The Sum of Standardized Residuals: Goodness of Fit Test for Binary Response Model

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Lu Chen, B.S

Graduate Program in Public Health

The Ohio State University

2017

Thesis Committee:
Grzegorz A. Rempala, Advisor
Chuck Song
Abstract

Binary response model is frequently used to analyze binary outcome variables [1]. This popularity leads to an increase in statistical research on the model [1]. One area of current research is developing new goodness of fit (GOF) tests to evaluate whether a model fits the data [1]. In this paper, a new GOF test statistic, the sum of standardized residuals \( C_n \), is proposed and its asymptotic distribution is described by following Windmeijer’s idea [2]. In addition, we illustrate, via numerical examples, the practical applications of the asymptotic result to some finite samples of size \( n \) and compare \( C_n \)’s performance with some other currently used statistics. Our results demonstrate that, compared to other statistics, the overall performance of \( C_n \) is satisfying and stable, and \( C_n \) can be calculated easily and interpreted intuitively, unlike its other competitors.
Vita

2014 ............................................. B.S. Pharmacy, China Pharmaceutical University
2015 to present ................................. School of Public Health, The Ohio State University

Fields of Study

Major Field: Public Health
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Vita</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>v</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Some Currently Used Goodness-of-Fit Statistics</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Hosmer-Lemeshow Statistic</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Stukel Statistic</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3 Unweighted Sum of Squares Statistic</td>
<td>3</td>
</tr>
<tr>
<td>1.1.4 Information Matrix Statistic</td>
<td>4</td>
</tr>
<tr>
<td>2. Sum of Standardized Residual Statistic</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Regularity Assumptions</td>
<td>5</td>
</tr>
<tr>
<td>2.2 The Asymptotic Distribution of $C_n$</td>
<td>7</td>
</tr>
<tr>
<td>3. Numerical Examples</td>
<td>8</td>
</tr>
<tr>
<td>3.0.1 Simulated Covariates</td>
<td>9</td>
</tr>
<tr>
<td>3.0.2 Real Covariates</td>
<td>11</td>
</tr>
<tr>
<td>3.0.3 Summary</td>
<td>18</td>
</tr>
<tr>
<td>Bibliography</td>
<td>19</td>
</tr>
<tr>
<td>Appendices</td>
<td>20</td>
</tr>
<tr>
<td>A. Proof of the Asymptotic Distribution of $C_n$</td>
<td>20</td>
</tr>
</tbody>
</table>
List of Tables

Table | Page
--- | ---
3.1 Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 50, 100 and 1000 with 1000 replications using simulated covariates | 10
3.2 Selected variables description in the Boston Marathon data | 11
3.3 Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 445 with 1000 replications in quadratic models using the Boston Marathon data | 13
3.4 Selected variables description in Cleveland data | 14
3.5 Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 303 with 1000 replications in interaction models using the Cleveland data | 15
3.6 Selected variables description in NHANES data | 16
3.7 Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 373 with 1000 replications in interaction models using the NHANES data | 17
Chapter 1: Introduction

The binary response model [3] is a popular method for analyzing binary data. In this model, we are interested in \( p_i \), where \( p_i \) is defined as the probability that a binary random variable \( Y_i = 1 \) given the covariates \( x_i \) such that

\[
p_i = P(Y_i = 1|x_i) = F(x_i^T \beta),
\]

where \( \beta \) is the \( k \)-vector of coefficients and \( F \) is a distribution function. Some popular choices of \( F \) include normal (probit model) or logistic (logit model) distributions.

After a model is fitted, we usually want to assess its goodness of fit (GOF). One of the several approaches is to perform a formal GOF test. We will introduce several currently used test statistics in the next section. Also, we will propose a new statistic, the sum of standardized residuals (\( C_n \)), and describe its asymptotic distribution in Chapter 2. In Chapter 3, we will illustrate, via numerical examples, the practical applications of the asymptotic result to some finite samples of size \( n \) and compare \( C_n \)'s performance with other statistics.

1.1 Some Currently Used Goodness-of-Fit Statistics

1.1.1 Hosmer-Lemeshow Statistic

In 1980, Hosmer and Lemeshow [4] proposed a grouping-based method to measure the goodness of fit for the binary response model. Specifically, the data are grouped
according to the $\hat{p}_i$’s, the estimated values of $p_i$. The observed and expected frequency are calculated for each group, and compared using the Pearson chi-square statistic to obtain

$$HL = \sum_{k=1}^{g} \left[ \frac{(O_{1k} - \hat{E}_{ik})^2}{\hat{E}_{1k}} + \frac{(O_{0k} - \hat{E}_{0k})^2}{\hat{E}_{0k}} \right]$$

