitive Benchmarks.* Sociological Methods & Research 2018; Vol. 47(3), 507-531.

[16] Allison, Paul D. *Measures of Fit for Logistic Regression.* SAS Global Forum; Paper 1485. 2014.

[17] David Cox and Nancy Wermuth. *A Comment on the Coefficient of Determination for Binary Responses.* The American Statistician 1992; 46(1):1-4.

[18] Glenn Hoetker. *The use of logit and probit models in strategic management research: Critical issues.* Strategic Management Journal 2007; Volume 28, Issue 4: 331-43.

[19] UCLA-Institute for Digital Research and Education. *R Data Analysis Examples: Logit Regression.* https://stats.idre.ucla.edu/r/dae/logit-regression/.

[20] Pearson, Karl. *On the criterion that a given System of Devaitions 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.* Philosophical Magazine. Series 5 (1900). 50 (302): 157–175.

[21] Alan Agresti. *Categorical Data Analysis, 1st Edition*. Wiley; Mar 22, 1990.

[22] Hosmer DW, Lemeshow S. *Goodness of fit tests for the multiple logistic regression model*. Communications in Statistics-Theory and Methods 1980; 9(10):1043-1069.

[23] D. W. Hosmer, T. Hosmer, S. LE Cessie and S. Lemeshow. *A Comparison of Goodness-of-fit Tests for the Logistic Regression Model*. Statistics in Medicine 1997, VOL. 16(9): 965-980.

[24] Copas JB. *Unweighted sum of squares test for proportions*. Applied Statistics 1989; 38(1):71-78.

[25] le Cessie S, van Houwelingen HC. *Testing the fit of a regression model via score tests in random effects models*. Biometrics 1995 Jun; 51(2):600-14.

[26] Osius G, Rojek D. *Normal Goodness-of-fit Tests for Multinomial Models with Large Degrees of Freedom*. Journal of the American Statistical Association 1992; 87(420): 1145-1152.

[27] Cressie NAC, Read TRC. *Multinomial Goodness-of-fit Tests*. Journal of the Royal Statistical Society, Series B 1984; 46(2):440-464.

[28] Royston P. *The USe of Cusums and other Techiques in Modeling Continuous Covariates in Logistic Regression*. Statistics in Medicine 1992; 11:1115-1129.

[29] Royston P. *Cusum Plots and Tests for Binary Variables*. STATA Technical Bulletin 1993. STB-12, 16-17.

[30] John Q. Su and L. J. Wei. *A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model*. Journal of the American Statistical Association, Vol. 86, No. 414, Jun., 1991, pp.420-426.

[31] David W. Hosmer and Nils Lid Hjort. *Goodness-of-fit processes for logistic regression: simulation results*. Statistics in Medicine 2002; 21:2723-2738.

[32] Oliver Kuss. *Global goodness-of-fit tests in logistic regression with sparse data.* Statistics in Medicine 2002; 21:3789-3801.

[33] Pregibon D. *Logistic regression diagnostics.* Annals of Statistics 1981; 9(2):705–724.

[34] Santner T. J. *The statistical analysis of discrete data.* Springer: New York 1989.

[35] David S. Moore and M. C. Spruill. *Unified Large-sample Theory of General Chi-square Statistics for Tests of Fit.* The Annals of Statistics 1975; Vol 3, No. 3, 599-616.

[36] Wei Yu, Wangli Xu, and Lixing Zhu. *A modified Hosmer-Lemeshow test for large data sets.* Communications in Statistics-Theory and Methods 2017; Vol 46, No. 23, 11813-11825.

[37] G. S. Watson. *Some Recent Results in Chi-Square Goodness-of-Fit Tests.* Biometrics 1959; Vol. 15, No. 3, pp. 440-468.

[38] D.W. Hosmer, S. Lemeshow, J. Klar. *Goodness-of-Fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small.* Biometrical Journal 1988; Vol 30, Issue 8, 911-924.

[39] Xu H. *Extensions of the Hosmer-Lemeshow goodness-of-fit test.* University of Massachusetts at Amherst 1996.

[40] Guido Bertolini, Roberto D'Amico, D Nardi, Angelo Tinazzi and G Apolone. *One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model.* Journal of Epidemiology and Biostatistics 2000; 5(4):251-3.

[41] Joseph G. Pigeon and Joseph F. Heyse. *An Improved Goodness of Fit Statistic for Probability Prediction Models.* Biometrical Journal 1999; Vol 41, Issue 1, 71-82.

