Goodness-of-Tests for Logistic Regression

Sutan Wu
THE FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

GOODNESS-OF-TESTS FOR LOGISTIC REGRESSION

By
SUTAN WU

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Fall Semester, 2010
The members of the Committee approve the Dissertation of Sutan Wu defended on August 19, 2010.

______________________________
Dan L. McGee
Professor Co-Directing Dissertation

______________________________
Jinfeng Zhang
Professor Co-Directing Dissertation

______________________________
Myra Hurt
University Representative

______________________________
Debajyoti Sinha
Committee Member

Approved:

______________________________
Dan McGee, Chair, Department of Statistics

______________________________
Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above named committee members.
To Dun and my parents
ACKNOWLEDGEMENTS

I would first like to acknowledge my major professor, Dr. Dan McGee for his guidance, patience and encouragement. His help was essential for the completion of my thesis. I have learned a lot from him and I wish I listened better and learned even more from his advice. Next, I would like to acknowledge to my co-adviser, Dr. Zhang for his encouragement and great help on computational part of this thesis. My gratitude also goes to Dr. Sinha and Dr. Hurt for serving in my committee and for their comments showing me different angles to look at my research.

I came to FSU in 2005. Living in a foreign country is never easy at the beginning. Here, I want to thank Dr. Niu for all his help and personal opinions. Radha has helped me enormously for my teaching and I learned a lot from her how to be an effective teacher. Dr. Ramiser offered me the great opportunity for being a statistics consultant, which gave me the chance to apply statistics in multiple disciplines. Also, Dr. Ramiser has made very important contributions to my consulting skills sharing his experience and ideas with me. My appreciation extends to Chauncey and Pam for always being helpful.

Finally I wish to express my deepest gratitude to my parents and my husband Dun, for their love during these years. I was very lucky to meet Dun at the gate of the flight to the United States in 2005. Thank Dun for always standing by me.
# TABLE OF CONTENTS

List of Tables ................................................................. vii
List of Figures ................................................................. ix
Abstract  ................................................................. x

1.  **INTRODUCTION** ............................................................... 1

2.  **LOGISTIC REGRESSION, GENERALIZED LINEAR MODEL AND GOODNESS-OF-FIT TESTS** ................................................................. 3

3.  **GOODNESS-OF-FIT TESTS FOR LOGISTIC REGRESSION** ................. 6
   3.1 Cornfield’s Early Work ................................................. 6
   3.2 Pearson’s chi-square and Deviance Tests .......................... 7
   3.3 Hosmer and Lemeshow Test .......................................... 10
   3.4 Unweighted Sum of Squares Test ................................... 12
   3.5 Cumulative Sums of Residuals ....................................... 13
   3.6 Other Tests .......................................................... 17

4.  **COMPARISONS-EMPIRICAL STUDIES** ...................................... 22
   4.1 Motivating Example ................................................... 22
   4.2 Empirical Studies ..................................................... 24
   4.3 Conclusions .......................................................... 36

5.  **COMPARISONS-SIMULATION STUDIES** ..................................... 37
   5.1 Null Distribution ..................................................... 38
   5.2 Functional Form of Covariates ...................................... 43
   5.3 Omitted Interactions ................................................ 48
   5.4 Link Function ........................................................ 50
   5.5 Conclusions .......................................................... 54
   5.6 Discussion ........................................................... 55

6.  **FUTURE WORK, EXTENSION TO THE PROPORTIONAL HAZARDS MODEL** ................................................................. 58
   6.1 Adaption G.O.F tests for Cox Proportional Hazards Model ........ 59
LIST OF TABLES

3.1 Part of Cornfield’s Early G.O.F Test ........................................... 7
3.2 Data Classification for Pearson chi-square and Deviance tests .......... 8
3.3 Observed (Obs) and Expected (Exp) Counts Within Each Decile of Risk,
    Defined by Fitted Values (Prob.) for y = 1 and y = 0 .................... 11
3.4 Hosmer-Lemeshow Test Results With Differenting Number of Groups .... 12
3.5 Steam Data [2] ................................................................. 17
4.1 Estimation Results for the linear Main-effect Model (UIS Study) ......... 23
4.2 P-values of goodness-of-fit tests for the fitted model in Table 4.1 ......... 23
4.3 P-values of goodness-of-fit tests for model with non-linear term (UIS study) 24
4.4 Estimation Results for the Final Model (UIS Study) ....................... 25
4.5 P-values of goodness-of-fit tests for the Final Model (UIS Study) ........ 25
4.6 Estimation Results for Main-Effect Model (LBW study) .................. 27
4.7 P-values of goodness-of-fit tests for the Main-Effect Model (LBW study) 28
4.8 Estimation Results for the Final Model (LBW study) ....................... 28
4.9 P-values of goodness-of-fit tests for the Final Model (LBW study) ....... 28
4.10 Estimation Results for the Final Model (ICU study) ..................... 31
4.11 P-values of goodness-of-fit tests for the Final Model (ICU study) ........ 31
4.12 Estimation Results for the Main-Effect Model (NHANES III study) ..... 33
4.13 P-values of goodness-of-fit tests for the Main-Effect Model (NHANES III study) 33
4.14 P-values of goodness-of-fit tests for the Transformed Model (NHANES III study) 34
4.15 Estimation Results for the Final Model (NHANES III study) ............ 34
4.16 P-values of goodness-of-fit tests for the Final Model (NHANES III study) ........................................... 36
5.1 Settings used to examine the null distribution ................................................................. 40
5.2 Simulated percent rejection at the $\alpha = 0.05$ for Null Distribution .......................... 42
5.3 Coefficients for detecting quadratic form of $x$ ................................................................. 43
5.4 Percentage Rejection with $\alpha = 0.05$ for Omission of Quadratic Term, 1000 Simulations ......................................................................................................................... 45
5.5 Simulated percent rejection at the $\alpha = 0.05$ level for the omitted cubic term .............. 46
5.6 Settings for non-linearity forms of $x$ .................................................................................. 46
5.7 Simulated percent rejection at the $\alpha = 0.05$ level for non-linear form of $x$ from $\chi^2(4)$ .................................................................................................................................. 47
5.8 Simulated percent rejection at the $\alpha = 0.05$ level for non-linear form of $x$ from $U(1, 20)$ .................................................................................................................................. 47
5.9 Simulated percent rejection at the $\alpha = 0.05$ level for the $W_x$ test ............................... 48
5.10 Settings for a omission of interaction term ...................................................................... 48
5.11 Simulated percent rejection at the $\alpha = 0.05$ level for an omission of interaction term I .................................................................................................................................. 50
5.12 Simulated percent rejection at the $\alpha = 0.05$ level for an omission of interaction term II .................................................................................................................................. 50
5.13 Simulated percent rejection at the $\alpha = 0.05$ level for Misspecified Link Functions I .................................................................................................................................. 51
5.14 Types of Misspecified Link Functions ............................................................................ 54
5.15 Simulated percent rejection at the $\alpha = 0.05$ level for Misspecified Link Functions II .................................................................................................................................. 54
5.16 Simulated percent rejection at the $\alpha = 0.05$ level for Parametric Bootstrap .................. 56
6.1 Data set for Stacked Logistic Regression ........................................................................... 62
6.2 Study Design ..................................................................................................................... 63
6.3 Comparison Results .......................................................................................................... 64
6.4 Comparison Results .......................................................................................................... 65
6.5 “Hosmer-Lemeshow” Test Results in Stacked Logistic Regression ................................... 66
LIST OF FIGURES

4.1 Cumulative Sums of Residuals for UIS Study. **Top**: Cumulative Sums of Residuals for the main effect model; **Bottom**: Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. ......................................................... 26

4.2 Cumulative Sums of Residuals for LBW Study. **Top**: Cumulative Sums of Residuals for the main effect model; **Bottom**: Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. ......................................................... 29

4.3 Cumulative Sums of Residuals for ICU Study. The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. .................. 32

4.4 Cumulative Sums of Residuals for NHANES III Study. **Top**: Cumulative Sums of Residuals for the main effect model; **Bottom**: Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. ................................. 35

5.1 Common Situations for $\eta(x)$ ............................. 39

5.2 Four Settings For the Quadratic Term ................................. 44

5.3 Setting for Cubic term of $x$ ........................................ 45

5.4 Three Settings For an Omitted Interaction term: Setting 2 display progressively more interaction than Setting 1; Setting 3 display progressively more interaction than Setting 2. .......................... 49

5.5 Link Function I: Logit, Probit and Complementary log-log Link .............. 52

5.6 Link Functions II: Symmetric or asymmetric distributions with either lighter or heavier tails ................................................. 53

6.1 Cox-Snell Overall Test .................................................. 66
ABSTRACT

The generalized linear model and particularly the logistic model are widely used in public health, medicine, and epidemiology. Goodness-of-fit tests for these models are popularly used to describe how well a proposed model fits a set of observations. These different goodness-of-fit tests all have individual advantages and disadvantages. In this thesis, we mainly consider the performance of the “Hosmer-Lemeshow” test, the Pearson’s chi-square test, the unweighted sum of squares test and the cumulative residual test. We compare their performance in a series of empirical studies as well as particular simulation scenarios. We conclude that the unweighted sum of squares test and the cumulative sums of residuals test give better overall performance than the other two. We also conclude that the commonly suggested practice of assuming that a p-value less than 0.15 is an indication of lack of fit at the initial steps of model diagnostics should be adopted. Additionally, D’Agostino et al. presented the relationship of the stacked logistic regression and the Cox regression model in the Framingham Heart Study [1]. So in our future study, we will examine the possibility and feasibility of the adaption these goodness-of-fit tests to the Cox proportional hazards model using the stacked logistic regression.
CHAPTER 1

INTRODUCTION

The generalized linear model and particularly the logistic model are widely used in public health, medicine, and epidemiology. Goodness-of-fit tests for these models are popularly used to describe how well a proposed model fits a set of observations. In logistic regression and the generalized linear model, there are many goodness-of-fit tests proposed with different purposes. They all have their individual advantages and disadvantages. The development of methods for judging whether a logistic regression or a generalized linear model fits the data well and their performance in particular situations remains a major research area. Hence, the questions we are interested in is under which situations should a particular methodology be preferred; and can we assess their performance under different data situations?

Hosmer et. al. previously compared the performance of goodness-of-fit tests in the ordinary binary data situation [3, 4] and Evans et. al. (2005) previously compared goodness-of-fit tests under the clustered binary data situation (logit GEE) [5]. In our study, we focus on examining the four most popular and important goodness-of-fit tests: the Pearson’s Chi-square test, the “Hosmer-Lemeshow” test, the unweighted sums of squares test and the cumulative sums-of-residuals test. We approach our examination through two steps. In the first step, we compare the performance of these tests in a series of empirical data sets with small and large sample size, multiple categorical variables, and perhaps requiring non-linear terms for describing the relationship between a characteristic and the outcome. In the second step, we compare their performance in a series of simulation scenarios. The simulation scenarios are designed to check the power of these goodness-of-fit tests under a variety of situations. Additionally, we consider adapting these goodness-of-fit tests for use with the Cox proportional hazards model. These adaptations are based on using the stacked logistic regression that has been shown by D’Agostino et. al. [1] to be, under certain conditions,
asymptotically equivalent to the Cox proportional hazards model.

This thesis is constructed as follows: In Chapter 2, there is a brief introduction to logistic regression, the generalized liner model, and goodness-of-fit tests. In Chapter 3, we introduce the four goodness-of-fit tests which we are most interested in, and briefly describe some additional tests. In Chapter 4 and Chapter 5, we compare the performance of these tests in empirical settings and simulation scenarios. Then, in the last Chapter, we discuss our initial work at adapting these goodness-of-fit tests for use with the Cox proportional hazard model and discuss our planned future work for these adaptations. All theoretical proofs are presented in appendices. We also provide a brief introduction to fractional polynomials as an appendix.
CHAPTER 2

LOGISTIC REGRESSION, GENERALIZED LINEAR MODEL AND GOODNESS-OF-FIT TESTS

In data analysis in many fields, the logistic regression model or another generalized linear model are the most widely-used statistical modeling methodologies. The logistic regression model was originally developed when the large observational studies such as the Framingham Heart Study began to mature. More recently, it has been used extensively in the medical and social sciences as well as marketing applications such as prediction of a customer’s propensity to purchase a product or cease a subscription.

Suppose we have followed \( n \) individuals for a period of time \( t \) and recorded their covariates at the beginning of the study and whether they have failed by the end of the study. For the \( i^{th} \) subject, the response variable \( y_i \) is binary - takes value 1 if the subject has failed, and 0 if he/she hasn’t failed. Denote \( X_i = (1, x_{i1}, ..., x_{ip})' \) as the covariate vector and \( \pi_i \) as the probability of failure by the end of the follow up.

The logistic regression model specifies:

\[
\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip},
\]

where \( \beta = (\beta_0, ..., \beta_p)' \) are the regression coefficients.

In a logistic regression model the parameter \( \beta_k, k = 1, ..., p \) associated with the explanatory variable \( x_k \) is such that \( \exp(\beta_k) \) is the odds that \( y = 1 \) when \( x_k \) increases by one unit, conditional on the other explanatory variables remaining the same. Maximum likelihood is used to estimate the parameters, the log-likelihood function being

\[
l(\beta; y_i) = \sum_{i=1}^{n} y_i \log[\pi_i(\beta'X_i)] + (1 - y_i)\log[1 - \pi_i(\beta'X_i)].
\]
This log-likelihood function is maximized numerically using an iterative algorithm such as the Newton-Raphson method.

The logistic regression model is a special case of the generalized linear model with a logit link function. The generalized linear model (GLM) is a flexible generalization of ordinary least squares regression. It generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

In GLM, the response variable of the \( i \)th subject, \( y_i \), is assumed to be generated from a particular distribution in the exponential family. Suppose the density function of \( y_i \) is expressed by

\[
f(y_i|\theta, \phi) = \exp\{[a(\theta)y_i - b(\theta)]/\phi + \zeta(y_i, \phi)\}.
\]

Here, \( \theta \) is the parameter of interest. \( \phi \) is called dispersion parameter, typically \( \phi \) is known and is usually related to the variance of the distribution. Let \( \mu(\theta) \) be the mean of \( y_i \) and \( h \) be the link function, then the GLM connects \( y_i \) with the covariate vector \( X_i \) by

\[
E(y_i) = h(\mu(\theta)) = \beta'X_i,
\]

where \( E(y_i) \) is the expected value of \( y_i \); \( \eta_i = \beta'X_i \) is the linear predictor for the \( i \)th subject; \( \beta = (\beta_0, ..., \beta_p)' \) is the coefficient vector. The unknown \( \beta \)'s are typically estimated by maximum likelihood or maximum quasi-likelihood techniques. Suppose that the data consists of \( n \) independent replicates of \( (y_i, X_i) \), \( i = 1, ..., n \). Then the likelihood score function for \( \beta \) takes the form:

\[
U(\beta) = \sum_{i=1}^{n} u(\beta'X_i)X_i(y_i - v(\beta'X_i)), \tag{2.1}
\]

where \( v(r) = h^{-1}(r) \), and \( u(r) = \partial\{h \circ \mu(r)\}^{-1}/\partial r \).

A goodness-of-fit (G.O.F) test describes how well a model 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. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions using e.g., the Kolmogorov-Smirnov test, or whether outcome frequencies follow a specified distribution, using a test such as Pearson’s chi-square test.

4
There are many methodologies in the G.O.F test area, but our study will focus on the methodologies proposed primarily for the logistic regression model and the generalized linear model. One reason is that most G.O.F tests come from the generalized linear model, and another is that the essential ideas of these tests can be well adapted to other more complicated models or situations. We introduce the most commonly used G.O.F tests in the next Chapter.
In this chapter, we introduce the most popular and important goodness-of-fit (G.O.F) tests for logistic regression and the generalized linear model. These G.O.F tests are proposed mainly for the overall measures of fit. The null hypothesis for these G.O.F tests is that the fitted model is correct in all aspects, and the alternative hypothesis is that the fitted model lacks fit.

Suppose we have \( n \) independent observations of the pair \((y_i, X_i)\), where \( y_i \) is the dichotomous outcome and \( X_i = (1, x_{1i}, ..., x_{pi})' \) is the \((p + 1) \times 1\) covariate vector, \( i = 1, .., n \). Denote the \( \hat{\pi}_i \) as 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.

Thus, a complete assessment of the fitted model involves both the calculation of summary measures of the distance between \( y_i \) and \( \hat{\pi}_i \), and a thorough examination of the individual components of these measures. In following sections, we will introduce some measures of fit for logistic regression and the generalized linear model in a strict binary situation.

### 3.1 Cornfield’s Early Work

The earliest goodness-of-fit test for logistic regression comes from Cornfield’s discriminant function approach to logistic regression considering the effect of serum cholesterol (Chol) and
systolic blood pressure (SBP) on the risk of coronary heart disease (CHD) [6]. Cornfield’s examination of goodness-of-fit consists of comparisons between the observed incidence of CHD events and the expected incidences based on the estimated model within different levels of SBP and Chol. The estimated incidence of CHD event was calculated as

\[
\frac{\text{population at risk}}{[1 + \exp(23.13 - 6.14 x_1 - 3.29 x_2)]},
\]

where \( x_1 = \log(\text{Chol}) \) and \( x_2 = \log(\text{SBP} - 75) \).

Table 3.1 presents part of the calculation of the goodness-of-fit test used by Cornfield in his original paper. From the results in this table, we can see that the proposed goodness-of-fit test appears to support that the estimated equation fits the data.

### 3.2 Pearson’s chi-square and Deviance Tests

Pearson’s chi-square test and the deviance test are two goodness-of-fit tests that we are most familiar with. They are constructed using two measures of the difference between the observed response variable \( y \) and its fitted values \( \hat{\pi} \): the Pearson residual and the deviance residual.