where $k$ denotes the $k^{th}$ group, $k = 1, 2, 3, ..., g$. $O_{1k} = \sum_{j=1}^{c_k} Y_i$, is the observed number of events ($Y_i = 1$) in the $k^{th}$ group, where $c_k$ is the number of covariate patterns in the $k^{th}$ group. $\hat{E}_{1k} = \sum_{j=1}^{c_k} m_j \hat{p}_j$ is the expected number of events in the $k^{th}$ group, where $m_j$ is the number of observations in $j^{th}$ covariate pattern, and $\hat{p}_j$ is the estimated probability for $j^{th}$ covariate pattern. Similarly, $O_{0k} = \sum_{j=1}^{c_k} (m_j - Y_j)$ is the observed number of non-events ($Y_i = 0$) in the $k^{th}$ group, and $\hat{E}_{0k} = \sum_{j=1}^{c_k} m_j (1 - \hat{p}_j)$ is the expected number of non-events in the $k^{th}$ group. Since $g = 10$ and $g = 20$ are commonly suggested in practice, we consider using these two setting to perform $HL$ tests in Section 3. With a bit of algebra, it can be shown that

$$HL = \sum_{k=1}^{g} \frac{(O_{1k} - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)}$$

where $n_k'$ is the number of observations in the $k^{th}$ group, $\bar{p}_k$ is the average estimated probability in the $k^{th}$ group. When the number of groups $g$ equals to the sample size $n$, the formula of $HL$ can be further rewritten as

$$HL = \sum_{i=1}^{n} \frac{(Y_i - \hat{p}_i)^2}{\hat{p}_i (1 - \hat{p}_i)}$$

which suggests the formula for another goodness of fit measure: the sum of squared standardized residuals.
1.1.2 Stukel Statistic

To perform Stukel GOF test [5], one or two additional covariates are included in the standard logistic model, and this new model is tested versus the original one. For example, let \( \hat{p}_i = F(x_i^\top \hat{\beta}) \) be the original model, where \( \hat{\beta} \) is the vector of estimated coefficients. Let \( z_a = g_i^2 I(g_i \geq 0) \) and \( z_b = g_i^2 I(g_i < 0) \) be the two covariates, where \( g_i = x_i^\top \hat{\beta} \). After including \( z_a \) and \( z_b \) into the model, it becomes \( \hat{p}_i = F(x_i^\top \hat{\beta} + \hat{\beta}_a z_a + \hat{\beta}_b z_b) \). Then, the Ward test or the Likelihood Ratio test can be performed to test the null hypothesis that both \( \beta_a \) and \( \beta_b \) are equal to 0.

1.1.3 Unweighted Sum of Squares Statistic

When some of the expected cell frequencies in the \( 2 \times n \) contingency table are too small, the weighted chi-squared test for proportions cannot be used. In this case, Copas [6] modified the chi-squared test as

\[
USS = \sum_{i=1}^{n} (Y_i - \hat{p}_i)^2,
\]

where \( \hat{p}_i \) is defined as before, \( n \) is the sample size. Under the null hypothesis, Hosmer et al. [1] have proved that this statistic will have an asymptotic normal distribution when standardized by its mean and the square root of its variance:

\[
\frac{USS - \text{trace}(\hat{V})}{\sqrt{\text{Var}(USS - \text{trace}(\hat{V}))}} \xrightarrow{d} N(0, 1)
\]

where

\[
\sqrt{\text{Var}(USS - \text{trace}(\hat{V}))} = \hat{d}_i^\top (I - \hat{M}) \hat{V} \hat{d}_i
\]

with

\[
\hat{d}_i = 1 - 2\hat{p}_i
\]
\[
\hat{V} = diag[\hat{p}_i(1 - \hat{p}_i)]
\]

\[
\hat{M} = \hat{V} x_i (x_i^\top \hat{V} x_i)^{-1} x_i^\top
\]

### 1.1.4 Information Matrix Statistic

In 1982, White [7] introduced a general test for model misspecification. The basic idea of this test is to compare two different estimates of the covariance matrix. If these two estimates are the same, there is no lack of fit. The formula for Information Matrix statistic is:

\[
IM = \sum_{i=1}^{n} \sum_{j=0}^{p} (Y_i - \hat{p}_i)(1 - 2\hat{p}_i)x_{ij}^2
\]

where \( x_{ij} \) is the \( j^{th} \) predictor for \( i^{th} \) observation. It has been shown by White that \( IM \) has a approximate chi-square distribution with \( p + 1 \) degree of freedom under the null hypothesis of the correct model.
Chapter 2: Sum of Standardized Residual Statistic

As we mentioned in Section 1, the sum of squared standardized residuals is also a goodness of fit measure in binary response model

\[ T_n = \sum_{i=1}^{n} \frac{(Y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)} = \sum_{i=1}^{n} W_i. \]

This statistic was proved to be asymptotically following the standard normal distribution when properly standardized [2]. However, the drawback of \( T_n \) statistic is that the variances of its individual components \( W_i \) depend heavily on \( p_i \) in such way that extreme values of \( p_i \) could produce unbalanced variance across the sample. One intuitive way to address this issue is to assign a variance-stabilizing weight to \( W_i \). This yields a new statistic

\[ C_n = \sum_{i=1}^{n} \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}. \]

We will analyze the asymptotic distribution of this statistic under the following assumptions.