[42] Horn, S. D. *Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale.* Biometrics 1977; 33, 237-247.

[43] D.Y. Lin, L. J. Wei, Z. Ying. *Model Checking Techniques Based on Cumulative Residuals*. Biometrics March 2002; 58, 1-12.

[44] Zhiying Pan and D.Y. Lin. *Goodness-of-fit Methods for Generalized Linear Mixed Models*. Biometrics December 2005; 61, 1000-1009.

[45] Stukel, T. A.. *Generalized logistic models*. Journal of American Statistical Association 1988; 83, 426-431.

[46] Brown, C. C.. *On a goodness-of-fit test for the logistic model based on score statistics*. Communications in Statistics 1982; 11, 1087—1105.

[47] Prentice, R. L.. *A generalization of the probit and logit methods for dose response curves*. Biometrics 1976; 32, 761—768.

[48] le Cessie, S. and van Houwelingen, J. C. *A goodness-of-fit test for binary data based on smoothing residuals*. Biometrics; 47, 1267—1282.

[49] McCullagh P. *On the asymptotic distribution of Pearson's statistic in linear exponential-family models*. International Statistical Review 1985; 53(1):61–67.

[50] Farrington C. P. *On assessing goodness of fit of generalized linear models to sparse data*. Journal of the Royal Statistical Society, Series B 1996; 58(2):349–360.

[51] Erik Pulkstenis and Timothy Robinson. *Two goodness-of-fit tests for logistic regression models with continuous covariates*. Statistics in Medicine 2002; 21:79-93.

[52] Tsiatis, A. A. *A Note on a Goodness-of-Fit Test for the Logistic Regression Model*. Biometrika, 67, 250-251.

[53] Jana D. Canary, Leigh Blizzard, Ronald P. Barry, David W. Hosmer and Stephen J. Quinn. *A comparison of the Hosmer–Lemeshow, Pigeon–Heyse, and Tsiatis goodness-of-fit tests for binary logistic regression under two grouping methods*. Communications in Statistics-Simulation and Computation 2017; Volume 46, Issue 3.

[54] Xian-Jin Xie, Jane Pendergast, William Clarke. *Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors.* Computational Statistics and Data Analysis 2008; Vol 52, Issue 5, 2703-2713.

[55] Dreiseitl, S., and Osl, M. *Effects of Data Grouping on Calibration Measures of Classifiers.* Computer Aided Systems Theory–EUROCAST 2011; 359-366.

[56] Halbert White. *Maximum Likelihood Estimation of Misspecified Models.* Econometrica 1982; 50 (1): 1–25.

[57] Kennan, J., and Neuman G. R. *Why Does the Information Matrix Test Reject So Often*? Hoover Institution, Stanford University, Working Papers in Economics 1988; E-88-10.

[58] Prabasaj Paul, Michael L. Pennell and Stanley Lemeshow. *Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets.* Statistics in Medicine 2013; 32: 67–80.

[59] Martin Schader and Friedrich Schmid. *Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution.* The American Statistician 1989; Volume 43, Issue 1, 23-24.

[60] William G. Cochran. *The X^2 test of the goodness-of-fit.* The Annals of Mathematical Statistics 1952; Vol 23, No. 3, 315-345.

[61] C. Arthur Williams. *On the choice of the number and width of classes for the chi-square test of the goodness-of-fit.* Journal of American Statistical Association 1950; Vol 45, 77-86.

[62] Harald Cramer. *Mathematical Methods of Statistics.* Princeton University Press, 1946.

[63] C. R. Rao. *Linear Statistical Inference and its Applications.* John Wiley, 1973.

[64] Herman Chernoff and E. L. Lehmann. *The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit.* Ann. Math. Statist 1954; Volume 25, Number 3, 579-586.

[65] G. S. Watson. *On Chi-Square Goodness-Of-Fit Tests for Continuous Distributions.* Journal of the Royal Statistical Society. Series B (Methodological) 1958; Vol. 20, No. 1, pp. 44-72.

[66] Luciano Molinari. *Distribution of the Chi-Squared Test in Nonstandard Situations.* Biometrika 1977; Vol. 64, No. 1, pp. 115-121.

[67] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. *Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients.* The Journal of the American Medical Association 1993; 270(20):2478-86.