Let \( J \) denote the number of possible distinct values of covariate vector \( X \). Denote the number of subjects with the covariate pattern, \( X = x_j \) by \( m_j, j = 1, ..., J \). Let \( y_j \) denote the number of positive responses, \( y = 1 \), among the \( m_j \) subjects with \( X = x_j \). \( \hat{\pi}_j \) is the fitted value for covariate pattern \( x_j \). Then, the Pearson’s residual is:
Table 3.2: Data Classification for Pearson chi-square and Deviance tests

<table>
<thead>
<tr>
<th>Covariate Pattern</th>
<th>y=1</th>
<th>y=0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>x₁</td>
<td>O₁₁</td>
<td>O₁₀</td>
<td>N₁</td>
</tr>
<tr>
<td>x₂</td>
<td>O₂₁</td>
<td>O₂₀</td>
<td>N₂</td>
</tr>
<tr>
<td>. . . . . .</td>
<td>. .</td>
<td>. .</td>
<td></td>
</tr>
<tr>
<td>xₖ</td>
<td>Oₖ₁</td>
<td>Oₖ₀</td>
<td>Nₖ</td>
</tr>
</tbody>
</table>

\[ r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}. \]

And the deviance residual is:

\[ d(y_j, \hat{\pi}_j) = \pm \left\{ 2\left[ y_j \log\left( \frac{y_j}{m_j\hat{\pi}_j} \right) + (m_j - y_j) \log\left( \frac{m_j - y_j}{m_j(1 - \hat{\pi}_j)} \right) \right] \right\}^{1/2}, \]

where the sign, + or −, is the same as the sign of \((y_j - m_j\hat{\pi}_j)\).

The Pearson’s Chi-square statistic and the deviance statistic are then defined as:

\[ \chi^2 = \sum_{i=1}^{J} (y_i - m_i\hat{\pi}_i)^2 / m_i\hat{\pi}_i(1 - \hat{\pi}_i) \quad (3.1) \]

\[ D = 2\left\{ \sum_{i=1}^{J} [y_i \ln(y_i/m_i\hat{\pi}_i) + (m_i - y_i) \ln(m_i - y_i)/(m_i(1 - \hat{\pi}_i))] \right\} \quad (3.2) \]

For a better understanding of \(\chi^2\) and \(D\), we illustrate these two tests in the conceptual framework of a contingency table. The essential idea is to consider the data as described by a \(J \times 2\) contingency table, where the 2 columns are defined by the values of the dichotomous outcome variable \(y\) (\(y = 0\) and \(y = 1\)), and \(J\) rows are defined by the number of possible distinct values of covariate vector \(X\). Table 3.2 presents the data classification for Pearson chi-square and deviance tests.

From the contingency table, \(\chi^2\) and \(D\) can be computed by following equations:
\[
\chi^2 = \sum_{i=1}^{J} \sum_{j=0}^{1} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

(3.3)

\[
D = 2 \sum_{i=1}^{J} \sum_{j=0}^{1} O_{ij} \log \frac{O_{ij}}{E_{ij}},
\]

(3.4)

where \(O_{ij}\) is the observed counts, and \(E_{ij}\) is the expected counts. \(E_{ij}\) is computed by
\(E_{i1} = N_i \hat{\pi}_i\) and \(E_{i0} = N_i - E_{i1}\).

Under the null hypothesis that the fitted logistic regression is correct in all aspects, the asymptotic distribution of \(\chi^2\) and \(D\) is a chi-square distribution with a degrees-of-freedom equal \(J - (p + 1)\). For the deviance test statistics, this statement follows directly from the fact that \(D\) is the likelihood ratio test statistic of a saturated model with \(J\) parameters versus the fitted model with \(p + 1\) parameters. We reject the hypothesis of no significant lack of fit when they exceed the corresponding upper \(\alpha\) critical value.

However, from the view of the contingency table, we know that both of the two tests require that the number of rows in Table 3.2 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. As a result, these two test statistics may not be appropriate for continuous covariates. Osius and Rojek (1992) derived a large-sample normal approximation for the Pearson chi-square test statistics [7], which is usually referred to the scaled Pearson chi-square. The following are the major steps in the computational procedure for this scaled Pearson chi-square as described in Osius and Rojek [7].

1. Save the fitted values from the model, denoted as \(\hat{\pi}_j\), \(j = 1, ..., J\).
2. Create the variable \(v_j = m_j \hat{\pi}_j (1 - \hat{\pi}_j)\), \(j = 1, ..., J\).
3. Create the variable \(c_j = \frac{(1 - 2 \hat{\pi}_j)}{v_j}\), \(j = 1, ..., J\).
4. Compute the Pearson chi-square statistics, namely, \(\chi^2 = \sum_{j=1}^{J} \frac{(y_j - m_j \hat{\pi}_j)}{v_j}\).
5. Perform a weighted linear regression of \(c\), defined in step 3, on \(X\), the model covariates, using weights \(v\), defined in step 2. Note that the sample size for this regression is \(J\), the number of covariate patterns. Let \(RSS\) denote the residual sum-of-square from this regression.
6. Compute the correction factor for the variance, denoted for convenience as \(A\), as...
follows:

\[ A = 2\left(J - \sum_{j=1}^{J} \frac{1}{m_j}\right). \]

7. Compute the standardized test statistic

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

8. Compute the test significance using the standard normal distribution.

In the following, we will use the scaled version of the Pearson’s chi-square test \( \chi^2 \) whenever we are considering continuous variables.

### 3.3 Hosmer and Lemeshow Test

Hosmer and Lemeshow proposed a goodness-of-fit test, now universally referred to as the Hosmer-Lemeshow test in 1980 [8]. The idea applies the concept of the contingency table but creates the table based on a partition of the estimated probability of positive response, \( \hat{\pi}_i \).

First, define a random variable \( W \) where \( w_i = j \), if \( c_{j-1} \leq \hat{\pi}_i < c_j, j = 1, \ldots, g, i = 1, \ldots, n \). The \( c_j \)'s are selected such that the marginal distribution of \( W \) is uniform [8]. That is, the cut points depend on the estimated probability \( \hat{\pi}_i \) and are determined so that \( n/g \) values of \( \hat{\pi}_i \) fall in each interval. This can be easily done. For instance, using \( g = 10 \) groups results in the first group containing the \( n_{1} = n/10 \) subjects having the smallest estimated probabilities, and the last group contains the \( n_{10} = n/10 \) subjects having the largest estimated probabilities.

Denote \( O_{kj} \) be the observed frequency of the occurrence of the pair \((y_i = k, w_i = j)\) in the data set, \( k = 0, 1 \) and \( j = 1, 2, \ldots, g \). \( E_{kj} \) is the expected frequency of the occurrence of the pair \((y_i = k, w_i = j)\), \( k = 0, 1; j = 1, 2, \ldots, g \). The expected frequency is computed by \( E_{1j} = (\sum_{i=1}^{n_j} \hat{\pi}_i) \), the sum of all the estimated probabilities of positive response in the \( j^{th} \) group, and \( E_{0j} = n_j - E_{1j} \), where \( n_j \) is the total number of subjects in the \( j^{th} \) group, \( j = 1, 2, \ldots, g \).

Then, a natural test statistics to calculate is:

\[ H_g = \sum_{j=1}^{g} \sum_{k=0}^{1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}, \quad (3.5) \]

where \( g \) is usually chosen to be 10. When \( g = 10 \) groups are formed, this grouping strategy is referred to “Deciles of Risk”.

10
Table 3.3: Observed (Obs) and Expected (Exp) Counts Within Each Decile of Risk, Defined by Fitted Values (Prob.) for $y = 1$ and $y = 0$

<table>
<thead>
<tr>
<th>Decile</th>
<th>Prob.</th>
<th>y=1 Obs</th>
<th>y=1 Exp</th>
<th>y=0 Obs</th>
<th>y=0 Exp</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.094</td>
<td>4</td>
<td>4.1</td>
<td>54</td>
<td>53.9</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>0.126</td>
<td>5</td>
<td>6.2</td>
<td>52</td>
<td>50.8</td>
<td>57</td>
</tr>
<tr>
<td>3</td>
<td>0.163</td>
<td>8</td>
<td>8.5</td>
<td>50</td>
<td>49.5</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>0.204</td>
<td>11</td>
<td>10.4</td>
<td>46</td>
<td>46.6</td>
<td>57</td>
</tr>
<tr>
<td>5</td>
<td>0.234</td>
<td>16</td>
<td>12.7</td>
<td>42</td>
<td>45.3</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>0.279</td>
<td>11</td>
<td>14.5</td>
<td>46</td>
<td>42.5</td>
<td>57</td>
</tr>
<tr>
<td>7</td>
<td>0.324</td>
<td>18</td>
<td>17.5</td>
<td>40</td>
<td>40.5</td>
<td>58</td>
</tr>
<tr>
<td>8</td>
<td>0.376</td>
<td>24</td>
<td>19.8</td>
<td>33</td>
<td>37.2</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>0.459</td>
<td>23</td>
<td>23.9</td>
<td>35</td>
<td>34.1</td>
<td>58</td>
</tr>
<tr>
<td>10</td>
<td>0.728</td>
<td>27</td>
<td>29.3</td>
<td>30</td>
<td>27.7</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 3.3 presents an example of the $10 \times 2$ contingency table obtained using the decile of risk strategy using maximum likelihood estimates and a logistic model.

Under the null hypothesis that the fitted logistic regression model is correct in all aspects, simulation studies suggest that $H_g$ has an approximate distribution: $\chi^2_{g-2}$ [8]. However, there are some problems: Is it a truly $\chi^2_{g-2}$? How many groups should be formed? Will a different number of groups change the final result?

Table 3.4 illustrates the problem of selecting the number of groups for partitioning the estimated probabilities. The data set we use here is the Framingham Heart Study (males only). There are three covariates in the model: Age, Systolic Blood Pressure (SBP) and Cholesterol (Chol). The outcome variable is whether the subject died from Coronary Heart Disease (CHD). The fitted logistic regression model is

$$
\text{logit}(Pr(\text{death from CHD})) = -6.57 + 0.03 \times Age + 0.008 \times Chol + 0.014 \times SBP
$$

None of the calculated values of the “Hosmer-Lemeshow” test in Table 3.4 suggest that the fitted model demonstrates lack of fit. Nonetheless, the p-values range from $p=0.0916$ (when 8 groups are formed) to $p=0.7695$ (when 15 groups are formed). This suggests that the results of the Hosmer-Lemeshow test depends on the number of groups formed.
Table 3.4: Hosmer-Lemeshow Test Results With Differenting Number of Groups

<table>
<thead>
<tr>
<th>Number of groups((g))</th>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.57</td>
<td>0.3797</td>
</tr>
<tr>
<td>8</td>
<td>10.90</td>
<td>0.0916</td>
</tr>
<tr>
<td>12</td>
<td>12.19</td>
<td>0.2726</td>
</tr>
<tr>
<td>15</td>
<td>9.05</td>
<td>0.7695</td>
</tr>
</tbody>
</table>

With too few groups, we run the risk that we will not have the sensitivity needed to distinguish observed from expected occurrences. Hosmer and Lemeshow point out when \(g\) is fewer than 6 groups it will almost always indicate that the model fits [9].

When ties exist in \(\hat{\pi}_i\), the value of \(H_g\) depends on how these ties are assigned to groups. Different software may use different strategies to handle ties. For example, STATA assigns ties to deciles in such a way that the resulting column totals are as close as to \(n/g\) as possible. Hosmer and Lemeshow suggest that the use of different methods to handle ties by different software is not likely to be an issue unless the number of covariate patterns is so small that assigning all tied values to one group results in a huge imbalance in group size [9]. In the following, we use 10 as the default number of groups.

### 3.4 Unweighted Sum of Squares Test

In 1989, Copas [10] proposed an unweighted sum of squares test for testing proportions. Here, we simplify Copas’ original test statistic and provide its asymptotic distribution under the strict binary situation.

The test statistic is:

\[
S = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2, \tag{3.6}
\]

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

Under the null hypothesis that the fitted logistic regression is correct in all aspects, the asymptotic moments of \(\hat{S}\) are \(E[\hat{S} - \text{trace}(V)] \cong 0\) and \(\text{var}[\hat{S} - \text{trace}(V)] \cong d'(I - M)Vd\). Here, \(d\) is the vector with general element \(d_i = (1 - 2\pi_i)\) and \(V = \text{diag}[\pi_i(1 - \pi_i)]\) is the \(n \times n\) covariance matrix. \(M = VX(X'VX)^{-1}X'\) is the logistic regression version of the hat matrix.
where $X$ is the design matrix. Therefore, after substituting the estimates of $\pi_i$, we can use the standardized statistic \[
\frac{[\hat{S} - \text{trace}(\hat{V})]}{\sqrt{\text{var}[\hat{S} - \text{trace}(\hat{V})]}}\] to assess the significance using the standard normal distribution [3].

### 3.5 Cumulative Sums of Residuals

In 1991 Su and Wei proposed a goodness-of-fit test for the generalized linear model based on the cumulative sums of residuals [2]. The cumulative sums of residuals is built on partitions of “some space” (e.g. covariate space). 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.

We introduce this goodness-of-fit test under the generalized linear model: $y$, the response variable, has density (from the exponential family):

\[
f(y|\theta, \phi) = \exp\{[a(\theta)y - b(\theta)]/\phi + \zeta(y, \phi)\}, \tag{3.7}\]

where $\theta$ is the parameters of interest. Let $X = (1, x_1, ..., x_p)'$ be the $(p + 1) \times 1$ covariate vector. $\mu(\theta)$ is the mean of $y$. $h$ is the link function, that is

\[
h(\mu(\theta)) = \beta'X, \tag{3.8}\]

where $\beta = (\beta_0, ..., \beta_p)'$ is the coefficient vector.

Suppose that the data consists of $n$ independent observations of $(y_i, X_i), i = 1, ..., n$. Then the likelihood score function for $\beta$ takes the form

\[
U(\beta) = \sum_{i=1}^{n} u(\beta'X_i)X_i(y_i - v(\beta'X_i)), \tag{3.9}\]

where $v(r) = h^{-1}(r)$, and $u(r) = \partial\{h \circ \mu(r)\}^{-1}/\partial r$. Define the residuals as $e_i = y_i - v(\hat{\beta}'X_i)$ and consider the following statistic,

\[
W_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_iI(X_i \leq t), \tag{3.10}\]

where $t = (1, t_1, ..., t_p)' \in 1 \times R^p$, and $I(.)$ is the indicator function.
Note that $W_n(t)$ is a function of $t$ 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 test statistic, $G_n = \sup_{t \in \mathbb{R}^p} |W_n(t)|$, leads to the conclusion of model misspecification [2].

Under mild conditions, Su and Wei proved that $G_n$ has the same asymptotic distribution as $\tilde{G}_n = \sup_{t \in \mathbb{R}^p} |\tilde{W}_n(t; \tilde{\beta})|$, where

$$\tilde{W}_n(t; \tilde{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i e_i \{I(X_i \leq t) + \eta'(t; \tilde{\beta}) \mathcal{J}^{-1}(\tilde{\beta}) X_i u(\tilde{\beta}' X_i)\}. \quad (3.11)$$

And

$$\eta(t; \beta) = -\sum_{i=1}^{n} I(X_i \leq t) \partial v(\beta' X_i) / \partial \beta$$

$$\mathcal{J}(\beta) = -\partial U(\beta) / \partial \beta'$$

$U(\beta)$ is the likelihood score function defined in equation 3.9.

Here $Z_1, ..., Z_n$ is a random sample from $N(0, 1)$, independent of $(y_i, X_i, i = 1, ..., n)$. Proof of the weak convergence of $G_n$ and $\tilde{G}_n$ is given in Appendix B.

Let $S_k$ be the set that consists of all the $k^{th}$ components from the observed covariate vectors $X_1, ..., X_n$, and let $S = \Pi_{k=1}^{p} S_k$ be the product space of $S_1, ..., S_p$. Then the test statistic, $G_n = \sup_{t \in \mathbb{R}^p} |W_n(t)| = \sup_{t \in \mathbb{R}^p} |W_n(t)|$. Denote the observed value of $G_n$ by $g_n$. Since $G_n$ has the same asymptotic distribution as $\tilde{G}_n$, the p-value of $G_n$, $Pr(G_n \geq g_n)$, is the same as the p-value of $\tilde{G}_n$, $Pr(\tilde{G}_n \geq g_n)$. Su and Wei presented a simulation method for computing the p-value, $Pr(\tilde{G}_n \geq g_n)$.

Let the $y$, $\beta$ and $\tilde{\beta}$ in $\tilde{W}_n$ be replaced with the observed values, $y$, $\hat{\beta}$ and $\tilde{\beta}$ respectively. Then generate random samples $Z_1, ..., Z_n$ from $N(0, 1)$ independently and compute $\tilde{W}_n$ repeatedly to estimate $Pr(\tilde{G}_n \geq g_n)$. Through simulation studies, Su and Wei show that this large sample approximation to the null distribution of $G_n$ is fairly satisfactory for moderate sample sizes [2].

The proposed test based on $G_n$ is consistent against the alternative hypothesis [2]. 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.
3.5.1 Adaptations of the Cumulative Sums of Residuals Test

In 2002 Lin et. al. proposed two new cumulative sums of residuals goodness-of-fit tests based on Su and Wei’s work [11].

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

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

(3.12)

where $x_{ij}$ is the $j^{th}$ component of $X_i$, $t \in R$. The corresponding test statistic is $G_j = \sup_{t \in R}|W_j(t)|$. Letting the indicator function $I(.)$ be $I((\hat{\beta}'X_i) \leq t), t \in R$. Then, we have

$$W_g(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I((\hat{\beta}'X_i) \leq t)e_i,$$

(3.13)

and the corresponding test statistic is $G_g = \sup_{t \in R}|W_g(t)|$. The p-values of $G_j$ and $G_g$ are still computed based on $G_n$, but with the original indicator function replaced by $I(x_{ij} \leq t)$ and $I((\hat{\beta}'X_i) \leq t)$, respectively.