2.1 Regularity Assumptions

Following Windmeijer [2] we assume that
1. Both the first-order derivative \( f \) and second-order derivative \( f' \) of \( F \) are continuous. There exists a positive \( M < \infty \) such that \( 0 < F(z) < 1 \) and \( f(z) > 0 \) for all \( |z| \leq M \).

2. \( \beta \in \mathcal{B} \), the parameter space which is a open subset of the Euclidean space of the same dimension \( k \) as \( \beta \).

3. There exists a finite number \( M \) such that \( ||x_i|| \leq M, \forall i \in \mathbb{N} \) and \( \lim_{n \to \infty} \sum_i x_i x_i^\top \) is a finite non-singular matrix. Also, the empirical distribution of \( x_i \) converges to a non-degenerate distribution function.

Under the assumptions 1-3, the following two results about MLE \( \hat{\beta}_n \) are known (see e.g., [8]).

**Result 1 (Asymptotic distribution of MLE)**

\[
n^\frac{1}{2}(\hat{\beta}_n - \beta_0) \overset{d}{\to} N(0, \Omega^{-1}) \tag{2.1}
\]

where \( \hat{\beta}_n \) is the maximum likelihood estimator of \( \beta_0 \) and \( \Omega \) is defined as

\[
\Omega = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n \frac{f^2_{i0}}{p_{i0}(1-p_{i0})} x_i x_i^\top,
\]

where \( f_{i0} = f(x_i^\top \beta_0) \).

By Taylor’s expansion one can also show the following.

**Result 2 (Score representation of MLE)**

\[
\hat{\beta}_n - \beta_0 = - \left( \frac{\partial \log L}{\partial \beta \partial \beta^\top} \bigg|_{\beta_0} \right)^{-1} \frac{\partial \log L}{\partial \beta} \bigg|_{\beta_0} \tag{2.2}
\]

where \( \beta_n^* \) lies on the segment between \( \hat{\beta}_n \) and \( \beta_0 \) and

\[
\log L = \sum_{i=1}^n \left( Y_i \log p_i + (1 - Y_i) \log (1 - p_i) \right)
\]

is the loglikelihood of the binary response model.
2.2 The Asymptotic Distribution of $C_n$

We describe the asymptotic distribution of $C_n$ in the following theorem, which we prove in the appendix.

**Theorem 1** Under the assumptions 1-3, it can be shown that

\[
\frac{n^{-\frac{1}{2}}C_n}{\sigma_n} \xrightarrow{d} N(0, 1),
\]

where

\[
\sigma_n^2 = 1 - \mathbf{v}_n^\top \Omega^{-1} \mathbf{v}_n
\]

with

\[
v_n = n^{-1} \sum_{i=1}^{n} \frac{f_{i0} x_i}{\sqrt{p_{i0}(1 - p_{i0})}}, \tag{2.3}
\]
Chapter 3: Numerical Examples

Based on the theorem, a new GOF test can be defined. The test statistic is:

\[ \tilde{C}_n = \frac{C_n}{\sqrt{n}\sigma_n} \]

which asymptotically follows the standard normal distribution. To show the application of this result and evaluate the power of \( C_n \) along with other GOF statistics, we provide several numerical examples under different scenarios. Although this asymptotic result is generalized to binary response models with different link functions, we consider only the logit link here since it is the most commonly-used one.

Statistics considered here include Hosmer-Lemeshow (HL) with \( g = 10 \) and \( g = 20 \), Stukel, Unweighted Sum of square (USS), Information Matrix (IM) and our new proposed statistic \( C'_n \). The tests are all performed at the \( \alpha \) level of 0.05.
3.0.1 Simulated Covariates

We first consider scenarios similar to those used by Windmeijer [2] in generating $x_i$ and $Y_i$:

$$p_i = \frac{\exp(1 + x_i)}{1 + \exp(1 + x_i)} \quad x_i \sim N\left(-\frac{1}{2}, 1\right) \quad \text{Model 1.0}$$

$$p_i = \frac{\exp(1 + x_i + \ln |x_i|)}{1 + \exp(1 + x_i + \ln |x_i|)} \quad x_i \sim N\left(-\frac{1}{2}, 1\right) \quad \text{Model 1.1}$$

$$p_i = \exp(\lambda (x_i - 3)) \quad \lambda = \frac{\sqrt{3}}{\pi} \quad x_i \sim U(-3, 3) \quad \text{Model 1.2}$$