[68] Lemeshow S, Le Gall JR. *Modeling the severity of illness of ICU patients. A systems update.* The Journal of the American Medical Association 1994; 272(13):1049-55.

[69] Ying Liu, Paul I. Nelson, Shie-Shien Yang. *An omnibus lack 0f fit test in logistic regression with sparse data.* Stat Methods Appl 2012; 21: 437-452.

[70] J.Andrew Royle, Richard B. Chandler, Rahel Sollmann and Beth Gardner. *Spatial Capture-Recapture.* Academic Press, Elsevier Inc. 2014.

Appendix A

A.1 Proof 1: D statistic in (2.3) is asymptotically equivalent to X^2 statistic in (2.2)

Th chi-square in (2.2) and deviance (2.3) statistics are asymptotically equivalent. We prove this use the Taylor series expansion of $Olog\frac{O}{E}$ about $O = E$, namely,

$$Olog\frac{O}{E} = (O-E) + \frac{1}{2}\frac{(O-E)^2}{E} + \cdots$$

by ignoring the effect of the remaining higher-order terms, then

$$\begin{aligned}
D &= 2\sum_{i=1}^{g}\left(y_i log\Big(\frac{y_i}{n_i\hat{\pi}_i}\Big) + (n_i - y_i)log\Big(\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}\Big)\right) \\
&= 2\sum_{i=1}^{g}\left((y_i - n_i\hat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + [(n_i - y_y) - n_i(1-\hat{\pi}_i)] + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\hat{\pi}_i)]^2}{n_i(1-\hat{\pi}_i)} + \cdots\right) \\
&= 2\sum_{i=1}^{g}\left(\frac{1}{2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\hat{\pi}_i)]^2}{n_i(1-\hat{\pi}_i)} + \cdots\right) \\
&= \sum_{i=1}^{g}\left(\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i(1-\hat{\pi}_i)} + \cdots\right) \\
&\approx \sum_{i=1}^{g}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \\
&= X^2.
\end{aligned}$$

A.2 Proof 2: Inequality (2.7)

Some notations:

$k = 1, \ldots, g$: the k^{th} bin/group

$j = 1, \ldots, c_k$: the k^{th} unique covariate pattern in the k^{th} bin

$n_k^{'}$: the total number of observations in the k^{th} bin

$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k^{'}}$: the average estimated probability of all subjects falling in the k^{th} bin

c_k: the number of covariate pattterns in the k^{th} bin

m_j: the number of subjetcs of covariate pattern j in the k^{th} bin

$$\begin{aligned}\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j) &= \sum_{j=1}^{c_k} m_j \hat{\pi}_j - \sum_{j=1}^{c_k} m_j \hat{\pi}_j^2 \\ &= n_k^{'} \bar{\pi}_k - \sum_{j=1}^{c_k} m_j \hat{\pi}_j^2. \end{aligned} \tag{A.1}$$

$$\begin{aligned}\sum_{j=1}^{c_k} m_j (\hat{\pi}_j - \bar{\pi}_k)^2 &= \sum_{j=1}^{c_k} m_j \hat{\pi}_j^2 - 2\bar{\pi}_k \sum_{j=1}^{c_k} m_j \hat{\pi}_j + \bar{\pi}_k^2 \sum_{j=1}^{c_k} m_j \\ &= \sum_{j=1}^{c_k} m_j \hat{\pi}_j^2 - 2\bar{\pi}_k n_k^{'} \bar{\pi}_k + n_k^{'} \bar{\pi}_k^2 \qquad (since \quad \sum_{j=1}^{c_k} m_j = n_k^{'}) \\ &= \sum_{j=1}^{c_k} m_j \hat{\pi}_j^2 - n_k^{'} \bar{\pi}_k^2 \end{aligned} \tag{A.2}$$

Add (A.1) and (A.2), we have

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j) + \sum_{j=1}^{c_k} m_j (\hat{\pi}_j - \bar{\pi}_k)^2 = n_k^{'} \bar{\pi}_k - n_k^{'} \bar{\pi}_k^2 = n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k). \tag{A.3}$$

Therefore, from ($A.3$) we obtain inequality (2.7)

$$\sum_{j=1}^{c_k} m_j \hat{\pi}_j (1 - \hat{\pi}_j) = n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k) - \sum_{j=1}^{c_k} m_j (\hat{\pi}_j - \bar{\pi}_k)^2$$

$$< n_k^{'} \bar{\pi}_k (1 - \bar{\pi}_k) \qquad (since \;\; m_j > 0 \;\; and \;\; (\hat{\pi}_j - \bar{\pi}_k)^2 > 0 \;\; \forall \, j \in \{1, \ldots, c_k\}).$$