The first $W_j(t)$ (Equation 3.12) is designed to test the functional form of one particular covariate $x_i, i = 1, ..., p$. The second $W_g(t)$ (Equation 3.13) is more informative about the link function, but in fact it assesses the linearity of the fitted model and the chosen link function at the same time [11]. $W_n(t)$ (Equation 3.10) is usually referred as the omnibus test.

In 2002, Hosmer and Hjort also adapted Su and Wei’s work and proposed a weighted cumulative sums of residuals goodness-of-fit test [4]. They consider the statistics:

$$W_w(t) = \frac{1}{n} \sum_{i=1}^{n} I((\hat{\beta}'X_i) \leq t)w(X_i, \hat{\beta})e_i,$$

where $w(X_i, \hat{\beta})$ is the weight function. The weight functions are chosen so that they maximize the power of the goodness-of-fit test to detect a particular type of alternative to the null model. That is, obtain the weight function that maximizes the value of the non-centrality parameter of the limiting chi-square distribution under a specific alternative [4].

In our work, we will focus primarily on the performance of $W_g(t)$ since it accesses not only the linearity of covariates but also the chosen link function at the same time. Simulation studies suggest that $W_g(t)$ has reasonable power against model misspecification such as the functional forms of covariates and omitted interaction terms [12]. Moreover, the computation
of $W_g(t)$ is much easier than the omnibus test $W_n(t)$. Therefore, in the following work, the cumulative sums of residuals test is referred as $W_g(t)$ unless otherwise specified. And we denote $W_g(t)$ as $W$. For the omnibus test $W_n(t)$, we will observe its performances and compare it to $W_g$ under some relatively simple situations. The functional form test $W_j(t)$ will be used as an additional method for detecting the “true” form of a covariate.

### 3.5.2 Computation of the p-value for $W_n(t)$

At the end of the introduction of the cumulative sums of residuals test, we briefly talked about how to compute the p-value of the omnibus test ($W_n(t)$) efficiently. As we mentioned before, the computation of any of the cumulative sums of residuals tests is time consuming, especially for the omnibus test. Because to compute $I(X \leq t)$, the indicator function in $W_n(t)$, we need to consider all the possible combinations of $t = (t_1, ..., t_p)'$ in multidimensional space. Therefore, one efficient way to compute the p-value of $W_n(t)$ is using Monte Carlo optimization.

Monte Carlo optimization is a computational algorithm that relies on repeated random sampling to compute the results. The common procedure of Monte Carlo optimization is: 1. Define a domain of $t = (t_1, ..., t_p)'$ (note here $W_n(t)$ is a function of $t$); 2. Generate a series of inputs of $t$ randomly from the domain using a certain specified probability distribution such as the normal distribution, or the uniform distribution; 3. Determine an objective function and the optimization goal such as maximum, minimum or least squares; 4. Choose initial values of $t$; 5. Compute the value of the objective function repeatedly until the optimization goal is reached. In the following, we use the “Steam Data” from Su and Wei (1991) as an illustrate example to show how to use Monte Carlo optimization to obtain the p-value.

As shown in Table 3.5, the data consists of 25 observations. Each observation is composed of three components: the pounds of steam used monthly ($y$), the operating days per month ($x_1$), and the average atmospheric temperature ($x_2$) \[2\].

The data was originally analyzed by a simple linear regression. Here, we use the omnibus test ($W_n(t)$) to recheck the fitted linear regression model. Based on the range of $x_1$ and $x_2$, the domain of $t = (t_1, t_2)'$ is $(11, 23)$ for $t_1$ and $(28, 77)$ for $t_2$. Since this is a small sample size and sparse data, the series of $t$ is generated from a Bernoulli distribution. That is the series of $t_1$ chosen from the domain based on $Bernoulli(-1, 1)$ and the series of $t_2$ based on $Bernoulli(-3, 3)$. The objective function is the absolute value of $W_n(t)$ ($\|W_n(t)\|$) and the
optimization goal is to find the maximum. For this case, in each Monte Carlo optimization, we use 250 irritations to find the maximum of $|W_n(t)|$, which is exactly the value of $G_n$. Then repeating the Monte Carlo optimization for 500, we obtain a series values of $G_n$. The final p-value we obtained is 0.044, which is very close to Su and Wei’s result ($p = 0.048$).

### 3.6 Other Tests

There are other goodness-of-fit tests which have been proposed for particular data or model structure. These goodness-of-fit tests are less popular than the tests introduced above.
But they might be more sensitive for some special cases. Hence, in this section, we briefly introduce several of these methods.

### 3.6.1 Stukel’s generalized logistic regression

Stukel (1988) proposed a class of generalized logistic regression models indexed by two shape parameters $\alpha_1$ and $\alpha_2$. 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.

Let $\pi$ be the probability of positive response and $\eta$ be the linear predictor $\beta'X$. The proposed generalization by Stukel is:

$$
\pi_\alpha(\eta) = \frac{e^{h_\alpha(\eta)}}{1 + e^{h_\alpha(\eta)}} \quad (3.14)
$$

or

$$
\log(\pi/(1-\pi)) = h_\alpha(\eta), \quad (3.15)
$$

where the function $h_\alpha(\eta)$ is strictly increasing function of $\eta$ indexed by two shape parameters, $\alpha_1$ and $\alpha_2$. It is defined as follows: For $\eta \geq 0$ ($\pi \geq 1/2$),

$$
h_\alpha = \begin{cases} 
\alpha_1^{-1}(e^{\alpha_1|\eta|} - 1) & \text{if } \alpha_1 > 0 \\
\eta & \text{if } \alpha_1 = 0 \\
-\alpha_1^{-1}\log(1-\alpha_1|\eta|) & \text{if } \alpha_1 < 0
\end{cases}
$$

and for $\eta \leq 0$ ($\pi \leq 1/2$),

$$
h_\alpha = \begin{cases} 
-\alpha_2^{-1}(e^{\alpha_2|\eta|} - 1) & \text{if } \alpha_2 > 0 \\
\eta & \text{if } \alpha_2 = 0 \\
\alpha_2^{-1}\log(1-\alpha_2|\eta|) & \text{if } \alpha_2 < 0
\end{cases}
$$

From the definitions, we note that the proposed generalization separates the standard logistic curve at the point of $\pi = 1/2$. Each shape parameter ($\alpha_1$, $\alpha_2$) governs the behavior of one tail of the curve. When $\alpha_1 = \alpha_2$, the corresponding probability curve $\pi(\eta)$ is symmetric; when $\alpha_1 \neq \alpha_2$, the two tails are treated asymmetrically. Besides, the $h_\alpha(\eta)$ functions are rescaled exponential ($\alpha_1, \alpha_2 > 0$) and logarithmic ($\alpha_1, \alpha_2 < 0$) functions that rise more quickly and slowly, respectively, than the standard logistic curve [13]. And the larger absolute value of $\alpha_1$, $\alpha_2$ indicates the larger deviation. Therefore, clearly, the ordinary logistic model has $\alpha_1 = \alpha_2 = 0$. Additionally, $\alpha_1 = 0.62$, $\alpha_2 = -0.037$ gives the log-log and complementary log-log model; $\alpha_1 = \alpha_2 = 0.165$ gives the probit model.
Through the delta algorithm, the estimates of \((\alpha_1, \alpha_2)\) could be computed from the maximum likelihood estimation. A score test is used to evaluate whether a generalized logistic model is better than a standard model fit to the data. The null hypothesis is:

\[ H_0 : \alpha_1 = \alpha_2 = 0 \]

Let \(\beta\) be the coefficients and \(l(\beta, \alpha_1, \alpha_2)\) \((l)\) be the log-likelihood function from \(n\) observations. Then \(s' = (s_1, s_2) = (\partial l/\partial \alpha_1, \partial l/\partial \alpha_2)\) is the score function evaluated at \((\hat{\beta}, 0, 0)\). Under the null hypothesis, the test statistics \(s'Var(s)^{-1}s\) has an asymptotic \(\chi^2_2\) distribution \([13]\).

### 3.6.2 Redefined Chi-square test and Deviance test by Pulkstenis and Robinson

Pulkstenis and Robinson (2002) proposed two goodness-of-fit tests for logistic regression models with continuous and categorical covariates analogous to the Pearson’s chi-square and deviance tests as well as the “Homser-Lemeshow” test \([14]\). Specifically, they proposed a two-level sub-grouping within each covariate pattern based on fitted probabilities within that covariate pattern, where covariate patterns are determined by only the categorical explanatory variables in the model \([14]\). This model requires sorting all responses by the fitted probabilities within each unique covariate pattern, as defined by only the categorical covariates, and then creating two subcategories within each covariate pattern, essentially splitting the category in two based on the median of fitted probabilities \([14]\). Based on such defined sub-grouping, the proposed test statistics are

\[
\chi^{*2} = \sum_{i=1}^{I} \sum_{h=1}^{2} \sum_{j=1}^{2} \frac{(O_{ihj} - E_{ihj})^2}{E_{ihj}}
\]

\[
D^{*2} = 2 \sum_{i=1}^{I} \sum_{h=1}^{2} \sum_{j=1}^{2} O_{ihj} \log \frac{O_{ihj}}{E_{ihj}}
\]

where \(i\) indexes covariate patterns (patterns of categorical variables), \(h\) indexes the substratification due to ordering by fitted probabilities, and \(j\) indexes columns: \(y = 0\) and \(y = 1\). Through simulations, the two proposed goodness-of-fit tests have the approximate chi-square distribution with degrees of freedom \(2I - k - 2\) under the null hypothesis. Here \(I\) is the
number of the unique covariate patterns and $k$ is the number of categorical variables in the model.

The two tests are proposed to detect omitted interaction between a continuous variable and categorical variable. However, they can only be used for models containing both continuous variables and categorical variables.

### 3.6.3 Smoothed Residual Based Tests

le Cessie and van Houwelingen (1995) proposed a class of tests based on the smoothed residuals. The central idea 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 probability [3].

The smoothed standardized residuals are $\hat{r}_{si} = \sum_{j=1}^{n} w_{ij} \hat{r}_{j}$, where $\hat{r}_{j} = (y_j - \hat{\pi}_j) / \sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}$. The test statistics is

$$\hat{T}_r = \sum_{i=1}^{n} \hat{r}_{si}^2 / \text{vár}(\hat{r}_{si}^2).$$

The weights, $w_{ij}$, are defined by the distance between subject $i$ and $j$, either in the “$X$” space (covariate space) or “$y$” space (outcome space).

Here we list two commonly used weights $w_{ij}$: the uniform kernel in the “$X$” space and a cubic weight in the “$y$” space. Let $X_i = (x_{i1}, ..., x_{ip})$ be the $i^{th}$ observed covariate vector, then the uniform kernel is $w_{ij} = \prod_{k=1}^{p} u(X_{ik}, X_{jk})$, 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 $x_{ik}, i = 1, ..., n$. Choice of the weight constant, $c_u$ is important. Its value used here is the one recommended by le Cessie: $c_u = \frac{1}{2} \left(\frac{4}{n} \right)^{1/(2p)}$ [3]. The cubic weights are given by the equation $w_{ij} = 1 - |\tilde{\pi}_i - \tilde{\pi}_j|^3 / c_{ci}$ if $|\tilde{\pi}_i - \tilde{\pi}_j| \leq c_{ci}$ or equals to zero otherwise. The constant $c_{ci}$ depends on $i$ and is chosen such that $\sqrt{n}$ weights are non-zero for each subject [3].

### 3.6.4 Royston testing procedures

Royston (1992) proposed two goodness-of-fit tests designed to detect departures from linearity in the logistic regression model that use partial sums of residuals. The first test statistics is $P\hat{R}_1 = \max_{1 \leq l \leq n} |q_l|$ where $q_l = -\sum_{i=1}^{l}(y_{(i)} - \hat{\pi}_{(i)})$, $\hat{\pi}_{(i)}$ is the $i^{th}$ largest estimated probability of the positive response and $y_{(i)}$ is the associated value of the outcome
variables [15]. This procedure is used to detect an overall departure from monotonicity in logistic regression. The second test is \( P\hat{R}_2 = \max_{1 \leq t \leq n/2} |q_t - q_{n-t}| \). The second one is often referred as “Royston quadratic” since it was formulated to detect a quadratic departure in the fitted model [3].
CHAPTER 4

COMPARISONS-EMPIRICAL STUDIES

In the previous chapter, we introduced the most important and popular goodness-of-fit (G.O.F) tests for logistic regression and the generalized linear model. Unlike linear regression, there is no standard goodness-of-fit test in logistic regression and the generalized linear model. The goodness-of-fit tests discussed in Chapter 3 all have their individual advantages and disadvantages. Natural questions are under which situations should a particular methodology be preferred and how do we assess their performance under different situations?

In this chapter, we examine the performance of these goodness-of-fit tests in some empirical data sets. The tests that we examine most thoroughly in the following are the “Hosmer-Lemeshow” test \( H - L \), the unweighted sums-of squares test \( S \), Pearson’s Chi-square test \( \chi^2 \) and the cumulative sums of residuals test \( W \). Other tests introduced in Chapter 3 will be examined under some specific situations.

4.1 Motivating Example

To illustrate the questions we first consider a simple example. The data is from the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study, referred to here as the UIS study, and described in more detail in Hosmer and Lemeshow [9]. Briefly, the UIS study was a randomized trial conducted at two sites to study the effect of a residential treatment program of two different durations on prevention of drug abuse and high-risk HIV behavior [9]. The dichotomous outcome variable of interest indicates whether or not the subject remained drug free for the duration of the treatment program. A total of 575 subjects had complete data on model covariates. Among these subjects, 147 remained drug free and 428 returned to drug use. The covariates include age of the subject (AGE), the number of prior drug treatments (NDRGTX), history of IV drug use (IVHX, 1=never, 2=previous, 3=recent) modeled with
Table 4.1: Estimation Results for the linear Main-effect Model (UIS Study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.051</td>
<td>0.0173</td>
<td>0.003</td>
</tr>
<tr>
<td>NDGRTX</td>
<td>-0.061</td>
<td>0.0256</td>
<td>0.016</td>
</tr>
<tr>
<td>IVHX₂</td>
<td>-0.604</td>
<td>0.2872</td>
<td>0.035</td>
</tr>
<tr>
<td>IVHX₃</td>
<td>-0.735</td>
<td>0.2524</td>
<td>0.004</td>
</tr>
<tr>
<td>RACE</td>
<td>0.224</td>
<td>0.2234</td>
<td>0.316</td>
</tr>
<tr>
<td>TREAT</td>
<td>0.442</td>
<td>0.1993</td>
<td>0.027</td>
</tr>
<tr>
<td>SITE</td>
<td>0.146</td>
<td>0.2173</td>
<td>0.501</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.412</td>
<td>0.5535</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.2: P-values of goodness-of-fit tests for the fitted model in Table 4.1

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>χ²</td>
<td>0.128</td>
</tr>
<tr>
<td>H-L</td>
<td>0.522</td>
</tr>
<tr>
<td>S</td>
<td>0.921</td>
</tr>
<tr>
<td>W</td>
<td>0.049</td>
</tr>
</tbody>
</table>

two dummy variables with “never” as the reference group, race (0=white, 1=non-white), treatment group (0=short, 1=long) and treatment site (0=A, 1=B) [9].

To examine the performance of these goodness-of-fit tests we fit the logistic model with linear main effect terms. In the published analysis of this data [9], the final model involved a fractional polynomial transformation for the number of prior dug treatments as well as two interaction terms. So for this data, assuming that the published analysis is correct, we expect the tests to demonstrate a lack of fit for the model containing only the linear main effect terms. Table 4.1 shows the results of the fitted linear main effect model and Table 4.2 presents the results of the goodness-of-fit tests. In Table 4.2, four G.O.F tests give us totally different p-values that vary from .049 to .921; only the cumulative sums of residuals test W yields a statistically significant result. In this case, we would make a different decision depending on the test we employed.

As noted above, the final published model for the UIS study involves a fractional polynomial transformation and two interaction terms. Assuming that the published model
is correct, for this particular situation, the cumulative sums of residuals test is most sensitive
and powerful with respect to discerning model misspecification in the presence of the need for a transformation and interaction terms. However, how generalizable this empirical example is still uncertain. We present additional examples in the following.

4.2 Empirical Studies

4.2.1 UIS Study, Continued

First, we continue the model fitting process for UIS study. As noted above, the final model in textbook by Hosmer and Lemeshow contained non-linear terms and interactions [9]. Hence, we continue our examination with a refitted model with transformed non-linear term.

Model with non-linear term

In the textbook, variable “NDRGTX” is suggested to have a 2-degree fractional polynomial transformation. Hence, this time, we refitted the main effect model but with “NDRGTX” replaced by: 1) \(numtreat1 = ((numtreat + 1)/10)^{-1}\); 2) \(numtreat2 = numtreat1 \times log((numtreat + 1)/10)\). And recalculated the goodness of fit statistics. Table 4.3 presents the test results of the four methods. None of these G.O.F tests reject the null hypothesis that the refitted model (with non-linear term) is correct in all aspects. The variability of the resulting p-values is much less than previously, but \(W\) still results in the smallest p-value.