Model 1.0 is the null model, Model 1.1 and Model 1.2 are two alternative models to assess power for detecting nonlinearity and incorrectly specified link function. For each simulation, 1000 replicates were obtained with sample sizes of 50, 100, and 1000. Table 3.1 shows the proportion of $H_0$ rejected by each of the tests.
Table 3.1: Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 50, 100 and 1000 with 1000 replications using simulated covariates

<table>
<thead>
<tr>
<th>Model</th>
<th>Test statistics</th>
<th>Proportion $H_0$ rejected</th>
<th>Sample size=50</th>
<th>Sample size=100</th>
<th>Sample size=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $H_0$: Model 1.0</td>
<td>$C_n$</td>
<td>0.045</td>
<td>0.051</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.05</td>
<td>0.048</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.027</td>
<td>0.039</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.037</td>
<td>0.054</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.092</td>
<td>0.076</td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>0.044</td>
<td>0.05</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>Under $H_a$: Model 1.1</td>
<td>$C_n$</td>
<td>0.224</td>
<td>0.419</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.253</td>
<td>0.659</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.099</td>
<td>0.435</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.198</td>
<td>0.302</td>
<td>0.656</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.51</td>
<td>0.773</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>0.255</td>
<td>0.562</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Under $H_a$: Model 1.2</td>
<td>$C_n$</td>
<td>0.178</td>
<td>0.618</td>
<td>0.904</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.123</td>
<td>0.437</td>
<td>0.784</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.128</td>
<td>0.361</td>
<td>0.711</td>
<td></td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.144</td>
<td>0.465</td>
<td>0.756</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.197</td>
<td>0.718</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>0.159</td>
<td>0.653</td>
<td>0.944</td>
<td></td>
</tr>
</tbody>
</table>

In Table 3.1, we see that when the fitted model is correct, in all but a few situations, all the tests reject the null hypothesis nearly at the 5% level. *Stukel* statistic has an inflated type one error for samples of size 50 and 100. This result suggests that when sample size is small, the highest power obtained by *Stukel* statistic should be considered as suspect since it is not at the same level as the others. Therefore, we will not consider *Stukel*’s subsequent performances when sample size is 50 and 100.

Table 3.1 also shows that there is an increasing trend in power as the sample size goes up. For a sample of size 1000, we see that when trying to detect the omission of a nature log term, almost all statistics, including $C_n$, attain a 100% power. When trying to detect an incorrect link function, we also see a very good power (over 90%)
for $C_n$, Stukel and IM. With a small sample size, though $C_n$ does not perform the best, it is still, at least, better than some of the currently used statistics.

### 3.0.2 Real Covariates

As the simulated covariates might not be biologically meaningful, we consider also the covariates based on values from a real dataset. In other words, we only generate the outcome variable $Y_i$. The covariates as well as their coefficients are chosen based on the original models. In this case, the power to detect the omission of a quadratic term or an interaction term was evaluated. For each simulation, the number of replication is also 1000.

**Linear vs. quadratic**

Covariates used in this section are derived from an epidemiological study conducted among runners in the 2002 Boston Marathon [9]. The goal of this study is to identify potential risks associated with hyponatremia which is defined as a serum sodium concentration level of 135 mmol/l or lower. Details about the selected covariates and the outcome variable are provided in Table 3.2. After excluding participants with missing values in selected covariates or the outcome variable, we are left with $n = 445$.

**Table 3.2: Selected variables description in the Boston Marathon data**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Codes/Values</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$0 = \leq 135$ mmol/liter, $1 &gt; 135$ mmol/liter</td>
<td>Serum sodium concentration at race completion</td>
</tr>
<tr>
<td>$s$</td>
<td>$0 =$ Female, $1 =$ Male</td>
<td>Sex</td>
</tr>
<tr>
<td>$w$</td>
<td>kg</td>
<td>Weight change during race</td>
</tr>
<tr>
<td>$r$</td>
<td>Hours</td>
<td>Race duration</td>
</tr>
<tr>
<td>$u$</td>
<td>$1 &gt; 3$ liter, $0 \leq 3$ liter</td>
<td>Fluid consumption during the race</td>
</tr>
</tbody>
</table>
Models we used to generate $Y_i$ are:

\[ p_i = \frac{\exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i)}{1 + \exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i)} \quad \text{Model 2.0} \]

\[ p_i = \frac{\exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i + 0.1w_i^2)}{1 + \exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i + 0.1w_i^2)} \quad \text{Model 2.1} \]

\[ p_i = \frac{\exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i + w_i^2)}{1 + \exp(-5 + 0.7s_i + 0.7w_i + 0.013r_i + 0.9u_i + w_i^2)} \quad \text{Model 2.2} \]