Appendix B

B.1 R function for data simulation in the study

```
simulate.Data<- function(beta, intercept, ss){
beta<-sample(beta, 20, replace = TRUE)
intercept<-sample(intercept, 1)

if (ss == "small") {
N <- nobs <- sample(200:500, 1)
} else if (ss == "moderate"){
N <- nobs <- sample(500:2000, 1)
} else if (ss == "large"){
N <- nobs <- sample(2000:5000, 1)
} else if (ss == "extralarge"){
N <- nobs <- sample(5000:15000, 1)
} else { N <- nobs <- sample(15000:30000, 1)}
nc<-sample(3:5, 1)
nd<-sample(2:5, 1)
ni<-sample(0:2, 1)
nq<-sample(0:2, 1)
collinearity=sample(c("null", "low", "moderate", "high"), 1)
continuous<-data.frame(x1=rnorm(N), x2=rnorm(N, 0, 4),
```

```
    x3=runif(N, -1, 1), x4=rbeta(N, 2, 2), x5=rnorm(N))
sqX<-continuous^2
colnames(sqX)<-paste0(names(continuous), "Sqr")
continuousX<-continuous[, 1:5]
Xsq<-sqX[, 1:5]
dichots<-data.frame(z1=rbinom(N, 1, 0.1), z2=rbinom(N, 1, 0.3),
    z3=rbinom(N, 1, 0.5), z4=rbinom(N, 1, 0.7), z5=rbinom(N, 1,
    0.9))
interacts<-data.frame(x1z1=continuousX[, 1]*dichots[, 1],
    x1z2=continuousX[, 1]*dichots[, 2], x1z3=continuousX[,
    1]*dichots[, 3], x1z4=continuousX[, 1]*dichots[, 4],
    x1z5=continuousX[, 1]*dichots[, 5], x2z1=continuousX[,
    2]*dichots[, 1], x2z2=continuousX[, 2]*dichots[, 2],
    x2z3=continuousX[, 2]*dichots[, 3], x2z4=continuousX[,
    2]*dichots[, 4], x2z5=continuousX[, 2]*dichots[, 5],
    x3z1=continuousX[, 3]*dichots[, 1], x3z2=continuousX[,
    3]*dichots[, 2], x3z3=continuousX[, 3]*dichots[, 3],
    x3z4=continuousX[, 3]*dichots[, 4], x3z5=continuousX[,
    3]*dichots[, 5], x4z1=continuousX[, 4]*dichots[, 1],
    x4z2=continuousX[, 4]*dichots[, 2], x4z3=continuousX[,
    4]*dichots[, 3], x4z4=continuousX[, 4]*dichots[, 4],
    x4z5=continuousX[, 4]*dichots[, 5], x5z1=continuousX[,
    5]*dichots[, 1], x5z2=continuousX[, 5]*dichots[, 2],
    x5z3=continuousX[, 5]*dichots[, 3], x5z4=continuousX[,
    5]*dichots[, 4], x5z5=continuousX[, 5]*dichots[, 5])
ccoef<-seq(0.1, 0.9, 0.1)
cc.null<-0
```

```
cc.low<-sample(ccoef[1:3], 1)
cc.mod<-sample(ccoef[4:6], 1)
cc.high<-sample(ccoef[7:9], 1)
corrX1 <- rmvnorm(N, mean=c(0, 0), sigma=matrix(c(1,
    cc.low*sqrt(2), cc.low*sqrt(2), 2),2,2), method = "chol")
corrX2 <- rmvnorm(N, mean=c(0, 0), sigma=matrix(c(1,
    cc.mod*sqrt(2), cc.mod*sqrt(2), 2),2,2), method = "chol")
corrX3 <- rmvnorm(N, mean=c(0, 0), sigma=matrix(c(1,
    cc.high*sqrt(2), cc.high*sqrt(2), 2),2,2), method = "chol")
corrX<-data.frame(corrX1, corrX2, corrX3)
names(corrX)<-c("corrlow", "corrlow2", "corrmod", "corrmod2",
    "corhigh", "corhigh2")

colcontinX<-colnames(continuousX)
colcontinXSamp<-sample(colcontinX, nc)
continX<-continuousX[, colcontinXSamp, drop=FALSE]
nameXs<-colnames(continuousX)
nameX<-colnames(continX)
nameXsq<-colnames(Xsq)
nameXsqs<-substr(nameXsq, 1, 2)
if (length(nameX)==length(nameXsqs)){