Final Model

In the final model of the UIS study previously presented [9], two interaction terms were deemed necessary: 1) “agenumtreat1”: interaction between age and “NDRGTX”; 2) “racesite”: interaction between race and site. Table 4.4 shows the estimation results, and

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\chi^2)</td>
<td>0.342</td>
</tr>
<tr>
<td>H-L</td>
<td>0.352</td>
</tr>
<tr>
<td>S</td>
<td>0.172</td>
</tr>
<tr>
<td>W</td>
<td>0.110</td>
</tr>
</tbody>
</table>
Table 4.4: Estimation Results for the Final Model (UIS Study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.051</td>
<td>0.0173</td>
<td>0.003</td>
</tr>
<tr>
<td>numtreat1</td>
<td>1.669</td>
<td>0.4071</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>numtreat2</td>
<td>0.433</td>
<td>0.1169</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>IVHX</td>
<td>-0.604</td>
<td>0.2872</td>
<td>0.035</td>
</tr>
<tr>
<td>IVHX3</td>
<td>-0.735</td>
<td>0.2524</td>
<td>0.004</td>
</tr>
<tr>
<td>RACE</td>
<td>0.224</td>
<td>0.2234</td>
<td>0.316</td>
</tr>
<tr>
<td>TREAT</td>
<td>0.442</td>
<td>0.1993</td>
<td>0.027</td>
</tr>
<tr>
<td>SITE</td>
<td>0.146</td>
<td>0.2173</td>
<td>0.501</td>
</tr>
<tr>
<td>agenumtreat1</td>
<td>-0.015</td>
<td>0.0060</td>
<td>0.011</td>
</tr>
<tr>
<td>racesite</td>
<td>-1.429</td>
<td>0.5298</td>
<td>0.007</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.412</td>
<td>0.5535</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.5: P-values of goodness-of-fit tests for the Final Model (UIS Study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.631</td>
</tr>
<tr>
<td>H-L</td>
<td>0.827</td>
</tr>
<tr>
<td>S</td>
<td>0.956</td>
</tr>
<tr>
<td>W</td>
<td>0.722</td>
</tr>
</tbody>
</table>

Table 4.5 presents the corresponding test results. All of the G.O.F tests support that the model fits the data. For this particular study the cumulative sums of residuals test $W$ is more sensitive to model misspecification, but none of the tests showed a significant result when the interaction terms were omitted.

In this study, test results of $W$ were computed from 1000 realizations of the zero-mean Gaussian process. Figure 4.1 presents the process of $W$: the black curves are 20 simulated realizations for $W$ and the dark grey curve is the observed $W$. If the test result is insignificant, the observed curve will fall within the range of the simulated curves.

4.2.2 Small Data Sets

In this section we present examples of empirical studies all of which have small sample size. Since the test results of all these G.O.F tests depend on their asymptotic distributions, these
Figure 4.1: Cumulative Sums of Residuals for UIS Study. **Top:** Cumulative Sums of Residuals for the main effect model; **Bottom:** Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. 26
studies provide us with an examination of their performance with small sample size.

**Low Birth Weight Study**

The “Low Birth Weight Study” (LBW study), also comes from the textbook by Hosmer and Lemeshow [9]. These data were collected at Baystate Medical Center in Springfield, Massachusetts in 1986 [9]. The outcome variable is whether birth weight is less than 2500 grams. Data on 189 births was collected, of which 59 were low birth weight and 130 were normal birth weight. The independent variables considered are: age of the mother (AGE), weight of the mother at the last menstrual period (LWT), race of the mother (White, Black, Others, coded with two dummy variables “RACE$$_1$$” and “RACE$$_2$$”, with White as the reference category), and whether the mother smoked (SMOKE, 1 = yes, 0 = no) [9].

Hosmer et al. analyzed this study and suggested that the final model should contain the main effects and two interaction terms: 1) “ageweight”: interaction between “AGE” and “LWT”; 2) “weightsmoke”: interaction between “SMOKE” and “LWT” [3].

To examine performance of the goodness of fit tests, we again start from the main-effects model with linear predictors. Table 4.6 presents the results of the fitted main-effects model and Table 4.7 presents the corresponding test results. (Although statistically insignificant, Hosmer et al. suggest keeping “AGE” in the regression model because of its clinical importance.)

For the main effect model, three of the tests would not reject the null hypothesis that the fitted model is correct in all aspects. Only the cumulative sums of residuals test, W, results in a statistically significant departure from the main effect model ($p = 0.032$).

We refit the model based on the original authors’ suggestion. Table 4.8 presents the new

---

### Table 4.6: Estimation Results for Main-Effect Model (LBW study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>-0.022</td>
<td>0.034</td>
<td>0.511</td>
</tr>
<tr>
<td>LWT</td>
<td>-0.013</td>
<td>0.006</td>
<td>0.050</td>
</tr>
<tr>
<td>RACE$_1$</td>
<td>1.236</td>
<td>0.517</td>
<td>0.017</td>
</tr>
<tr>
<td>RACE$_2$</td>
<td>0.946</td>
<td>0.416</td>
<td>0.023</td>
</tr>
<tr>
<td>SMOKE</td>
<td>1.054</td>
<td>0.380</td>
<td>0.006</td>
</tr>
<tr>
<td>Constant</td>
<td>0.332</td>
<td>1.118</td>
<td>0.764</td>
</tr>
</tbody>
</table>
Table 4.7: P-values of goodness-of-fit tests for the Main-Effect Model (LBW study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.102</td>
</tr>
<tr>
<td>H-L</td>
<td>0.349</td>
</tr>
<tr>
<td>S</td>
<td>0.071</td>
</tr>
<tr>
<td>W</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 4.8: Estimation Results for the Final Model (LBW study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.028</td>
<td>0.163</td>
<td>0.860</td>
</tr>
<tr>
<td>LWT</td>
<td>-0.01</td>
<td>0.030</td>
<td>0.722</td>
</tr>
<tr>
<td>$RACE_1$</td>
<td>1.230</td>
<td>0.519</td>
<td>0.017</td>
</tr>
<tr>
<td>$RACE_2$</td>
<td>0.878</td>
<td>0.426</td>
<td>0.039</td>
</tr>
<tr>
<td>SMOKE</td>
<td>-0.552</td>
<td>1.683</td>
<td>0.743</td>
</tr>
<tr>
<td>ageweight</td>
<td>-0.0004</td>
<td>0.001</td>
<td>0.749</td>
</tr>
<tr>
<td>weightsmoke</td>
<td>0.013</td>
<td>0.012</td>
<td>0.328</td>
</tr>
<tr>
<td>Constant</td>
<td>0.332</td>
<td>1.118</td>
<td>0.511</td>
</tr>
</tbody>
</table>

Table 4.9: P-values of goodness-of-fit tests for the Final Model (LBW study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.922</td>
</tr>
<tr>
<td>H-L</td>
<td>0.768</td>
</tr>
<tr>
<td>S</td>
<td>0.317</td>
</tr>
<tr>
<td>W</td>
<td>0.528</td>
</tr>
</tbody>
</table>

estimation results and Table 4.9 shows the corresponding test results. All the tests support that the fitted logistic regression model is correct. Figure 4.2 presents the results of $W$, which were computed from 500 realizations of the zero-mean Gaussian process.

However, we found out that the two suggested interaction terms are not statistically significant based on the likelihood ratio test ($p = 0.85$) and that the fractional polynomials method in STATA also supports the linearity of “AGE” and “LWT” in the logistic regression.
Figure 4.2: Cumulative Sums of Residuals for LBW Study. **Top**: Cumulative Sums of Residuals for the main effect model; **Bottom**: Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. 

29
Therefore, for this particular study, we could not reproduce the published results. A possible explanation is that these two interaction terms have important clinical meaning. But, purely from the point of view of model fitting, it seems that the cumulative sums of residuals test $W$ does not give us the expected result for this particular study. To examine the case further, we also fit an intermediate model with one interaction term ("ageweight") only. For this model, the p-value from $W$ test is 0.064, a border line case.

**Intensive Care Unit Study**

The second study we examine is the “Intensive Care Unit Study” (ICU study). This data set is part of a large study on survival of patients following admission to an adult intensive care unit (ICU) [9]. The major goal of this study is to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. A total of 200 subjects were included, of which 40 died and 160 lived and were discharged from the hospital. The independent variables are: age of the subject at ICU admission ("AGE"), systolic blood pressure at ICU admission ("SBP"), heart rate at ICU admission ("HRA"), cancer part of present problem (1 = yes, 0 = no, “CAN”), previous admission to an ICU (1 = yes, 0 = no, “PRE”), type of admission (1 = Emergency, 0 = Elective, “TYP”), PH from initial blood gases (0 = $\geq$ 7.25, 1 = $<$ 7.25, “PH”), PCO2 from initial blood gases (0 = $\leq$ 45, 1 = $>$ 45, “PCO2”), level of consciousness at ICU admission (0 = No coma/Deep Stupor, 1 = Deep stupor, 2 = Coma, “CONS”).

In the literature, a number of publications have appeared that analyze the original study. The most recent ones are Lemeshow et al. (1993) [16] and Lemeshow and Le Gall (1994) [17]. With their findings as our reference, we attempt to find the best model for this small data set. First of all, following previous analyses, we use categorical versions of independent continuous variables other than “AGE”. In the literature, “HRA” is grouped into greater than 150 beats/min ("HRA" = 1) or less ("HRA" = 0) and “SBP” is grouped into less than 90 mmHg (“SBP” = 1) or greater (“SBP” = 0). However, in our data set, only 8 subjects had “HRA” greater than 150 beats/min, and “HRA” is highly insignificant from the Wald test in the initial screening, so we remove “HRA” from the independent variable set. In addition, only 5 out of the 200 subjects had deep stupor and 2 out of 200 had Coma at ICU admission. We regrouped “CONS” into two groups: 1 is “no coma or deep stupor”; 2 is “coma or deep stupor”. Finally, “AGE” is the only continuous variable in the data set,
Table 4.10: Estimation Results for the Final Model (ICU study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.039</td>
<td>0.014</td>
<td>0.005</td>
</tr>
<tr>
<td>SBP</td>
<td>1.859</td>
<td>0.722</td>
<td>0.010</td>
</tr>
<tr>
<td>CAN</td>
<td>2.353</td>
<td>0.922</td>
<td>0.011</td>
</tr>
<tr>
<td>TYP</td>
<td>2.922</td>
<td>0.952</td>
<td>0.002</td>
</tr>
<tr>
<td>PH</td>
<td>1.730</td>
<td>0.875</td>
<td>0.048</td>
</tr>
<tr>
<td>PCO2</td>
<td>-2.371</td>
<td>0.976</td>
<td>0.015</td>
</tr>
<tr>
<td>CONS</td>
<td>4.189</td>
<td>0.982</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Constant</td>
<td>-7.034</td>
<td>1.399</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 4.11: P-values of goodness-of-fit tests for the Final Model (ICU study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.656</td>
</tr>
<tr>
<td>H-L</td>
<td>0.684</td>
</tr>
<tr>
<td>S</td>
<td>0.725</td>
</tr>
<tr>
<td>W</td>
<td>0.537</td>
</tr>
</tbody>
</table>

and the lower level of all categorical variables are treated as the reference sets.

With the literature references and our own findings, the final model for the ICU study is the main effect model with one continuous variable ("AGE") and 6 categorical variables. Table 4.10 presents the results for the fitted model, and Table 4.11 presents the corresponding goodness of fit test results. All of the test results support that the fitted logistic regression model is correct in all aspects and in this example the p-values are close to each other. Figure 4.3 presents the results of $W$, which were computed from 500 realizations of the zero-mean Gaussian process.

4.2.3 Large Sample Size Study

National Health and Nutrition Examination Study III

As the last section of the empirical studies, we examine the performance of these G.O.F tests in a study with a large sample size. The data we used is a subset data from the National Health and Nutrition Examination Survey III (NHANES III), conducted by the National
Figure 4.3: Cumulative Sums of Residuals for ICU Study. The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$.

Center for Health Statistics between 1988 and 1994. Hosmer and Lemeshow analyzed this original survey data as a numerical example for logistic regression with complex sample surveys. For our analysis, we randomly sampled 3000 records from the original data set, conducted the model fitting, and compared the performance of the goodness-of-fit tests.

The purpose of this study is to use the logistic regression to model the relationship between reported high blood pressure and other clinical factors. A total of 3000 subjects were included, of which 582 subjects reported having high blood pressure in the survey. There were 6 clinical factors: age ("AGE"), sex ("SEX", 0 = Female, 1 = Male), race ("RACE", 1 = White, 2 = Black, treated as a dichotomous variable with 1 = White as the reference set), body weight ("WEIGHT"), standing height ("HEIGHT"), and smoking
As the first step of model fitting, as previously we fitted the main-effects model for the data. The variable “SMOKE” was not statistically significant in this data set and was removed from the regression model. Table 4.12 presents the results of the main-effect model and Table 4.13 presents the results of the goodness-of-fit tests.

All the G.O.F. tests suggest there is significant lack of fit for the main-effect model (all $p < 0.001$). In the Hosmer and Lemeshow analysis (based on a different version of the data), they suggested that the linear form was not appropriate for the “AGE” variable or the “WEIGHT” variable [9]. The fractional polynomials method in STATA also suggested that there should be a cubic power for the “AGE” variable and the linear term of “WEIGHT” is appropriate in our study. Therefore, a cubic term for “AGE” is included in our model (“AGE1”: the linear term, “AGE3”: the cubic term $AGE^3$). Table 4.14 presents the testing results for the modified model. After these changes to the model, the cumulative sums of residuals test, $W$, still yields a significant result ($p = 0.021$), which suggests that the modified model may need more modification. Despite that, we note that the p-value from

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>0.071</td>
<td>0.003</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SEX</td>
<td>0.356</td>
<td>0.151</td>
<td>0.018</td>
</tr>
<tr>
<td>RACE</td>
<td>0.557</td>
<td>0.125</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>0.008</td>
<td>0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>-0.06</td>
<td>0.021</td>
<td>0.003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>H-L</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>S</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>W</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Table 4.14: P-values of goodness-of-fit tests for the Transformed Model (NHANES III study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.561</td>
</tr>
<tr>
<td>H-L</td>
<td>0.633</td>
</tr>
<tr>
<td>S</td>
<td>0.121</td>
</tr>
<tr>
<td>W</td>
<td>0.021</td>
</tr>
</tbody>
</table>

Table 4.15: Estimation Results for the Final Model (NHANES III study)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE1</td>
<td>0.218</td>
<td>0.025</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AGE2</td>
<td>-0.0008</td>
<td>0.0002</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>SEX</td>
<td>0.396</td>
<td>0.151</td>
<td>0.009</td>
</tr>
<tr>
<td>RACE</td>
<td>0.260</td>
<td>0.359</td>
<td>0.468</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>0.003</td>
<td>0.0005</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HEIGHT</td>
<td>-0.008</td>
<td>0.002</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>agerace</td>
<td>-0.0025</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>weightrace</td>
<td>-0.0008</td>
<td>0.0003</td>
<td>0.009</td>
</tr>
</tbody>
</table>

The unweighted sums of squares test, $S$, is only 0.121, which is also a weak evidence for the model adequacy. Based on this information (Table 4.14), as well as the possibility of interactions among those independent variables, we build our final models.

The final model includes a cubic term of “AGE” and two interaction terms: interaction between “AGE” and “RACE” (“agerace”), and interaction between “WEIGHT” and “RACE” (“weightrace”). Table 4.15 presents the results of our final model in this NHANES III study and the corresponding test results are presented in Table 4.16. For this final model, all of the G.O.F tests suggest the fitted model sufficiently fits the data. Overall, for this particular study, the cumulative sums of residuals test, $W$, appears most sensitive to the omitted interaction terms. Figure 4.4 presents the process of $W$. In this study, test results of $W$ were computed from 6000 realizations of the zero-mean Gaussian process. The black curves are 20 simulated realizations for $W$ and the dark grey curve is the observed $W$. We can see that the observed curve falls within the range of the simulated curves suggesting the test result is insignificant.
Figure 4.4: Cumulative Sums of Residuals for NHANES III Study. **Top:** Cumulative Sums of Residuals for the main effect model; **Bottom:** Cumulative Sums of Residuals for the final model; The grey curve stands for the observed $W$, and the black curve stands for the simulated $W$. 
Table 4.16: P-values of goodness-of-fit tests for the Final Model (NHANES III study)

<table>
<thead>
<tr>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>0.532</td>
</tr>
<tr>
<td>H-L</td>
<td>0.941</td>
</tr>
<tr>
<td>S</td>
<td>0.826</td>
</tr>
<tr>
<td>W</td>
<td>0.638</td>
</tr>
</tbody>
</table>

4.3 Conclusions

In this Chapter, we examined the performance of the G.O.F tests in a series of empirical studies. Generally speaking, these G.O.F tests all support the final version of the fitted models (the p-values of these tests for the final model in each studies all exceed 0.05.)

However, there are several kinds of departures from the fitted model in which the performance of these G.O.F tests is of interest. The two major departures we examined in our the empirical examples were the need for transformation (non-linearity on the logit scale) and the need for interaction terms. Based on these empirical examples, the cumulative sums of residuals test $W$ appears to have the best performance for detecting these departures from fit. $W$ gave us a significant result in most cases where a non-linearity of logit departure was suggested. Similarly, in the “NHANES III study” where we suspected, based on previous analysis, that there was a need for interactions terms, $W$ gave us a significant result while the other three tests all accepted the fitted model. Therefore, we expect the cumulative sums of residuals test $W$ to perform better than the other three tests. Based on these empirical examinations, the unweighted sum of squares test $S$ performs less well than $W$ but better than the other two tests.

Also based on these empirical examples, for all G.O.F tests considered, it appears more difficult to detect an omitted interaction term than a non-linear term in the fitted model. In addition, we found out that a p-value less than 0.15 of these G.O.F tests may suggest a lack of fit of the model, in initial steps of model fitting.

However, we can not determine the true model in empirical studies so it is difficult to generalize our results to determine which method performs better in a given situation. Therefore, we examine these issues further using simulation studies in the next chapter.
CHAPTER 5

COMPARISONS-SIMULATION STUDIES

In the previous chapter, we examined the performance of these goodness-of-fit tests in several empirical studies which included small and large sample sizes, multiple categorical variables, and required one or both non-linear terms and interactions. Based on these examples, the ability to detect departures from the fitted model varies somewhat. Based only on these empirical data sets it is difficult to say which test is more powerful or superior to others for one or two particular types of departures. Therefore, in this Chapter, we access the performance (size and power) of these G.O.F tests under some particular simulated settings.