Model 2.0 contains no quadratic term, Model 2.1 and Model 2.2 contain quadratic terms with different coefficients which are to emphasize or deemphasize the effect of the quadratic term [10]. Results comparing the $H_0$ rejection proportion among different tests are showed in Table 3.3.
Table 3.3: Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 445 with 1000 replications in quadratic models using the Boston Marathon data

<table>
<thead>
<tr>
<th>Model</th>
<th>Test statistics</th>
<th>Proportion $H_0$ rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $H_0$: Model 2.0</td>
<td>$C_n$</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>0.051</td>
</tr>
<tr>
<td>Under $H_a$: Model 2.1</td>
<td>$C_n$</td>
<td>0.354</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>0.239</td>
</tr>
<tr>
<td>Under $H_a$: Model 2.2</td>
<td>$C_n$</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.751</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.763</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>IM</td>
<td>1</td>
</tr>
</tbody>
</table>

In Table 3.3, we see that under null, nearly all test statistics reject the null at about the $\alpha$ level of 0.05 except Stukel and $HL(g = 10)$. Table 3.3 also shows that under Model 2.1 which is added a quadratic term with a coefficient of 0.1, rejection proportions of all statistics increase. It is worth mentioning that $C_n$ has the highest power here. Under Model 2.2, as the value of the coefficient increases, the power of detecting such departure becomes much greater. Indeed the tests based on $C_n$, Stukel and IM all have the power exceeding 90%.
Linear vs. interaction

Additional set of covariates is used to measure power against superficial interaction and comes from the Cleveland dataset which is a subset of a large Heart Disease dataset [11]. Only $s$ and $t$ are chosen to be included in our simulation based on their significant relationship with the response variable. After excluding observations with missing values in any of the three variables, a total of 303 observations is left in our study. Table 3.4 shows a description about these three variables.

Table 3.4: Selected variables description in Cleveland data

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Codes/Values</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>0 = Absence, 1 = Presence</td>
<td>Heart Disease Diagnosis</td>
</tr>
<tr>
<td>$s$</td>
<td>0 = Female, 1 = Male</td>
<td>Sex</td>
</tr>
<tr>
<td>$t$</td>
<td>mm Hg, recorded at baseline</td>
<td>Resting blood pressure</td>
</tr>
</tbody>
</table>

We used two models to generate $Y_i$:

$$p_i = \frac{\exp(-4 + 1.4s_i + 0.02t_i)}{1 + \exp(-4 + 1.4s_i + 0.02t_i)} \quad \text{Model 3.0}$$

$$p_i = \frac{\exp(-4 + 1.4s_i + 0.02t_i - 0.04s_i \times t_i)}{1 + \exp(-4 + 1.4s_i + 0.02t_i - 0.04s_i \times t_i)} \quad \text{Model 3.1}$$

Model 3.0 is the null model which contains two covariates $s$ and $t$. Model 3.1 is the alternative model which contains not only $s$ and $t$, but also an interaction term between the two covariates. Table 3.5 shows the percent of time $H_0$ was rejected by each of the six tests except Hosmer-Lemeshow with group=20 because of the limited number of covariate patterns in this situation.
Table 3.5: Proportion \( H_0 \) rejected at the \( \alpha = 0.05 \) using sample size of 303 with 1000 replications in interaction models using the Cleveland data

<table>
<thead>
<tr>
<th>Model</th>
<th>Test statistics</th>
<th>Proportion ( H_0 ) rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under ( H_0 ): Model 3.0</td>
<td>( C_n )</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>( HL(g = 10) )</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>( USS )</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>( Stukel )</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>( IM )</td>
<td>0.038</td>
</tr>
<tr>
<td>Under ( H_a ): Model 3.1</td>
<td>( C_n )</td>
<td>0.613</td>
</tr>
<tr>
<td></td>
<td>( HL(g = 10) )</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td>( USS )</td>
<td>0.181</td>
</tr>
<tr>
<td></td>
<td>( Stukel )</td>
<td>0.538</td>
</tr>
<tr>
<td></td>
<td>( IM )</td>
<td>0.498</td>
</tr>
</tbody>
</table>

We see in Table 3.5 that, as expected, the proportions of rejecting the null hypothesis are still close to 0.05 when the fitted model is correct. Also, when an interaction term is added to the true model, rejection proportions are greater than those under null. In this setting, while both \( HL \) with a group of size 10 and \( USS \) attain a poor power, \( C_n \) performs noticeably better than the others with a very good power of 61.3%. However, it is unclear whether the increase in power holds when the true model departs further from the fitted model.