Xsq2<-Xsq
} else {Xsq2<-Xsq[, -which(!nameXsqs%in%nameX)]}
missX<-setdiff(nameXs, nameX)
coldichots<-colnames(dichots)
coldichotsSamp<-sample(coldichots, nd)
dichot<-dichots[, coldichotsSamp, drop=FALSE]
```

```
nameds<-colnames(dichots)
named<-colnames(dichot)
missd<-setdiff(nameds, named)
nameint<-colnames(interacts)

if (length(missd)==0) { nameint1<-nameint
} else {
missint1<-NULL
for (i in 1:length(missd)) {
index1<-which(grepl(missd[i], nameint))
missint1<-c(missint1, index1)}
nameint1<-nameint[-missint1]
}

if (length(missX)==0) {nameint2<-nameint1
} else {
missint2<-NULL
for (i in 1:length(missX)) {
index2<-which(grepl(missX[i], nameint1))
missint2<-c(missint2, index2)}
nameint2<-nameint1[-missint2]
}
int<-interacts[, nameint2, drop=FALSE]
colint<-colnames(int)
colintSamp<-sample(colint, ni)
interX<-int[, colintSamp, drop=FALSE]
colXsq<-colnames(Xsq2)
```

```
colXsqSamp<-sample(colXsq, nq)
Xsqs<-Xsq2[, colXsqSamp, drop=FALSE]

if (collinearity=="null") {
covars<-data.frame(continX, dichot, interX, Xsqs)
L<-sample(beta, dim(covars)[2], replace=TRUE)
df = data.frame(matrix(intercept, nrow=N), t(t(covars) * L))
s=rowSums(df)
pr = 1/(1+exp(-s))
y = ifelse(runif(N) < pr, 1, 0)
data<-data.frame(y, covars)
nSuccess<-sum(y)
ldt<-list(N, nSuccess, nc, nd, ni, nq, collinearity, cc.high,
    data, L)
} else if (collinearity=="low") {
covars<-data.frame(continX, dichot, interX, Xsqs, corrX[, 1:2])
L<-sample(beta, dim(covars)[2], replace=TRUE)
df = data.frame(matrix(intercept, nrow=N), t(t(covars) * L))
s=rowSums(df)
pr = 1/(1+exp(-s))
y = ifelse(runif(N) < pr, 1, 0)
data<-data.frame(y, covars)
nSuccess<-sum(y)
ldt<-list(N, nSuccess, nc, nd, ni, nq, collinearity, cc.high,
    data, L)

} else if (collinearity=="moderate") {
```

```
covars<-data.frame(continX, dichot, interX, Xsqs, corrX[, 3:4])
L<-sample(beta, dim(covars)[2], replace=TRUE)
df = data.frame(matrix(intercept, nrow=N), t(t(covars) * L))
s=rowSums(df)
pr = 1/(1+exp(-s))
y = ifelse(runif(N) < pr, 1, 0)
data<-data.frame(y, covars)
nSuccess<-sum(y)
ldt<-list(N, nSuccess, nc, nd, ni, nq, collinearity, cc.high,
    data, L)
} else if (collinearity=="high") {
covars<-data.frame(continX, dichot, interX, Xsqs, corrX[, 5:6])
L<-sample(beta, dim(covars)[2], replace=TRUE)
df = data.frame(matrix(intercept1[1], nrow=N), t(t(covars) * L))
s=rowSums(df)
pr = 1/(1+exp(-s))
y = ifelse(runif(N) < pr, 1, 0)
data<-data.frame(y, covars)
nSuccess<-sum(y)
ldt<-list(N, nSuccess, nc, nd, ni, nq, collinearity, cc.high,
    data, L)
}
return(ldt)
}

# This function can be easily extended by including set.seed()
    method for reproducible research purpose.
```

```
# usage:
beta<-c(-4, -2, -1.5, -1, -0.75, -0.5, -0.2, -0.1, 0.1, 0.2, 0.5,
    0.75, 1, 1.5, 2, 4);
intercept<-c(-2, 2);
ldt<-simulate.Data(beta, intercept, "small");
# not run
```