Two major aspects of the performance of the goodness-of-fit tests that can not be inferred from the empirical examples are: 1) Size of the test: the Type I error of the G.O.F test when the proposed model is right; 2) Power of the test: the probability of the G.O.F test to detect deviation from the specified model when the model is incorrect. In this chapter, we present the results of series of simulation scenarios designed to address these two aspects of the tests. Sample size of $n = 100, 500, 1000$ are chosen to represent small, moderate and the large sample size studies. In all situations, we use 1000 replications to compute the size and power of the test.

Since all the four methods ($H - L$, $S$, $\chi^2$, and $W$) are mainly proposed for the logistic regression model, our simulation scenarios are designed for this model. We consider a binary outcome $y$. We denote by $\pi(x)$ as the probability of a positive response ($y = 1$) and by $\eta(x) = \beta'x$ as the linear predictor. We denote the estimated probability of a positive response as $\hat{\pi}(x)$.

In our simulation, the outcome variable $y$ is generated by comparing an independently generated $U(0, 1)$ random variable $u$, to the true logistic probability using the rule $y = 1$ if $u \leq \pi(x) = \exp(\eta(x))/1 + \exp(\eta(x))$ and $y = 0$ otherwise.
5.1 Null Distribution

We began by examining the empirical size of the goodness of fit tests when the logistic model is known to be correct. Table 5.1 displays the settings used for this examination. The choice of these settings is based on our experience in the empirical studies. We consider two aspects of the model, the distribution of the covariates and the values of the parameters in the model. We consider the four univariate models, two bivariate models and two multivariate models described in Table 5.1. Note that the outcome variable $y$ is generated depending on $\pi(x)$, and $\pi(x)$ is computed from the linear predictor, $\eta(x)$. Therefore, the key point is the possible distribution of $\eta(x)$. From previous studies, we found that the most common distributions of $\eta(x)$ are bell-shaped, left-skewed or right-skewed no matter how many covariates are fitted (Figure 5.1).

Therefore, to represent these common distributions of $\eta(x)$, the covariate(s) $x$ is generated from the normal distribution, the uniform distribution, the chi-square distribution, the Bernoulli distribution and their combinations. The coefficients $\beta$ are chosen to assure sufficient incidence to allow estimation. Table 5.1 describes the null situations where the distribution of the covariate(s) is given, along with the true coefficients for the logistic model and the average expected values for the minimum, maximum, and three quartiles for the resulting distribution of logistic probabilities for a sample size of 100. That is, for example, $Q_1$ in Table 5.1 is the average value of the first quartiles in all simulated distributions with a sample size of 100. Among all the settings, the chi-square distribution yields the most highly right skewed distribution.
Figure 5.1: Common Situations for $\eta(x)$
Table 5.1: Settings used to examine the null distribution

<table>
<thead>
<tr>
<th>Covariate distribution</th>
<th>Logistic Coefficients</th>
<th>Quartiles of the logistic probabilities (n = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min.</td>
</tr>
<tr>
<td>$N(0, 1.5)$</td>
<td>$\beta_0 = 0, \beta_1 = 0.8$</td>
<td>0.060</td>
</tr>
<tr>
<td>$U(-6, 6)$</td>
<td>$\beta_0 = 0, \beta_1 = 0.8$</td>
<td>0.009</td>
</tr>
<tr>
<td>$\chi^2(2)$</td>
<td>$\beta_0 = -2.5, \beta_1 = 1$</td>
<td>0.076</td>
</tr>
<tr>
<td>$\chi^2(4)$</td>
<td>$\beta_0 = -4.9, \beta_1 = 0.65$</td>
<td>0.007</td>
</tr>
<tr>
<td>Independent $N(0, 1.5)$ and $B(0.5)$</td>
<td>$\beta_0 = 0, \beta_1 = 0.8, \beta_2 = 0.2$</td>
<td>0.013</td>
</tr>
<tr>
<td>Independent $\chi^2(4)$ and $B(0.5)$</td>
<td>$\beta_0 = 1, \beta_1 = -0.4, \beta_2 = 0.5$</td>
<td>0.001</td>
</tr>
<tr>
<td>Independent $N(0, 1.5)$ and $U(-6, 6)$ and $\chi^2(4)$</td>
<td>$\beta_0 = -1.3, \beta_1 = \beta_2 = 0.8/3, \beta_3 = 0.65/3$</td>
<td>0.068</td>
</tr>
<tr>
<td>Normal-Bernoulli model$^1$</td>
<td>$\beta_0 = 0, \beta_1 = 0.8, \beta_2 = -0.8, \beta_3 = \log(2)$</td>
<td>0.019</td>
</tr>
</tbody>
</table>
Table 5.2 shows the percentage of time each of the goodness-of-fit test rejects the hypothesis of fit at the $\alpha = 0.05$ level. Overall, with a sample size of 500 or up, all G.O.F tests have about the right empirical size when we use $\alpha = 0.05$. However, with a small sample size of 100, the results are less clear as might be expected since all of the tests rely on asymptotic results.
Table 5.2: Simulated percent rejection at the $\alpha = 0.05$ for Null Distribution

<table>
<thead>
<tr>
<th>Setting/ Sample size</th>
<th>Pearson-chi-square $\chi^2$</th>
<th>Unweighted sum of squares $S$</th>
<th>Hosmer-Lemeshow $H - L$</th>
<th>Cumulative Sums of Residuals $W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 500 1000</td>
<td>100 500 1000</td>
<td>100 500 1000</td>
<td>100 500 1000</td>
</tr>
<tr>
<td>$N(0, 1.5)$</td>
<td>4.2 5.0 3.8 5.0 6.4</td>
<td>4.6</td>
<td>6.2 4.4 4.6</td>
<td>5.6 5.6 5.8</td>
</tr>
<tr>
<td>$U(-6, 6)$</td>
<td>3.6 4.8 4.4 4.8 5.8</td>
<td>5.8</td>
<td>5.8 4.8 6.2</td>
<td>4.8 5.2 5.6</td>
</tr>
<tr>
<td>$\chi^2(2)$</td>
<td>3.3 3.2 4.6 3.4 5.4</td>
<td>4.6</td>
<td>4.4 5.4 4.2</td>
<td>3.6 4.2 4.8</td>
</tr>
<tr>
<td>$\chi^2(4)$</td>
<td>3.4 4.1 3.8 3.6 5.2</td>
<td>5.4</td>
<td>7.4 6.6 4.6</td>
<td>6.6 5.0 5.2</td>
</tr>
<tr>
<td>Independent $N(0, 1.5)$ and $B(0.5)$</td>
<td>3.8 4.2 4.5 4.4 4.2</td>
<td>4.8</td>
<td>6.8 4.8 5.4</td>
<td>5.2 5.0 4.8</td>
</tr>
<tr>
<td>Independent $\chi^2(4)$ and $B(0.5)$</td>
<td>3.6 5.0 6.2 5.2 6.5</td>
<td>4.4</td>
<td>4.2 3.6 3.8</td>
<td>5.0 4.8 6.2</td>
</tr>
<tr>
<td>Independent $N(0, 1.5)$ and $U(-6, 6)$ and $\chi^2(4)$</td>
<td>3.4 4.4 5.0 3.6 4.8</td>
<td>4.2</td>
<td>4.4 4.4 5.0</td>
<td>6.2 5.8 5.2</td>
</tr>
<tr>
<td>Normal-Bernoulli model</td>
<td>4.4 5.0 5.3 5.6 4.6</td>
<td>5.0</td>
<td>3.4 4.8 5.2</td>
<td>4.8 5.2 6.0</td>
</tr>
</tbody>
</table>
Table 5.3: Coefficients for detecting quadratic form of $x$

<table>
<thead>
<tr>
<th>Settings</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\beta_0 = -0.12, \beta_1 = 5, \beta_2 = 1$</td>
</tr>
<tr>
<td>2</td>
<td>$\beta_0 = -0.5, \beta_1 = 3, \beta_2 = 1$</td>
</tr>
<tr>
<td>3</td>
<td>$\beta_0 = -0.5, \beta_1 = 2, \beta_2 = 1$</td>
</tr>
<tr>
<td>4</td>
<td>$\beta_0 = -0.5, \beta_1 = 1, \beta_2 = 1$</td>
</tr>
</tbody>
</table>

5.2 Functional Form of Covariates

In this section, we evaluate the power of these tests to detect a series of non-linear functional forms of the covariates, which we might encounter in practice.

5.2.1 Quadratic Form of $x$

First, we examine the power of the goodness of fit tests to detect a quadratic departure from the linear form of the model. We generate the outcome variable by using a logistic model with logit link $\text{logit}(\pi(x)) = \eta(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, but fit the generated data with the linear term of $x$ only. The distribution of $x$ is $N(0, 1.5)$, and the values of the coefficients $(\beta_0, \beta_1, \beta_2)$ are chosen such that the lack of linearity in the fitted logistic model on the logit scale becomes progressively more pronounced. Table 5.3 presents the chosen settings and Figure 5.2 demonstrates that the departure from linearity becomes progressively more pronounced with the settings.

Table 5.4 presents the power, the percent of time each of these G.O.F tests rejects the hypothesis of fit, at the $\alpha = 0.05$ level. From this table, we see that all of these tests have poor power (rejection $\leq 35\%$) to detect the model which is quite close to the linear logistic model (Setting 1), even with a sample of size 1000. As the departure from linearity in the fitted logistic model increases, the power increases rapidly. And also, as the sample size increases, the power increases rapidly. For a moderate departure from the linear model (Setting 2 and 3) and a moderate sample size ($n = 500$), all G.O.F tests have a good power for rejection (probability of rejection $\geq 76\%$). Moreover, all tests have a nearly 100% power to detect the extreme case (Setting 4). Overall, the $S$ test and the $\chi^2$ test yield similar performance, and have slightly higher power for detecting a quadratic departure than the
Figure 5.2: Four Settings For the Quadratic Term

$H - L$ test or the $W$ test.

In addition, we also examine the power of these G.O.F tests for one more case with an omitted cubic term. The true logistic regression model is $\text{logit}(\pi(x)) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, but the fitted model omits the cubic term. The $x$ is from $N(0, 1.5)$, and coefficients are chosen as $\beta_0 = 2, \beta_1 = 3.935, \beta_2 = -4.5, \beta_3 = 1$ (See Figure 5.5). Table 5.5 presents the power, the percent of time each of these G.O.F tests rejects the hypothesis of fit, at the $\alpha = 0.05$ level. For this case, the $S$ test has the best performance, while the $\chi^2$ test has the poorest power. With a large sample size, except the $\chi^2$ test, others all have good power for rejection.
Table 5.4: Percentage Rejection with $\alpha = 0.05$ for Omission of Quadratic Term, 1000 Simulations

<table>
<thead>
<tr>
<th>Settings</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>7.4</td>
<td>22.6</td>
<td>34.8</td>
<td>6.2</td>
</tr>
<tr>
<td>2</td>
<td>49.6</td>
<td>76.4</td>
<td>99.8</td>
<td>59.2</td>
</tr>
<tr>
<td>3</td>
<td>51.2</td>
<td>97.6</td>
<td>100.0</td>
<td>57.8</td>
</tr>
<tr>
<td>4</td>
<td>64.0</td>
<td>94.6</td>
<td>100.0</td>
<td>78.0</td>
</tr>
</tbody>
</table>

Figure 5.3: Setting for Cubic term of $x$
Table 5.5: Simulated percent rejection at the $\alpha = 0.05$ level for the omitted cubic term

<table>
<thead>
<tr>
<th>$n$</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>8.7</td>
<td>11.8</td>
<td>1.1</td>
<td>6.3</td>
</tr>
<tr>
<td>500</td>
<td>47.6</td>
<td>54.6</td>
<td>3.5</td>
<td>48.2</td>
</tr>
<tr>
<td>1000</td>
<td>75.6</td>
<td>84.0</td>
<td>4.5</td>
<td>74.5</td>
</tr>
</tbody>
</table>

Table 5.6: Settings for non-linearity forms of $x$

<table>
<thead>
<tr>
<th>Settings</th>
<th>Coefficients for $\chi^2(4)$</th>
<th>Coefficients for $U(1, 20)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log(x)$</td>
<td>$\beta_0 = -1.7, \beta_1 = 1.2$</td>
<td>$\beta_0 = -1.5, \beta_1 = 0.6$</td>
</tr>
<tr>
<td>$\sqrt(x)$</td>
<td>$\beta_0 = -1.7, \beta_1 = 1.2$</td>
<td>$\beta_0 = -1.5, \beta_1 = 0.6$</td>
</tr>
<tr>
<td>$x^{-1}$</td>
<td>$\beta_0 = -0.5, \beta_1 = 1.25$</td>
<td>$\beta_0 = -0.3, \beta_1 = 1.5$</td>
</tr>
<tr>
<td>$x^{-1/2}$</td>
<td>$\beta_0 = -2.5, \beta_1 = 4.25$</td>
<td>$\beta_0 = -1.5, \beta_1 = 3$</td>
</tr>
</tbody>
</table>

5.2.2 Other Non-Linear Forms of Covariates

In this section, we examine four additional non-linear forms: $\log(x)$, $\sqrt{x}$, $x^{-1}$ and $x^{1/2}$. Here, the distribution of $x$ is chosen from 1) $\chi^2(4)$, since the log transform and square root transform are mostly used in right-skewed data; 2) $U(1, 20)$, since we want to see whether the covariate’s distribution will affect the power of these G.O.F. tests. Table 5.6 presents settings for the four non-linear forms.

First, let’s look at their performance when $x$ is from a distribution of $\chi^2(4)$. Table 5.7 presents the power, the percent of time each of the tests rejects the hypothesis of fit, at the $\alpha = 0.05$ level. Similar to the quadratic situation, with a sample size of 100, none of these tests have a good power to detect any of these specified non-linear forms of $x$. Nevertheless, with a sample size of 500, the $W$ test has a power of 66.2% for “log” form of $x$, and a power of 80.4% for “inverse” form of $x$; The $H - L$ test has a power of 51% for “log” form of $x$ and a power of 69.6% for “inverse” form of $x$. With a sample size of 1000, the power of the $W$ test goes up to 92.4% for “log” form of $x$ and up to 99.6% for “inverse” form of $x$; The power of the $H - L$ test goes up to 84.6% for “log” form of $x$ and up to 98% for “inverse” form of $x$. All tests have very high power to detect a form of $x^{-1/2}$ with a sample size of 500 or up but low power to detect a form of $x^{1/2}$.
Table 5.7: Simulated percent rejection at the $\alpha = 0.05$ level for non-linear form of $x$ from $\chi^2(4)$

| Settings | $H - L$ |  |  |  |  |  |  |  |  |
|----------|---------|---|---|---|---|---|---|---|
|          | 100     | 500 | 1000 | 100 | 500 | 1000 | 100 | 500 | 1000 |
| $log(x)$ | 10.4    | 51.0 | 84.6 | 15.0 | 28.6 | 51.6 | 15.0 | 27.8 | 43.0 |
| $\sqrt{x}$ | 7.2    | 11.6 | 23.8 | 10.2 | 26.2 | 44.2 | 7.2 | 13.8 | 23.2 |
| $x^{-1}$  | 13.8    | 69.6 | 98.0 | 17.2 | 32.8 | 49.8 | 16.0 | 30.8 | 47.2 |
| $x^{-1/2}$ | 18.8   | 88.0 | 100.0 | 25.2 | 75.8 | 93.8 | 19.8 | 59.6 | 79.6 |

Table 5.8: Simulated percent rejection at the $\alpha = 0.05$ level for non-linear form of $x$ from $U(1,20)$

| Settings | $H - L$ |  |  |  |  |  |  |  |  |
|----------|---------|---|---|---|---|---|---|---|
|          | 100     | 500 | 1000 | 100 | 500 | 1000 | 100 | 500 | 1000 |
| $log(x)$ | 6.2     | 11.8 | 22.8 | 7.8 | 25.4 | 46.6 | 7.8 | 25.6 | 46.6 |
| $\sqrt{x}$ | 6.2     | 9.6  | 12.0 | 6.4 | 12.4 | 15.0 | 6.6 | 11.8 | 14.6 |
| $x^{-1}$  | 15.2    | 69.2 | 97.4 | 15.8 | 28.2 | 35.4 | 15.4 | 28.4 | 35.6 |
| $x^{-1/2}$ | 11.0    | 27.6 | 59.0 | 11.8 | 44.4 | 68.0 | 10.8 | 43.0 | 66.0 |

Table 5.8 presents the power of these G.O.F tests when $x$ is from $U(1,20)$ distribution. In this case, all these G.O.F tests also have high power to detect a form of $x^{-1}$ and $x^{-1/2}$, and low power to detect a form of $x^{1/2}$. However, this time, the power of the $H - L$ test and $W$ test decreased significantly for “log” form of $x$. Additionally, we note that the power of the two tests decreased as well for $x^{-1/2}$.

Integrating all the findings in quadratic form and other non-linear forms of $x$, we think that the ability of the different G.O.F tests are affected by both the “true” form of the covariates and also the distribution of the data. All tests show low power in most of the cases. Finally, we think the cumulative sums of residual test, $W$, seems to be the best performing overall. The unweighted sums of squares test, $S$ ranks second. Note that in Chapter 3 we mentioned another cumulative sums of residual test, $W_x$, which is proposed for detecting the non-linear form of a particular covariate. Here, we also examine the performance of $W_x$ under the same situations. From Table 5.9, we can see the $W_x$ test has a consistent
Table 5.9: Simulated percent rejection at the $\alpha = 0.05$ level for the $W_x$ test

<table>
<thead>
<tr>
<th>Settings</th>
<th>$\chi^2(4)$</th>
<th></th>
<th>$U(1,20)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>$\log(x)$</td>
<td>16.8</td>
<td>66.6</td>
<td>93.6</td>
<td>7.3</td>
</tr>
<tr>
<td>$\sqrt{x}$</td>
<td>7.4</td>
<td>17.8</td>
<td>35.8</td>
<td>7.4</td>
</tr>
<tr>
<td>$x^{-1}$</td>
<td>21.4</td>
<td>81.8</td>
<td>99.4</td>
<td>16.2</td>
</tr>
<tr>
<td>$x^{-1/2}$</td>
<td>29.2</td>
<td>95.8</td>
<td>100.0</td>
<td>11.4</td>
</tr>
</tbody>
</table>

Table 5.10: Settings for an omission of interaction term

<table>
<thead>
<tr>
<th>Settings</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\beta_0 = -1.5, \beta_1 = 1, \beta_2 = 1, \beta_3 = 0.22$</td>
</tr>
<tr>
<td>2</td>
<td>$\beta_0 = -1.5, \beta_1 = 1, \beta_2 = 2.25, \beta_3 = 0.5$</td>
</tr>
<tr>
<td>3</td>
<td>$\beta_0 = -1.5, \beta_1 = 1, \beta_2 = 4.05, \beta_3 = 0.9$</td>
</tr>
<tr>
<td>4</td>
<td>$\beta_0 = 0, \beta_1 = 1, \beta_2 = 1, \beta_3 = 0.1$</td>
</tr>
</tbody>
</table>

performance with the $W$ test in the examined cases.