To find out whether further departure yields higher power, another set of covariates is used. It comes from the National Health and Nutrition Survey which contains 17030 observations and 16 variables. Two independent variables, \( m \) and \( c \), are chosen to run simulation. They are significantly associated with the dependent variable \( y \) simultaneously. We randomly select 373 observations with no missing value in \( m \), \( c \) and \( y \). Table 3.6 gives a description of these three variables.
# Table 3.6: Selected variables description in NHANES data

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Codes/Values</th>
<th>Variable description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>0 = PEPMNK1R ≤140, 1 = PEPMNK1R &gt;140</td>
<td>High Blood Pressure</td>
</tr>
<tr>
<td>( m )</td>
<td>0 = Female, 1 = Male</td>
<td>Sex</td>
</tr>
<tr>
<td>( c )</td>
<td>0 = Yes, 1 = No</td>
<td>Serum Cholesterol</td>
</tr>
</tbody>
</table>

We used three models to generate \( Y_i \):

\[
p_i = \frac{\exp(-1.5 + 0.005c_i + 1.24m_i)}{1 + \exp(-1.5 + 0.005c_i + 1.24m_i)} \quad \text{Model 4.0}
\]

\[
p_i = \frac{\exp(-0.6 + 0.0007c_i + 0.5m_i + 0.003c_i \times m_i)}{1 + \exp(-0.6 + 0.0007c_i + 0.5m_i + 0.003c_i \times m_i)} \quad \text{Model 4.1}
\]

\[
p_i = \frac{\exp(-0.6 + 0.0007c_i + 0.5m_i + 0.01c_i \times m_i)}{1 + \exp(-0.6 + 0.0007c_i + 0.5m_i + 0.01c_i \times m_i)} \quad \text{Model 4.2}
\]

Model 4.0 is the model under null containing only the linear term, and Model 4.1 and Model 4.2 are two alternative models designed to assess the power of detecting the omission of an interaction component. The variation in the coefficients of the last two models is used to emphasize and deemphasize the effect of omitting an interaction term [10]. Table 3.7 shows the the percent of time \( H_0 \) was rejected by each of the six tests using the NHANES data.
Table 3.7: Proportion $H_0$ rejected at the $\alpha = 0.05$ using sample size of 373 with 1000 replications in interaction models using the NHANES data

<table>
<thead>
<tr>
<th>Model</th>
<th>Test statistics</th>
<th>Proportion $H_0$ rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $H_0$: Model 4.0</td>
<td>$C_n$</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>$IM$</td>
<td>0.046</td>
</tr>
<tr>
<td>Under $H_a$: Model 4.1</td>
<td>$C_n$</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.105</td>
</tr>
<tr>
<td></td>
<td>$IM$</td>
<td>0.092</td>
</tr>
<tr>
<td>Under $H_a$: Model 4.2</td>
<td>$C_n$</td>
<td>0.326</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 10)$</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td>$HL(g = 20)$</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>USS</td>
<td>0.332</td>
</tr>
<tr>
<td></td>
<td>Stukel</td>
<td>0.294</td>
</tr>
<tr>
<td></td>
<td>$IM$</td>
<td>0.238</td>
</tr>
</tbody>
</table>

In Table 3.7, we see that under null, proportions of $H_0$ being rejected among all statistics remain close to the nominal level of 0.05. Except two Hosmer-Lemeshow tests, the power of the other tests increases to around 10% when omitting an interaction term with coefficient of 0.03. As the coefficient increases to 0.01, the power of almost all statistics increases to around 30%. Moreover, only $C_n$ attains high power in both situations.
3.0.3 Summary

In the current work, we have proposed a new statistic for GOF in a binary response model. This statistic denoted $C_n$ is based on the sum of standardized residuals. Under the regularity assumptions on the model, we have shown that $C_n$ asymptotically follows a normal distribution which can be used to perform p-value calculations for GOF. Under the null model, we see that $H_0$ is rejected about 5% of the time as long as the sample size is not too small. Under the alternative, the overall performance of $C_n$ is satisfying and stable across most of the scenarios considered. When the binary response model contains multiple covariates, the overall performance of $C_n$ dominates the performance of alternative GOF statistics considered. Based on our results, it appears that $C_n$ is a well-powered and trustworthy goodness of fit statistic which is also simple to use.
Bibliography


Appendix A: Proof of the Asymptotic Distribution of $C_n$

Define

$$S_n(\hat{\beta}_n) = \sum_{i=1}^{n} \frac{(Y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}$$