5.3 Omitted Interactions

In this section, we examine the power of the G.O.F tests to detect the omission of an interaction term. The outcome variable is generated from a logistic model with logit link $\logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. The distribution of $x_1$ is chosen as $N(0,1.5)$. Two attributes of $x_2$ will be examined: 1) a categorical variable generated from $Bernoulli(0.5)$; 2) a continuous variable generated from $U(-3,3)$. Table 5.10 presents the settings for this examination. Setting 1, 2, and 3 are situations containing interaction between a continuous variable ($x_1$) and a categorical variable $x_2$ from $Bernoulli(0.5)$. The coefficients of the three settings are chosen such that the models display progressively more interaction. That is, Setting 2 displays more interaction than Setting 1, and Setting 3 displays more interaction than Setting 2. Figure 5.4 shows how the three settings display progressively more interaction. The last setting, Setting 4 is the situation for interaction between $x_1$ and continuous $x_2$.

Table 5.11 shows the power, the percent of time each of these tests rejects the hypothesis
of fit at the $\alpha = 0.05$ level. On the whole, all tests have poor power to detect the omission of an interaction term. Even under a “severe” situation (Setting 3) with a sample size of 1000, the highest power from the cumulative sums of residuals test, $W$, is only 52%. The poor performance is consistent with our finding in the empirical studies. The omitted interaction term is a “difficult” task for all these G.O.F tests. Additionally, the scaled chi-square test yields the overall lowest power.

In section “Other G.O.F Tests” Chapter 2, we introduced two additional G.O.F tests which are proposed by Pulksteins et al. [14]. These two tests, referred to “Continuous Test”, are expected to have good power to detect the omitted interaction between a continuous variable and a categorical variable. Here, we also examined the two tests’ performance under
Table 5.11: Simulated percent rejection at the $\alpha = 0.05$ level for an omission of interaction term I

<table>
<thead>
<tr>
<th>Settings</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>6.2</td>
<td>5.2</td>
<td>8.2</td>
<td>4.8</td>
</tr>
<tr>
<td>2</td>
<td>6.6</td>
<td>7.8</td>
<td>21.0</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>8.8</td>
<td>25.2</td>
<td>44.0</td>
<td>7.2</td>
</tr>
<tr>
<td>4</td>
<td>4.8</td>
<td>6.4</td>
<td>11.0</td>
<td>3.4</td>
</tr>
</tbody>
</table>

Table 5.12: Simulated percent rejection at the $\alpha = 0.05$ level for an omission of interaction term II

<table>
<thead>
<tr>
<th>Test</th>
<th>Setting 1</th>
<th>Setting 2</th>
<th>Setting 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100 500 1000</td>
<td>100 500 1000</td>
<td>100 500 1000</td>
</tr>
<tr>
<td>Continuous Test</td>
<td>10.2 13.8 20.8</td>
<td>15.4 36.2 60.8</td>
<td>14.4 61.2 87.4</td>
</tr>
<tr>
<td>LR Test</td>
<td>8.2 15.4 27.6</td>
<td>16.6 50.4 78.8</td>
<td>25.8 86.6 98.6</td>
</tr>
</tbody>
</table>

the same settings as above. Because the two proposed tests yield almost identical results, so we just state the results from the modified Deviance test here. Table 5.12 presents the results from the “Continuous Test”, as well as the results from the likelihood ratio test (LR Test). From Table 5.12, we can see that the two Continuous Tests have much higher power than other four G.O.F tests, but a slightly lower power than the likelihood ratio test. This finding is to be expected since the two tests are proposed to detect the interaction term between a continuous variable and a categorical variable. In summary, the likelihood ratio test has the highest power and the power of LR test under the continuous situation (Setting 4) is 12.4% ($n = 100$), 37.8% ($n = 500$) and 62.8% ($n = 1000$).

5.4 Link Function

As the last simulation scenario, we examine the performance of these G.O.F tests when the link function is misspecified. Binary outcomes may be modeled with three link functions, the logit link, the complementary log-log link and the probit link. Therefore, in the first part of this section, we examine whether the G.O.F tests can distinguish the departure from
Table 5.13: Simulated percent rejection at the $\alpha = 0.05$ level for Misspecified Link Functions

<table>
<thead>
<tr>
<th>Settings</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>log(-log)</td>
<td>3.8</td>
<td>14.8</td>
<td>41.2</td>
<td>3.6</td>
</tr>
<tr>
<td>Probit</td>
<td>7.4</td>
<td>5.2</td>
<td>6.6</td>
<td>3.0</td>
</tr>
</tbody>
</table>

the logistic model when the other two link functions are correct. In the second part of this section, we examine the performance of the goodness-of-fit tests under some special situations in which the logit is an asymmetric distribution, or has a long-tail or a short tail distribution. The outcome variables are simulated based on Stukel’s generalized logistic regression model, since the generalized model can easily represent symmetric or asymmetric distribution by choosing the values of two tail parameters $\alpha_1$, $\alpha_2$. In this way, we first generate $\eta$, the linear predictor in Stukel’s model, by $\eta = 0.8x_1$, where $x_1$ is from $N(0, 1.5)$. Then we choose different sets of $(\alpha_1, \alpha_2)$ to represent various settings.

5.4.1 Complementary Log-Log and Probit Link

First, we examine the performance of the G.O.F tests when the true link function is a complementary log-log function or a probit function. The values of $(\alpha_1, \alpha_2)$ of these two link functions are $(0.62, -0.037)$ and $(0.165, 0.165)$, respectively [13]. Table 5.13 presents the power, the percent of time each of the tests rejects the hypothesis of fit at the $\alpha = 0.05$ level. All G.O.F tests have very low power (power $\leq 8\%$) to identify lack of fit when the true link is the probit link function. This is not surprising since the probit link function is very close to the logit link function in many situations (See Figure 5.5). The power is highly increased with a sample size of 500 or up when the link function is a complementary log-log link function. The complementary log-log link function is not a symmetric function and has obvious departure from the logit link function in the right tail (See Figure 5.5). However, the increased power is still below 50% with a sample size of 1000. Among all these tests, the cumulative sums of residuals test, $W$, results in the highest power in all settings.
5.4.2 Other Special Situations for the Link Function

As noted in Section 3.6.1, Stukel [13] introduced a generalized logistic regression model that is indexed by two shape parameters $\alpha_1$ and $\alpha_2$. 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. In this section, we examine the performance under four additional cases (Table 5.14) based on Stukel’s formulation. In three of these cases, the chosen link functions have a symmetric distribution, and one has an asymmetric distribution. Based on Stukel’s model, the link function with symmetric distribution has the same values of $(\alpha_1, \alpha_2)$ (Setting 1, 2 and 3), while the one with an asymmetric distribution has different values of $\alpha_1$ and $\alpha_2$ (Setting 4). The absolute value of the $\alpha$’s determine the degree that the specified link function deviates from the “logit” link. For instance, Setting 2 and 3 have
larger deviation from the “logit” link function compared to Setting 1. And Setting 4 has the largest deviation since the simulated distribution is asymmetric. Figure 5.6 presents the degree of deviation from the standard logistic model for each of the specified settings.

Table 5.12 presents the power, the percentage of time each of these tests rejects the hypothesis of fit at the $\alpha = 0.05$ level. The power under Setting 1 is very low for any of the G.O.F tests since this case is still quite close to the linear logit model. As the deviation increases, the power of these tests increases slightly in Setting 2 and 3. The power of all tests is very high with a sample size of 500 or up under Setting 4. This is because the link function in Setting 4 not only has an asymmetric distribution, but also deviates from the linear logit model in both tails. We also note that the unweighted sum of square test, $S$ has high power in Setting 2, which exceeds our expectation.
Table 5.14: Types of Misspecified Link Functions

<table>
<thead>
<tr>
<th>Settings</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\alpha_1 = 0.25, \alpha_2 = 0.25$</td>
</tr>
<tr>
<td>2</td>
<td>$\alpha_1 = 1, \alpha_1 = 1$</td>
</tr>
<tr>
<td>3</td>
<td>$\alpha_1 = -1, \alpha_1 = -1$</td>
</tr>
<tr>
<td>4</td>
<td>$\alpha_1 = -1.5, \alpha_1 = 1$</td>
</tr>
</tbody>
</table>

Table 5.15: Simulated percent rejection at the $\alpha = 0.05$ level for Misspecified Link Functions II

<table>
<thead>
<tr>
<th>Settings</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
<th>$W$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>500</td>
<td>1000 100 500 1000</td>
<td>100 500 1000</td>
</tr>
<tr>
<td>1</td>
<td>7.6</td>
<td>5.8</td>
<td>8.2   4.2 8.6 16.2</td>
<td>3.8 3.2 5.6 7.4 7.8 13.6</td>
</tr>
<tr>
<td>2</td>
<td>6.6</td>
<td>14.6</td>
<td>30.8 5.6 45.6 80.7</td>
<td>0.8 9.6 43.4 8.6 25.8 45.2</td>
</tr>
<tr>
<td>3</td>
<td>5.4</td>
<td>6.6</td>
<td>10.6 7.6 18.2 33.6</td>
<td>8.4 17.4 32.8 7.8 14.2 28.2</td>
</tr>
<tr>
<td>4</td>
<td>15.4</td>
<td>85.2</td>
<td>99.6 17.4 86.2 97.2</td>
<td>4.8 71.8 84.4 20.4 87.2 99.6</td>
</tr>
</tbody>
</table>

To detect a misspecified link function is one of the most difficult tasks in model diagnostics. Integrating all the findings, all the G.O.F tests have low power to determine small or modest deviations from the linear logit. Only when the misspecified logit has an asymmetric distribution and deviates greatly from the linear logit in both tails (Setting 4) do any of the tests perform well. Compared to others, the unweighted sum of squares test, $S$ has the best overall performance, while the cumulative sums of residuals test, $W$, ranks second. The “Homser-Lemeshow” test is a little bit worse than these two, but the difference is small. The scaled chi-square test, $\chi^2$, has the worse performance.

5.5 Conclusions

In this chapter we examined the performance of the G.O.F test under different simulation scenarios. First of all, based on different distributions of the data and functional forms of the logit, all of the G.O.F tests appear to have the right size under the null hypothesis, with a sample size of 500 or up. This is consistent with our findings in empirical studies and previous published work. It was less clear that the correct size resulted when the sample
size was small (n=100). But, as noted this should probably not be surprising given that all of the tests rely on asymptotic results.

In examining the power of the tests against specific alternative formulation, the misspecified link function and the omitted interaction appeared to be the most difficult cases for any of the tests. All of the G.O.F tests had low power to detect a misspecified link function or an omitted interaction, unless the misspecified link function is extremely divergent from the linear logit link or the omitted interaction is extreme. In the omitted interaction case, we also examined the performance of the likelihood ratio test (LR test) and two goodness-of-fit tests proposed by Pulksteins et. al. The likelihood ratio test gives us the best performance overall. It will usually pick up the significant interaction. However, this is a test for the significance of interaction rather than a goodness of fit test.

For detecting a misspecified functional form of \( x \), we were unable to draw a definite conclusion about their relative performance in this setting. All of the tests have good power to detect an omitted quadratic term of \( x \). However, their power to detect other forms of \( x \) not only depends on the true form of \( x \) but also the distribution of the data. Overall, the cumulative sums of residuals test, \( W \) appears to give us the best performance, and the unweighted sum of squares test \( S \) appears to be second best.

Considering their performances under the different scenarios, the unweighted sum of squares test \( S \) and the cumulative sums of residual test \( W \) have better performance than the other two. From the computational standpoint, the \( S \) test is much easier and less time consuming to compute than the \( W \) test. In practice, it appears that the commonly suggested practice of assuming that a p-value less than 0.15 is an indication of lack of fit at the initial steps of model diagnostics should be adopted.

5.6 Discussion

In last section, we summarized our findings, and in this section, we will discuss some issues we found in our study. First of all, with a small sample size like 100, these G.O.F tests might not have a right empirical size in some situations. Also, they all lack power at small sample size though this is to be expected. Hence, when the appropriateness of an asymptotic distribution is in doubt, something else such as parametric bootstrap, exact distribution, or permutation tests should be examined. Here, we reexamine the power of the \( H - L \) test, the \( S \) test as well as the \( \chi^2 \) test using parametric bootstrap resampling method. Because
Table 5.16: Simulated percent rejection at the $\alpha = 0.05$ level for Parametric Bootstrap

<table>
<thead>
<tr>
<th>Settings</th>
<th>$H - L$</th>
<th>$S$</th>
<th>Pearson’s $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log(x)$</td>
<td>11.7</td>
<td>12.5</td>
<td>13.7</td>
</tr>
<tr>
<td>$\sqrt{x}$</td>
<td>5.4</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>$x^{-1}$</td>
<td>13.7</td>
<td>14.3</td>
<td>10.1</td>
</tr>
<tr>
<td>$x^{-1/2}$</td>
<td>19.4</td>
<td>21.2</td>
<td>20.3</td>
</tr>
</tbody>
</table>

The p-value of the $W$ test is already computed from sampling method, we don’t examine it again.

The parametric bootstrap simulates samples $i.i.d$ from the estimated parametric model. In our case, we first estimate the probability of positive response $\pi_i$, then sample the data from the fitted $\hat{\pi}_i$. Let $u_i, i = 1, ..., n$ be random variables from $U(0, 1)$ and $s_{obs}$ be the observed test statistic from the original data. The procedure of parametric bootstrap is:

1. Generate a random sample of new binary outcomes, $y^*_i, i = 1, ..., n$, using the fitted values of $\hat{\pi}_i$:
   $$y^*_i = \begin{cases} 
   1 & \text{if } u_i \leq \hat{\pi}_i \\
   0 & \text{otherwise}
   \end{cases}$$

2. Fit the model using the data $(y^*_i, X_i), i = 1, ..., n$.

3. Calculate the test statistics $s^*$.

4. Repeat steps 1-3 for $M$ times.

5. Calculate the p-value as $p = \frac{\sum_{i=1}^{M} I(s^*_i \geq s_{obs})}{M}$.

In our study, the number of replications for the parametric bootstrap is 1000. Table 5.16 presents the power, the percentage of time each of the three goodness-of-fit tests rejects the hypothesis of fit at the $\alpha = 0.05$ level. For cases we’ve examined, with a sample size of 100, the power from parametric bootstrap is very close to the power from the asymptotic distribution, for each of these goodness-of-fit tests.

Secondly, we will briefly talk about “overdispersion” as this is an important concept in the analysis of categorical data. The problem of overdispersion arises when the observations are neither independent nor identically distributed. Overdispersion means that the data...
shows evidence that the variance of the response variable $y_i$ is bigger than its theoretical value $\pi_i(1 - \pi_i)$. Then the “true” variance of the data needs to be adjusted by a “scale parameter” $\sigma$, $Var(y_i) = \sigma^2 \pi_i (1 - \pi_i)$. As a consequence, all overall goodness-of-fit tests will be distorted and corresponding adjustments should be made. So in future, we will examine the performance of the adjusted G.O.F tests under the situation with overdispersion. In practice, the problem of overdispersion may also be confounded with the problem of omitted covariates. So before considering overdispersion, we should make sure we’ve included all the necessary covariates and their interactions. Note here underdispersion is also theoretically possible, but rare in practice.

The last point we’d like to mention is the number of groups in the $H - L$ test. In our study, we choose 10 as the default number of group since in practice people choose 10 in majority times. In Section 3.3 Chapter 3, we also mention that the number of group has certain impact on the performance of the $H - L$ test. Therefore, our comparison would in all likelihood change somewhat if a different number of group has been chosen.
CHAPTER 6

FUTURE WORK, EXTENSION TO THE PROPORTIONAL HAZARDS MODEL

In this thesis we examined the performance of some G.O.F tests in situations involving a binary outcome and assuming that the data came from a cohort study. That is a study with baseline measurements and a specific follow-up period. There are other ways in which data result where the logistic regression and generalized linear model are widely used, including case-control study (matched or unmatched), studies with categorical response, ordinal response and etc. Whatever the sampling structure or data structure is, the central idea of goodness-of-fit test is the same and involves measurement of the distance between prediction and observation. All the G.O.F tests we’ve examined could be either applied or adapted for these situations. There are also some special G.O.F tests proposed for some particular situations. For instance, Qin and Zhang proposed a Kolmogorov-Smirnov type test statistic \cite{18}, and Zhang proposed a chi-square type test statistics for case-control studies \cite{19}. Arbogast and Lin adapted the cumulative sums of residual test, $W$ for stratified case-control studies \cite{12}. Also, Pulkstenis and Robinson extended their work to ordinal response regression models \cite{20}. Brockwell proposed a kind of residual, “Universal residuals”, as the goodness-of-fit test for a very general class of probability models \cite{21}. Therefore, part of our future task will be to examine the performance of the introduced G.O.F tests in these additional settings.

The goodness-of-fit tests studied in this thesis could be adapted for time to event data. Lin, Wei, and Ying \cite{22} proposed a cumulative sums of Martingale-Based residuals test, and May and Hosmer \cite{23} proposed an overall goodness-of-fit test for the Cox proportional hazards model using the grouping idea of the Hosmer-Lemeshow test. A simple adaption that we have initially developed and will develop further in the future is to use the asymptotic
equivalence of stacked logistic model and the proportional hazards model to develop general
goodness of fit tests for the proportional hazards model.

D’Agostino et al. presented the relationship of stacked logistic regression and Cox
regression model in the Framingham Heart Study \[1\]. If the data structure satisfies with
certain conditions, then stacked logistic regression is equivalent to the proportional hazards
model. Based on this result, our future work is to examine the possibility and the feasibility
of such adaptions. We present here our initial work, starting with a short review of Cox
regression model.

6.1 Adaption G.O.F tests for Cox Proportional
Hazards Model

The Cox proportional hazards model is the most popular model used in survival analysis, but
most goodness of fit diagnostics for this model rely on the plots of various kinds of residuals.
A general test for fit of the proportional hazard is, therefore, desirable.

6.1.1 Diagnostics for Cox Proportional Hazards Model

In this section, we introduce the most widely used diagnostic tests for the Cox proportional
hazards model. Model checking for Cox regression is complicated by the fact that easy-to-use
residuals such as discussed for linear regression models are not available. There are several
commonly used diagnostic methods for the Cox proportional hazard model that involve
specific definitions of a residual.

- The Cox-Snell residual is used to assess the overall fit of a Cox model. It is defined as:

\[
r_i = \hat{H}_0(t_i)\exp(\sum_{k=1}^{p} \beta' x_{ik}), \quad i = 1, ..., n.
\]

where $\hat{H}_0(t_i)$ is the estimated integrated hazard function at time $t_i$, the observed
survival time of subject $i$. If the fitted Cox model is correct, $r_i$ follows an exponential
distribution with mean 1. Thus, a plot of the estimated cumulative hazard rate of the
$r_i$’s, versus $r_i$ should be a straight line through the origin with a slope of 1.
• Martingale residuals are useful for determining the functional form of the covariates. It is the difference between the event indicator $\delta_i$ (equal to 1 if the $i^{th}$ subject experienced the event and 0 otherwise) and the Cox-Snell residual:

$$\hat{M}_i = \delta_i - r_i, \ i = 1, \ldots, n.$$ 

To find a functional form of the covariate, $x_k$, we fit a Cox model with linear predictor and compute the martingale residuals, $\hat{M}_i, i = 1, \ldots, n$. These residuals are plotted against the value of $x_k$ and a smoothed-fitted curve in such a plot gives an indication of the functional form of $x_k$.

• Deviance residuals are defined as:

$$D_i = \text{sign}(\hat{M}_i) \sqrt{-2[\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)]}, \ i = 1, \ldots, n.$$ 

Subjects with large positive or negative deviance residuals are poorly predicted by this model. Deviance residuals are also used to identify potential outliers.

• Other possible residuals are Schoenfeld residuals, score residuals and efficient score residuals which are suggested to be useful for checking and testing the proportional hazard assumption, examining leverage points, and identifying outliers.

### 6.2 The Interconnection between the Proportional Hazards Model and Stacked Logistic Regression

D’Agostino et al. [1] presented the relationship of stacked logistic regression to the time-dependent Cox proportional hazards model based on analyses of the Framingham Heart Study. The central idea is that in stacked logistic regression we model the probability of a subject developing the event of interest by time $t + \Delta(t)$ given he/she is event free at time $t$, which is exactly the definition of the instantaneous hazard. In other words, if both of them modele the instantaneous hazard, then what would their relationship be? D’Agostino et. al. showed that under the conditions that the time interval between measurements of covariates are short, the probability of an event within an time interval is small, and the intercept for stacked logistic regression is constant across all time intervals, then the stacked
logistic regression and the time-dependent covariate Cox regression model are asymptotically equivalent.

Note that the sampling design employed in cohort follow-up studies such as the Framingham Heart Study sometimes includes repeated measures on risk factors, but only one measurement is recorded on the outcome variable. The outcome variable is typically the survival time. Let $t_i$ be the $i^{th}$ measurement time of covariates and $[t_{i-1}, t_i)$ be the $i^{th}$ time interval. Follow-up starts at time $t_0$ and ends at $t_k$. Then, the follow-up time is broken down into small time intervals $[t_{i-1}, t_i)$, $i = 1, ..., k$. These time intervals can be treated as mini follow-up studies. Within the $i^{th}$ time interval, the $p \times 1$ covariate vector $X(t_{i-1})' = (x_1(t_{i-1}), ..., x_p(t_{i-1}))$ is constant over the entire time interval. Here, all time intervals have the same length such as 1-year or half year. (The situation that time intervals have unequal lengths could be considered in future study.)

In following subsections, we first introduce the stacked logistic regression, and then the work and results from D’Agostino et al. In the end of this chapter, we use the Framingham Heart Study data as our example to extend D’Agostino’s results into situations without time-dependent covariates.

The stacked logistic regression is built on the $i^{th}$ time interval, $i = 1, ..., k$. Denote $R_i$ as the number of subjects at risk at time $t_{i-1}$. $D_i$ is the number of events within the $i^{th}$ time interval. Let $q_i(X(t_{i-1}))$ be the conditional probability of developing an event by time $t_i$ given that the subject is event free at time $t_{i-1}$.

Mathematically the stacked logistic regression model is written:

$$\text{logit}[q_i(X(t_{i-1}))] = \alpha_i + \gamma_1 x_1(t_{i-1}) + ... + \gamma_p x_p(t_{i-1}),$$  \hspace{1cm} (6.1)

where $\Gamma' = (\gamma_1, ..., \gamma_p)$ is the coefficient vector for $X(t_{i-1})' = (x_1(t_{i-1}), ..., x_p(t_{i-1}))$. $\alpha_i$, the intercept, is a function of the $i^{th}$ time interval. It is often assumed to be constant over time, in which case $\alpha_i$ is replaced with $\alpha$.

The estimation of $\Gamma$ is very simple. First, the original data set is expanded by breaking down each subject’s survival history into a set of discrete time intervals which are treated as distinct observations. After pooling these observations, the next step is to estimate an ordinary logistic regression predicting whether an event did or did not occur in each time interval. The procedure is best explained by way of an example.

Suppose we have 3 subjects in risk set at the study time $t_0$. Risk factors are measured at
time $t_1, t_2, t_3$, and the study ends at time $t_3$. Subject 1 experienced the event in time interval $[t_2, t_3)$. Subject 2 was lost to follow-up in time interval $[t_1, t_2)$. Subject 3 experienced the event in time interval $[t_1, t_2)$.

The disease status is 1, if a subject experienced the event in the $i^{th}$ time interval, and is zero otherwise. Then, following the above steps, the expanded data set is shown in Table 6.2.

We now fit the ordinary logistic regression to the expanded data set, and obtain the estimates of $\Gamma$ in the stacked logistic regression.

The rationale for assuming independent observations in the expanded data set is based on a simple application of the definition of conditional probability. We can easily factor the likelihood function by $P_t$ the conditional probability of an event at time $t$, given that an event has not already occurred. For instance, if $t = 5$, we have

$$Pr(T = 5) = P_5(1 - P_4)(1 - P_3)(1 - P_2)(1 - P_1)$$

Each of the five terms in above equation may be treated as though it came from a distinct, independent observation.

### 6.3 The Relationship of the Two Models

To examine the relationship, the estimated coefficients from the stacked logistic regression and Cox proportional hazards model are compared. Denote $\gamma$ and $\beta$ as the coefficients for stacked logistic regression and Cox proportional hazards model, respectively. D’Agostino et. al. proposed that two kinds of measurements be used for comparison [1]:

Table 6.1: Data set for Stacked Logistic Regression

<table>
<thead>
<tr>
<th>Subject</th>
<th>Time Interval</th>
<th>Disease Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 6.2: Study Design

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>Endpoint</th>
<th>Incidence Rate</th>
<th>Follow-up Time (years)</th>
<th>Major Risk Factor</th>
<th>Non-significant Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2078</td>
<td>CHD</td>
<td>783(37.7)</td>
<td>30</td>
<td>Age,SBP,CSM</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1154</td>
<td>CHD</td>
<td>353(30.6)</td>
<td>18</td>
<td>Age,SBP,CSM</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1154</td>
<td>CHD</td>
<td>353(30.6)</td>
<td>18</td>
<td>Age,SBP,CSM</td>
<td>SCL,MRW,GLI</td>
</tr>
<tr>
<td>4</td>
<td>1333</td>
<td>Lung cancer</td>
<td>66(5.0)</td>
<td>18</td>
<td>Age,CSM</td>
<td>SBP</td>
</tr>
<tr>
<td>5</td>
<td>107</td>
<td>CVD</td>
<td>50(46.7)</td>
<td>18</td>
<td>SBP,CSM</td>
<td>Age</td>
</tr>
<tr>
<td>6</td>
<td>334</td>
<td>Stroke</td>
<td>44(13.2)</td>
<td>18</td>
<td>Age,SBP</td>
<td>CSM</td>
</tr>
</tbody>
</table>

1. Percentage differences:
   \[ R_1 = 100 \frac{\hat{\gamma} - \hat{\beta}}{\hat{\gamma}}. \]

2. A Comparison of the relative risk for the Cox model to the odds ratio of the logistic model:
   \[ R_2 = 100 \frac{\exp(\hat{\gamma}u) - \exp(\hat{\beta}u)}{\exp(\hat{\gamma}u)}, \]
   where \( u \) is the unit of increment.

Numerical examples covering a variety of sample sizes, event rates and number of risk factors were examined by D’Agostino et al. and they concluded that stacked logistic regression and Cox regression model are “asymptotic equivalent” under conditions [1]:

1. the length of time interval is short enough
2. the incident rate of an event within a time interval is low
3. the intercept of stacked logistic regression is constant over all time intervals.

Table 6.3 summarizes the numerical examples considered by D’Agostino et al. Table 6.3 presents the results for major risk factors: Age, Systolic blood pressure (SBP) and Number of cigarettes per day (CSM).

All of the results suggest that agreements in \( R_2 \), which compares relative risks to odds ratios, are much better than the direct comparison of coefficients \( R_1 \). Age has the largest
Table 6.3: Comparison Results

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Maximum R1</th>
<th>Maximum R2</th>
<th>Minimum R1</th>
<th>Minimum R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>32.4</td>
<td>20.3</td>
<td>9.8</td>
<td>4.3</td>
</tr>
<tr>
<td>SBP</td>
<td>5.7</td>
<td>0.9</td>
<td>-0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>CSM</td>
<td>-12.0</td>
<td>-5.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

difference in all examples which may be partly due to its confounding effect with the follow-up time. Overall, D’Agostino et. al. show that stacked logistic regression and the Cox model are asymptotically equivalent under conditions listed above. The theoretical proof of this result is given in Appendix A.

D’Agostino et al. assumed time-dependent covariates, but based on the theoretical proof, time-dependent covariates are not necessary. That is, for time-independent covariates, we still have a close relationship under the above conditions. We can see this through an empirical example.

We use a data set: The Framingham Heart Study (Males) to illustrate results similar to D’Agostino et. al. and that the time-dependent covariates are unnecessary. The data set has 2009 males at the beginning of the study. The outcome variable is the survival time to coronary heart disease (CHD) death or censoring. Three major risk factors (Age, Systolic Blood Pressure (SBP) and Cholesterol (Chol)) are included in the analysis of the 2009 subjects of whom 452 (22.5%) ultimately developed CHD. But unlike the numerical examples in the D’Agostino paper, none of the covariates is time-dependent.

Using the above procedure, a subject’s survival history is broken down into a set of discrete time intervals. To assure that the length of time interval is short enough, we expanded the original data set in 2-year, 1-year, 6-month and 3-month time intervals. Table 6.4 presents the results for the four different time intervals. $R_1$ and $R_2$ here are computed by switching the positions of $\hat{\gamma}$ and $\hat{\beta}$ in above formulas. This is because we plan to see which length of time interval is short enough by treating the estimated coefficients from the Cox proportional hazards model as the reference set.
Table 6.4: Comparison Results

<table>
<thead>
<tr>
<th></th>
<th>2-year</th>
<th>1-year</th>
<th>6-month</th>
<th>3-month</th>
<th>Cox</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coeff.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.0490</td>
<td>0.0490</td>
<td>0.0490</td>
<td>0.0490</td>
<td>0.0665</td>
</tr>
<tr>
<td>Chol</td>
<td>0.0061</td>
<td>0.0061</td>
<td>0.0060</td>
<td>0.0061</td>
<td>0.0058</td>
</tr>
<tr>
<td>SBP</td>
<td>0.0170</td>
<td>0.0170</td>
<td>0.0170</td>
<td>0.0170</td>
<td>0.0200</td>
</tr>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R1</td>
<td>R2</td>
<td>R1</td>
</tr>
<tr>
<td>Age</td>
<td>26.32</td>
<td>1.74</td>
<td>26.32</td>
<td>1.74</td>
<td>26.32</td>
</tr>
<tr>
<td>Chol</td>
<td>-5.17</td>
<td>-0.03</td>
<td>-5.17</td>
<td>-0.03</td>
<td>-3.45</td>
</tr>
<tr>
<td>SBP</td>
<td>15.00</td>
<td>3.00</td>
<td>15.00</td>
<td>3.00</td>
<td>15.00</td>
</tr>
</tbody>
</table>

We obtain the same results as the ones in D’Agostino paper: agreements in $R^2$ are much better than the direct comparison of coefficients $R_1$; Age has the largest difference in all examples. So for this data set, the stacked logistic regression is close to the Cox proportional hazards model under these four different lengths of time interval.

If the conditions suggested by D’Agostino et al. are satisfied, we expect a close relationship between the results of the stacked logistic regression and the Cox proportional hazards model. Additionally, in some situations such as when there are many ties in the data or the survival time is discrete, stacked logistic regression is more computationally efficient than the Cox proportional hazards model.

### 6.4 Numerical Examples

We again use the Framingham Heart Study data with 2009 males described above as our example. The outcome variable is the survival time to coronary heart death (CHD) or censoring, and the covariates included are age of the subject (Age), Systolic Blood Pressure (SBP) and Cholesterol (Chol). The Cox proportional hazards model as estimated in STATA is $h(t) = h_0(t)exp(0.067 \times Age + 0.0058 \times Chol + 0.02 \times SBP)$. The Schoenfeld residual test suggests that the proportional hazards assumption is satisfied. Figure 6.1 presents The graphical result of Cox-Snell overall test given by STATA.

The Cox-Snell test suggests that the fitted Cox proportional hazards model is correct. Table 6.5 presents the results of the “Hosmer-Lemeshow” test using stacked logistic regression, estimated in STATA and using 10 groups for the “Hosmer-Lemeshow” test.
Figure 6.1: Cox-Snell Overall Test

Table 6.5: “Hosmer-Lemeshow” Test Results in Stacked Logistic Regression

<table>
<thead>
<tr>
<th>Interval</th>
<th>Test Statistics</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-year</td>
<td>9.53</td>
<td>0.2999</td>
</tr>
<tr>
<td>1-year</td>
<td>9.75</td>
<td>0.2831</td>
</tr>
<tr>
<td>6-months</td>
<td>9.70</td>
<td>0.2870</td>
</tr>
<tr>
<td>3-months</td>
<td>9.50</td>
<td>0.3022</td>
</tr>
</tbody>
</table>
The “Hosmer-Lemeshow” test draws the same inference as the Cox-Snell overall test suggesting that the fitted Cox proportional hazard model is a correct model. For other tests, because the sample size of the stacked data exceeds the computational limits in R, we could only compute the results of the other G.O.F tests with the stacked data based on 2-year time intervals (note here 2-year time interval is short enough for this case from Table 6.4). The computed p-value for these tests were: 0.26 for the $W$ test, 0.53 for the unweighted sums of squares test $S$, and 0.58 for the Scaled $\chi^2$ test. All of these G.O.F tests suggest the fitted Cox regression model is correct.

One advantage of the goodness-of-fit test using stacked logistic regression is that it gives us a single, easily interpretable test result, while the graphical way requires subjective visual interpretation. Additionally, under some special situations such as time-dependent variables, the traditional diagnostic tests for Cox regression model are complicated. Stacked logistic regression can be considered an alternative testing methodology for the overall fit in the Cox proportional hazards model.

### 6.5 Future Problems to be Considered

Many problems exist that will require careful consideration. Note that in the Cox proportional hazards model, the usual diagnostics tests focus on particular areas of concern:

1. Whether the proportional hazards assumption is correct versus the alternative that the covariates have a time-dependent effect;

2. The functional form of covariates

3. The link function.

The final two concerns are also considered in goodness-of-fit tests for logistic regression. However, the first concern is a challenging problem since usually logistic regression does not consider the time-dependent effect of covariates. One approach to this concern would be to add a time variable as an independent covariate in the stacked logistic regression. This is similar to the standard method used in the proportional hazards model and was originally suggested by Cox in his original paper. We may also consider that the change of coefficients under different time-interval partitions is another inference about the appropriateness of the proportional hazard assumption.
In our future research, we will expand our initial adaption to examine the possibility and feasibility of these adaptions through more numerical examples and simulation scenarios. Particularly, we will examine the power of the stacked logistic regression to discern cases when the Cox model is incorrect and compare the test results from these adaptions with other proposed G.O.F tests such as the method suggested by May and Hosmer [23].
APPENDIX A

RELATIONSHIP OF STACKED LOGISTIC REGRESSION AND COX PROPORTIONAL HAZARDS MODEL

In this appendix we show that under appropriate assumptions, regression coefficients, likelihood function, and a score test for testing $\beta = 0$ based on a time-interval Cox proportional hazards model are close to those based on stacked logistic regression.