Local expansion of $n^{-\frac{1}{2}}S_n(\hat{\beta}_n)$ around $\beta_0$ gives:

$$n^{-\frac{1}{2}}S_n(\hat{\beta}_n) = n^{\frac{1}{2}}S_n(\beta_0) + n^{-1} \frac{\partial S_n(\beta)}{\partial \beta} |_{\beta_0} n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) + o_p(1)$$

where $S_n(\beta_n) = \sum_{i=1}^{n} \frac{(Y_i - p_i)}{\sqrt{p_i(1 - p_i)}}$.

$$n^{-1} \frac{\partial S_n(\beta)}{\partial \beta} |_{\beta_0} = n^{-1} \sum_{i=1}^{n} f_{i0} x_i \left( -\frac{1}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}} - \frac{1}{2} (Y_i - p_0)(p_0(1 - p_0))^{-\frac{3}{2}}(1 - 2p_0) \right)$$

$$= n^{-1} \sum_{i=1}^{n} \left( -\frac{1}{2} f_{i0} x_i (Y_i - p_0)(p_0(1 - p_0))^{-\frac{3}{2}}(1 - 2p_0) - n^{-1} \sum_{i=1}^{n} \frac{f_{i0} x_i}{\sqrt{p_0(1 - p_0)}} \right)$$

$$=: U_n - v_n$$

where

$$U_n = -n^{-1} \sum_{i=1}^{n} \frac{1}{2} f_{i0} x_i (Y_i - p_0)(p_0(1 - p_0))^{-\frac{3}{2}}(1 - 2p_0)$$

$$v_n = n^{-1} \sum_{i=1}^{n} \frac{f_{i0} x_i}{\sqrt{p_0(1 - p_0)}}.$$
We show that now \( \text{plim} \ (U_n) = 0 \) since \( E(U_n) = 0 \) and for each component of vector \( U_n \), say \( U_{j,n} \), we have \( E(U_{j,n}^2) \to 0 \) as \( n \to \infty \):

\[
E(U_{j,n}^2) = \frac{1}{4n^2} \sum_{i=1}^{n} \frac{(1 - 2p_i)^2}{(p_i(1 - p_i))^2} f_{i0} x_{j,i}
\]

\[
\leq \frac{1}{4n^2 A} \sum_{i=1}^{n} f_{i0} x_{j,i}
\]

\[
\leq \frac{D}{4nA} \left( \frac{\sum_{i=1}^{n} x_{j,i}}{n} \right)
\]

where \( 0 < A \leq p_i, f_{i0} \leq D < \infty \) and \( \sum_{i=1}^{n} x_{j,i} \) converges because of assumption 1-3.

Furthermore \( v_n \) is bounded, since

\[
||v_n||^2 = ||n^{-1} \sum_{i=1}^{n} f_{i0} x_i \sqrt{p_{i0}(1 - p_{i0})}||^2
\]

\[
\leq n^{-1} \sum_{i=1}^{n} ||f_{i0} x_i \sqrt{p_{i0}(1 - p_{i0})}||^2
\]

\[
= n^{-1} \sum_{i=1}^{n} \frac{f_{i0}^2}{p_{i0}(1 - p_{i0})} ||x_i||^2
\]

\[
= n^{-1} \text{tr} \sum_{i=1}^{n} \frac{f_{i0}^2}{p_{i0}(1 - p_{i0})} x_i x_i^\top
\]

and \( n^{-1} \text{tr} \sum_{i=1}^{n} x_i x_i^\top \) converges under our assumptions. Therefore, it follows that \( n^{-\frac{1}{2}} S_n(\hat{\beta}_n) \) has the same limiting distribution as

\[
n^{-\frac{1}{2}} S_n(\beta_0) - v_n^\top n^{\frac{1}{2}} (\hat{\beta}_n - \beta_0)
\]

Using Result 1 in Section 2 we know that \( \hat{\beta}_n - \beta_0 = -(\frac{\partial \log L}{\partial \beta} |_{\beta_n})^{-1} \frac{\partial \log L}{\partial \beta} |_{\beta_0} \), and therefore

\[
v_n^\top n^{\frac{1}{2}} (\hat{\beta}_n - \beta_0) = v_n^\top n^{-\frac{1}{2}} (A^*_n)^{-1} \frac{\partial \log L}{\partial \beta} |_{\beta_0}
\]

Where \( A^*_n = -n^{-1} \frac{\partial \log L}{\partial \beta} |_{\beta_n} \).

Since \( v_n \) is bounded, \( n^{\frac{1}{2}} (\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Omega^{-1}) \) and \( \text{plim} \ \Omega^{-1} A^*_n = I \), we have

\[
\text{plim} \ v_n^\top (\Omega^{-1} A^*_n - I) n^{\frac{1}{2}} (\hat{\beta}_n - \beta_0) = 0
\]
By the Theorem 1.10 from Shao’s book [12], \( n^{-\frac{1}{2}} S_n(\hat{\beta}_n) \) has the same limiting distribution as

\[
    n^{-\frac{1}{2}} S_n(\beta_0) - v_n^\top n^{-\frac{1}{2}} \Omega^{-1} \frac{\partial \log L}{\partial \beta} |_{\beta_0}.
\]

Since \( \frac{\partial \log L}{\partial \beta} |_{\beta_0} = \sum_{i=1}^{n} \frac{Y_i - p_{i0}}{p_{i0}(1 - p_{i0})} f_{i0} x_i \) then

\[
    n^{-\frac{1}{2}} S_n(\beta_0) - v_n^\top n^{-\frac{1}{2}} \Omega^{-1} \frac{\partial \log L}{\partial \beta} |_{\beta_0} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{Y_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}} - v_n^\top n^{-\frac{1}{2}} \Omega^{-1} \sum_{i=1}^{n} \frac{Y_i - p_{i0}}{p_{i0}(1 - p_{i0})} f_{i0} x_i
\]

\[
    = n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{(Y_i - p_{i0})(\sqrt{p_{i0}(1 - p_{i0})} - v_n^\top \Omega^{-1} f_{i0} x_i)}{p_{i0}(1 - p_{i0})}.
\]

Define

\[
    Z_{ni} = \frac{(Y_i - p_{i0})(\sqrt{p_{i0}(1 - p_{i0})} - v_n^\top \Omega^{-1} f_{i0} x_i)}{p_{i0}(1 - p_{i0})}
    = \frac{Y_i - p_{i0}}{\sqrt{p_{i0}(1 - p_{i0})}} \left( 1 - \frac{v_n^\top \Omega^{-1} f_{i0} x_i}{\sqrt{p_{i0}(1 - p_{i0})}} \right)
\]

\( Z_{n1}, ..., Z_{nn} \) are independent random variables. The moments of \( Z_{ni} \) are

\[
    E(Z_{ni}) = 0
\]

\[
    E(Z_{ni}^2) = \left( 1 - \frac{v_n^\top \Omega^{-1} f_{i0} x_i}{\sqrt{p_{i0}(1 - p_{i0})}} \right)^2 = \sigma_{ni}^2
\]

\[
    E(|Z_{ni}|^3) = \frac{(1 - 2p_{i0} + 2p_{i0}^2)}{\sqrt{p_{i0}(1 - p_{i0})}} \left| 1 - \frac{v_n^\top \Omega^{-1} f_{i0} x_i}{\sqrt{p_{i0}(1 - p_{i0})}} \right|^3 = m_{ni}
\]

22
By Liapounov’s Central Limit Theorem, if \( \lim_{n \to \infty} \frac{\sum_{i=1}^{n} m_{ni}}{(\sum_{i=1}^{n} \sigma_{ni}^2)^{3/2}} = 0 \) is satisfied:

\[
\frac{\sum_{i=1}^{n} m_{ni}}{(\sum_{i=1}^{n} \sigma_{ni}^2)^{3/2}} \leq \frac{\sum_{i=1}^{n} m_{ni}}{(n \min(\sigma_{ni}^2)^{3/2})}
= \frac{\sum_{i=1}^{n} m_{ni}}{n^{3/2}(\min \sigma_{ni})^3}
\leq \frac{1}{\sqrt{n}} \max_{n} m_{ni}
= O \left( \frac{1}{\sqrt{n}} \right)
\]

Then a sum of weighted residuals converges in distribution to the standard normal random variable, as \( n \) goes to infinity:

\[
\sum_{i=1}^{n} \frac{Y_i - p_i}{\sqrt{p_i(1-p_i)}} \left( 1 - \frac{v_n^T \Omega^{-1} f_0 x_i}{\sqrt{p_i(1-p_i)}} \right) \left( \sum_{i=1}^{n} \left( 1 - \frac{v_n^T \Omega^{-1} f_0 x_i}{\sqrt{p_i(1-p_i)}} \right)^{-2} \right)^{1/2} \xrightarrow{d} N(0, 1)
\]

Finally, by result 1 and equation 2.3, \( \sigma_n^2 \) can be rewritten as:

\[
\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left( 1 - \frac{v_n^T \Omega^{-1} f_0 x_i}{\sqrt{p_i(1-p_i)}} \right)^2
= 1 - 2v_n^T \Omega^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{f_0 x_i}{\sqrt{p_i(1-p_i)}} + v_n^T \Omega^{-1} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f_i^2}{p_i(1-p_i)} x_i x_i^T \right) \Omega^{-1} v_n
= 1 - v_n^T \Omega^{-1} v_n
\]