A.1 Regression coefficients

Let $h(t)$ be the instantaneous hazard rate at time $t$ and $S(t)$ denote the probability of being event-free up to time $t$. $X' = (x_1, x_2, ..., x_p)$ is the covariates vector. In Cox model, the hazard rate at time $t$ is a function of covariate $X$ such that

$$ h(t|X) = h_0(t) \exp[\beta'X], $$

where $\beta' = (\beta_1, ..., \beta_p)$ are the coefficients, and $h_0(t)$ is the baseline hazard rate. Assume that observations are grouped into $k$ time intervals and let $I_i = [t_{i-1}, t_i)$ denote the $ith$ interval. $p_i(X)$ is the conditional probability that a subject will survive up to time $t_i$, given that he/she is event-free up to time $t_{i-1}$. Then,

$$ p_i(X) = Pr(t > t_i|t \geq t_{i-1}) = \frac{S(t_i|X)}{S(t_{i-1}|X)} = \exp\left\{ - \int_{t_{i-1}}^{t_i} h_0(u) \exp[\beta'X] du \right\} $$

Now denote $H_i(X) = \int_{t_{i-1}}^{t_i} h_0(u) \exp[\beta'X] du$. Assume the event rate is low and the length of time interval is short enough, then $H_i(X)$ is small for $t$ varying in the $ith$ time interval. By Taylor expansion, we have
\[ p_i(X) = \exp(-H_i(X)) \]
\[ = 1 - H_i(X) + \frac{H_i(X)^2}{2!} - \frac{H_i(X)^3}{3!} + \frac{H_i(X)^4}{4!} \cdots \]
\[ = 1 - H_i(X) + o(H_i(X)) \approx 1 - H_i(X). \]

And
\[ H_i(X) = \int_{t_{i-1}}^{t_i} h_0(u) \exp[\beta'X] du = \exp\{[\log \int_{t_{i-1}}^{t_i} h_0(u) du] + \beta'X\}. \]

Now, let \( q_i(X) = 1 - p_i(X) \) be the conditional probability of developing an event by time \( t_i \), given the subject is event-free up to time \( t_{i-1} \). Then, the stacked logistic model yields
\[
\text{logit}(q_i(X)) = \alpha_i + \Gamma'X,
\]
where \( \Gamma' = (\gamma_1, \ldots, \gamma_p) \) are the coefficients for \( X \) and \( \alpha_i \) is the effect due to the \( ith \) time interval. Denote \( G_i(X) = \exp[\alpha_i + \Gamma'X] \). Again, \( G_i(X) \) is small by assuming that the event rate is low and the time interval is short enough. Consequently, by Taylor expansion,
\[
p_i(X) = 1 - q_i(X)
= \frac{1}{1 + \exp[\alpha_i + \Gamma'X]}
= \frac{1}{1 + G_i(X)}
= 1 - G_i(X) + G_i(X)^2 - G_i(X)^3 \cdots
= 1 - G_i(X) + o(G_i(X))
\approx 1 - G_i(X).
\]

Therefore, in stacked logistic model, \( p_i(X) \) can be approximated by \( 1 - G_i(X) \), while in Cox regression model, \( p_i(X) \) can be approximated by \( 1 - H_i(X) \). Note that both functions \( H_i(X) \) and \( G_i(X) \) are log-linear in terms of \( X \); hence the two corresponding little \( o \) functions, namely \( o(H_i(X)) \) and \( o(G_i(X)) \), are of same order. Then, we have \( 1 - H_i(X) \approx 1 - G_i(X) \). Comparing the component coefficients of the covariate vector \( X \) in \( H_i(X) \) and \( G_i(X) \), we can conclude that \( \gamma_j \) and \( \beta_j \) are approximately equal for each \( j = 1, \ldots, p \). Thus, \( \Gamma = \beta \). Similarly, \( \alpha_i \approx \log \int_{t_{i-1}}^{t_i} h_0(u) du \).
A.2 Likelihood functions

Assume that censoring takes place at the end of each interval only, and assume that there are no tied observations. Let \( R_i \) denote the risk set at time \( t_i - 1 \), \( D_i \) denote the set of subjects in whom events are observed in interval \([t_{i-1}, t_i)\), and \( C_i \) the set of subjects censored in \( i \)th time interval. Applying approximation results derived in previous section, it can be shown that when time interval lengths approach zero and the probability of an event occurring in a short time interval is small, the likelihood function for a time-interval Cox regression model is approximately equal to that of a stacked logistic regression model.

Using the notation given above, the likelihood function of a time-interval Cox regression model is given by

\[
\prod_{i=1}^{k} \left\{ \prod_{l \in D_i} \left( 1 - \exp(-H_i(X_l)) \right) \prod_{m \in R_i - D_i} \exp(-H_i(X_m)) \right\}
\]

\[
\approx \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} H_i(X_l) \prod_{m \in R_i - D_i} (1 - H_i(X_m)) \right\}
\]

\[
\approx \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} G_i(X_l) \prod_{m \in R_i - D_i} (1 - G_i(X_m)) \right\}
\]

\[
= \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} \frac{G_i(X_l)}{1 - G_i(X_l)} \prod_{m \in R_i} (1 - G_i(X_m)) \right\}
\]

\[
\approx \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} G_i(X_l) \prod_{m \in R_i} \frac{1}{1 + G_i(X_m)} \right\}
\]

\[
= \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} \exp[\alpha_l + \beta'X_l] \prod_{m \in R_i} \frac{1}{\exp[\alpha_l + \beta'X_m]} \right\},
\]

which is the likelihood function for a stacked logistic regression model.

For cases with tied observations or with censoring taking place other than at the end of each time interval, modifications can be made as by Thompson(1977) [24].

A.3 Score test statistics for testing \( \beta = 0 \)

By above equations, we can write
\[ p_i(X) = \{exp(- \int_{t_{i-1}}^{t_i} h_0(u)du)\}^{e^{\beta'X}} \]
\[ = p_i(0)^{e^{\beta'X}}, \]

where \( p_i(0) = exp(- \int_{t_{i-1}}^{t_i} h_0(u)du) \). Values of \( p_i(0) \) lie between 0 and 1; therefore a reparameterization is considered to remove this range restriction. Define new parameters \( \theta_i = log(-logp_i(0)), i = 1, ..., k \). Then, the log-likelihood function of a time-interval Cox regression model can be written as

\[
logL(\beta, \theta) = \sum_{i=1}^{k} \{ \sum_{l \in D_i} \log[exp\{exp(\theta_i + \beta'X_l)\}] - 1 \} - \sum_{l \in R_i} exp(\theta_i + \beta'X_l) \}
\]

Hence one can follow Prentice and Gloeckler’s [25] procedure to derive a partial score test [1]. It can be shown that the elements involved in computing the partial score test do not differ greatly from those for the stacked logistic regression model when time interval lengths are short and the event considered is a rare event [1]. The derivation is straightforward but tedious and hence is omitted.
In this appendix, we show that $G_n$ and $\tilde{G}_n$ in Chapter 2 have the same asymptotic distribution.

First recall the definitions and equations in Chapter 2, we have: $X = (1, x_1, ..., x_p)'$ is the $(p + 1) \times 1$ covariate vector; $\mu(\theta)$ is the mean of $y$. $h$ is the link function such that $h(\mu(\theta)) = \beta'X$.

Under some mild regularity conditions, the asymptotic properties of the maximum likelihood estimator $\hat{\beta}$ for $\beta$ were carefully studied by Fahrmeir and Kaufmann(1985) [26]. In particular, if $X$ has a compact support and the matrix $E(XX')$ is positive definite, then with the nature link function $h$ (i.e. $h \circ \mu$ is the identity function) $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically equivalent to $\mathcal{I}^{-1}(\beta)U(\beta)$, where

$$U(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u(\beta'X_i)X_i(y_i - v(\beta'X_i))$$

and

$$\mathcal{I}(\beta) = -\sqrt{n}\partial U(\beta)/\partial \beta',$$

where $v(r) = h^{-1}(r)$, and $u(r) = \partial\{h \circ \mu(r)\}^{-1}/\partial r$ [2]. Note that the definitions of $U(.)$ and $\mathcal{I}(.)$ differ with the ones in Chapter 2 by multiplied a function of $n$. This is only for the convenience of the following asymptotic proof. All the function of $n$ can be crossed over by each other in the equation of $W_n(t)$ and $\tilde{W}_n(t)$.

Define the residuals as $e_i = y_i - v(\hat{\beta}'X_i)$. Then $G_n = \sup_{t \in 1 \times R^p} |W_n(t)|$, where $W_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i I(X_i \leq t), t = (1, t_1, ..., t_p)' \in 1 \times R^p$. $\tilde{G}_n = \sup_{t \in 1 \times R^p} |\tilde{W}_n(t; \hat{\beta})|$, where $\tilde{W}_n(t; \hat{\beta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i e_i \{ I(X_i \leq t) + \eta(t; \hat{\beta}) \mathcal{I}^{-1}(\hat{\beta}) X_i u(\hat{\beta}'X_i) \}$. And $\eta(t; \beta) = -\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t) \partial v(\beta'X_i)/\partial \beta$, also multiplied by a function of $n$.  

Now, to show $G_n = \sup_{t \in [1 \times R^p]} |W_n(t)|$ converges weakly to $\tilde{G}_n = \sup_{t \in [1 \times R^p]} |\tilde{W}_n(t; \hat{\beta})|$, we first give a large sample approximation to the null distribution of $G_n = \sup_{t \in [1 \times R^p]} |W_n(t)|$. If $\eta''(.)$ is bounded, under $H_0$, $W_n(t)$ is asymptotically equivalent to $V_n(t) + \eta'(t; \hat{\beta})\sqrt{n}(\hat{\beta} - \beta)$, where

$$V_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - v(\beta'X_i))I(X_i \leq t).$$

Approximating $\sqrt{n}(\hat{\beta} - \beta)$ by $J^{-1}(\beta)U(\beta)$,

$$W_n(t) \approx V_n(t) + \eta'(t; \hat{\beta})J^{-1}(\beta)U(\beta).$$

Secondly, we show for a large $n$, the distribution of $W_n(t)$ can be approximated by that of $\tilde{W}_n(t; \hat{\beta})$. It follows that $G_n(t)$ has the same asymptotic distribution as that of $\tilde{G}_n(t)$.

Without loss of generality, let all the components of $X$ be between 0 and 1. Then, we split the covariate vector $X$ into two parts $X_1$ and $X_2$, where $X_1$ consists of all the discrete covariates and $X_2$, a $q \times 1$ vector, consists of all the continuous components of $X$. Let $t' = (t_1', t_2')$. Then, $V_n(t_1', t_2) = \sum_{k \leq t_1} V_n(k, t_2)$, where the $\sum$ denotes the summation over all possible outcomes $k$, generated from $X_1$, that are less than $t_1$, and

$$V_n(k, t_2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - v(\beta'X_i))I(X_{1i} = k, X_{2i} \leq t_2).$$

Note that these $V_n(k, t_2)$'s are independent of each other and are in the space $D[0, 1]^q$. It follows that the process $W_n(t) = W_n(t_1, t_2)$ can be approximated by $\sum_{k \leq t_1} V_n(k, t_2) + \eta'(t; \beta)J^{-1}(\beta)U(\beta)$.

Let $\ell$ be the collection of all possible outcomes generated from $X_1$. Also, let $W(.) = \{W_0, V(k, .); k \in \ell\}$ be a multidimensional Gaussian process, where $V(k, .)$'s are Gaussian processes and are independent of each other, the covariance function of $V(k, .)$ for $t_2^{(1)}$ and $t_2^{(2)}$ is $E\{(y - v(\beta'X))^2I(X_1 = k, X_2 \leq (t_2^{(1)} \wedge t_2^{(2)}))\}$, $W_0$ is a normal random vector with covariance matrix $E\{XX'u^2(\beta'X)(y - v(\beta'X))^2\}$, and the covariance between $V(k, t_2)$ and $W_0$ is $E\{Xu(\beta'X)(y - v(\beta'X))^2I(X_1 = k, X_2 \leq t_2)\}$. Now, for any finite collection of points $t_2^{(1)}, \ldots, t_2^{(m)}$ in $[0, 1]^q$, using Cramer-Wald Theorem, the proof of the convergence for $\{U(\beta), V_n(k, t_2^{(j)}); j = 1, \ldots, m, k \in \ell\}$ to the normal random vector $\{W_0, V(k, t_2^{(j)}); j = 1, \ldots, m, k \in \ell\}$ is straightforward.
And then, this implies that the multidimensional process \( \{U(\beta), V_n(k, t_2), k \in \ell \} \) converges weakly to the normal process \( \{W_0, V_n(k, t_2), k \in \ell \} \). Furthermore, if \( v'(.) \) is bounded, then for each \( k \), \( \sup_{t_2} |\hat{v}(k, t_2; \beta) - \eta(k, t_2; \beta)| \to 0 \), as \( n \to \infty \), where 
\[
\eta(k, t_2; \beta) = -E[v'(\beta'X)XI(X_1 = k, X_2 \leq t_2)].
\]
Therefore, the process \( \{V_n(k, t_2) + \eta'(k, t_2; \beta)\mathcal{J}^{-1}(\beta)U(\beta); k \in \ell \} \) converges weakly to \( \{V(k, t_2) + \eta'(k, t_2; \beta)\mathcal{J}^{-1}(\beta)W_0; k \in \ell \} \).

Since this limiting process is in \( C[0, 1]^q \), \( \|W_n(t)\| \) converges weakly to \( \sup_{t \in R^q} |\sum_{k \leq t_1} \{V(k, t_2) + \eta'(k, t_2; \beta)\mathcal{J}^{-1}(\beta)W_0\}|. \)

Now, define
\[
\tilde{V}_n(k, t_2; \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i(y_i - v(\beta'X_i))I(X_{1i} = k, X_{2i} \leq t_2),
\]
and
\[
\tilde{U}_n(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i((y_i - v(\beta'X_i))u(\beta'X_i)X_i,
\]
where the \( Z \)'s are independent of \( \{y_i, X_i; i = 1, ..., n\} \). Then, using the above argument, we can show that
\[
\|\tilde{W}_n(t_1, t_2; \beta)\| = \|\sum_{k \leq t_1} \tilde{V}_n(k, t_2; \beta) + \eta'(k, t_2; \beta)\mathcal{J}^{-1}(\beta)\tilde{U}_n(\beta)\|
\]
converges weakly to the random variable. Furthermore, assume that \( v''(.) \) and \( u''(.) \) are continuous and bounded. Since \( \hat{\beta} \to \beta \) almost surely as \( n \to \infty \), and \( Z \)'s are independent of \( \{y_i, X_i; i = 1, ..., n\} \), it follows that \( \tilde{V}_n(k, t_2; \hat{\beta}) - \tilde{V}_n(k, t_2; \beta) \to 0 \), \( \eta(k, t_2; \hat{\beta}) - \eta(k, t_2; \beta) \to 0 \), \( \mathcal{J}^{-1}(\hat{\beta}) - \mathcal{J}^{-1}(\beta) \to 0 \), and \( \tilde{U}_n(\hat{\beta}) - \tilde{U}_n(\beta) \to 0 \), in distribution, as \( n \to \infty \). Hence, \( G_n = \|W_n(t)\| \) and \( \tilde{G}_n = \|\tilde{W}_n(t; \hat{\beta})\| \) have the same asymptotic distribution.
APPENDIX C
FRATIONAL POLYNOMIAL METHOD

Royston and Altam [27] and Royston et al. [28] proposed fractional polynomials as a means to investigate the functional forms of continuous covariates. In this appendix, we will briefly introduce the method of fractional polynomials.

Let $\mathbf{\beta} = (\beta_0, \beta_1)$ is the vector of coefficients, and $x$ is the single covariate. $g$ is the link function such that

$$g(x, \beta) = \beta_0 + \beta_1 x.$$ 

One way to generalize this function is to specify it as

$$g(x, \beta) = \beta_0 + \sum_{j=1}^{J} F_j(x) \beta_j$$

The function $F_j(x)$'s are some particular types of power function [27]. The first function is $F_1(x) = x^{p_1}$, and the remaining functions are defined as

$$F_j(x) = \begin{cases} x^{p_j} & \text{if } p_j \neq p_{j-1} \\ F_{j-1}(x) \log(x) & \text{if } p_j = p_{j-1} \end{cases}$$

$j = 2, ..., J$. $p_j$ here is the power for function $F_j(x)$. Royston and Altman proposed restricting the powers to be among those in the set $p = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. And $p = 0$ denotes the log transformation [27].

For example, if we chose $J = 2$ with $p_1 = 0$ and $p_2 = -0.5$, then the model with fractional polynomials is

$$g(x, \beta) = \beta_0 + \beta_1 \log(x) + \beta_2 \frac{1}{\sqrt{x}}.$$
The model is quadratic in $x$ when $J = 2$ with $p_1 = 1$ and $p_2 = 2$. And in most applied settings, $J$ is chosen to be 1 or 2 [27]. The method of fractional polynomials can be implemented in STATA.
REFERENCES


[9] Hosmer D and Lemeshow S. *Applied Logistic Regression*(2nd edn.). Wiley: New York, 2000. 3.3, 4.1, 4.2.1, 4.2.1, 4.2.2, 4.2.2, 4.2.3


78


[17] Lemeshow S and Le Gall J. Modeling the severity of illness of icu patients. *JAMA*, 272, 1994. 4.2.2


BIOGRAPHICAL SKETCH

Sutan Wu

Sutan Wu was born in Wuhu, Anhui Province, China in 1983. She received her B.S degree in applied math from the University of Science and Technology of China, in 2005. After that, she came to Statistics Department, FSU to continue her graduate study. She obtained her Master degree in applied statistics in 2007 and her Ph.D degree in 2010 under the advisement of Dr. McGee and Dr. Zhang.

Sutan’s research interests include data exploration, modeling, clinical trial design and design of experiment. In life, Sutan likes animals and all outdoor activities.