Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 434-445
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Asymptomatic Distribution of Goodness-of-Fit Tests in Logistic Regression Model

Nuri H. Salem Badi

Mathematics Department, Faculty of Art and science-Alabyar University of Benghazi, Benghazi, Libya
Email: nuri.badi@uob.edu.ly

How to cite this paper: Badi, N.H.S. (2017) Asymptomatic Distribution of Goodness-of-Fit Tests in Logistic Regression Model. *Open Journal of Statistics*, 7, 434-445.
https://doi.org/10.4236/ojs.2017.73031

Received: April 17, 2017
Accepted: June 9, 2017
Published: June 12, 2017

Copyright © 2017 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

The logistic regression model has been become commonly used to study the association between a binary response variable; it is widespread application rests on its easy application and interpretation. The subject of assessment of goodness-of-fit in logistic regression model has attracted the attention of many scientists and researchers. Goodness-of-fit tests are methods to determine the suitability of the fitted model. Many of methods proposed and discussed for assessing goodness-of fit in logistic regression model, however, the asymptotic distribution of goodness-of-fit statistics are less examine, it is need more investigated. This work, will focus on assessing the behavior of asymptotic distribution of goodness-of-fit tests, also make comparison between global goodness-of-fit tests, and evaluate it by simulation.

Keywords

Logistic Regression Model, Goodness-of-Fit Tests

1. Introduction

The goal of a logistic regression analysis is to find the best fitting model to describe the relationship between an outcome and covariates where the outcome is dichotomous, [1] considered the logistic regression model is a member of the class of the generalized linear models. Many assumptions and more details considered about the behavior of logistic model see [2] [3], also for more application see [4] [5] [6] [7]. The goodness-of-fit is very important to decide if the more succinct model is adequate. After fitting the logistic regression model, the next step is to examine the proposed model how well fits the observation data and to know how effective the model is; this is called as its goodness-of-fit. Goodness-of-fit tests for the logistic regression can be split into three types: 1) Those based an examination of residuals; 2) Those based a test which groups the observation; 3) Those which do not group observation. Methods in 1) are more general and subjective assessments of a model and are not considered in this work. This is not to undervalue then they are often the most valuable approach to model assessment. The observed values for Bernoulli regression are just 0 s and 1 s and this makes graphical approaches less easy to handle. The focus of this work is the test statistics. In next section, tests using grouping are considered, with those that do not need to group the data being discussed in section 3. Investigate the behavior of the asymptotic distribution of goodness-of-fit tests is considered in section 4 with comparisons between some goodness-of-fit tests, evaluated by simulation data with two different sample sizes. The simulation in this work was designed according to simulation that made by [8], which made comparisons between some goodness-of-fit tests in logistic regression models with sparse data. The results of his simulation showed that some goodness-of-fit tests have reasonable power compared with other tests. However, Kuss did not give information about the asymptotic distribution of these statistics. This paper supposes to show the behavior of the asymptotic distribution of goodness-of-fit tests for logistic regression model. Finally, conclusion and further discussion made in the last section.

2. Goodness-of-Fit Tests with Grouping

[9] proposed and developed approaches involving grouping based on the values of the estimated probabilities obtained from the fitted logistic model. Two grouping methods were proposed. The first approach is based on grouping the data according to percentiles of the estimated probabilities, and the second approach is based on grouping the data according to fixed cutoff values of the estimated probabilities. Tests with grouping based on estimated probabilities were proposed and developed by [9] [10] [11]. [12] developed a score test statistic which essentially compares two fitted model.

Hosmer and Lemeshow Test $\hat{C}$ The calculation of this test dependent upon grouping of estimated probabilities $\hat{\pi}(x_i)$ which use g groups. The first group contains the $n_1 = n/g$ observations which have the smallest estimated probabilities, the second group contains $n_2 = n/g$ values have the next smallest estimated probabilities and the last group contains the $n_g = n/g$ observation with the largest $\hat{\pi}(x_i)$: here n is the size of the sample and g the total number of groups. Before defining a formulae to calculate $\hat{C}$ we will consider some notions. The statistic test $\hat{C}$ is obtained by calculating Pearson chi-square statistic from the $2 \times g$ table with two rows and g columns of observed and expected frequencies. In the row with $y = 1$ summing of the all estimated probabilities in a group give the estimated expected value. In the row with $y = 0$ estimated expected value is obtained by summing one minus the estimated probabilities over all subjects in the group. We can denotes the observed number of subjects have had the event present $(y=1)$ and absent $(y=0)$ respectively in each group columns g $(s = 1, 2, 3, \cdots, g)$:

$$O_{1s} = \sum_{i=1}^{n_s} y_i,\ O_{0s} = \sum_{i=1}^{n_s} (1 - y_i)$$

where n_s is the number of the observation in group *g*. The expected number of subjects of present and absent respectively is denoted by:

$$E_{1s} = \sum_{i=1}^{n_s} \hat{\pi}_i,\ E_{0s} = \sum_{i=1}^{n_s} (1 - \hat{\pi}_i)$$

Then $\hat{C}$ is simply obtained by calculation the Pearson χ^2 statistic for the observed and expected frequencies from the $2 \times g$ table as:

$$\hat{C} = \sum_{s=1}^{g} \sum_{j=0}^{1} \frac{(O_{js} - E_{js})^2}{E_{js}}$$

from which it following

$$\hat{C} = \sum_{s=1}^{g} \left(\frac{(O_{0s} - E_{0s})^2}{E_{0s}} + \frac{(O_{1s} - E_{1s})^2}{E_{1s}} \right)$$

and finally we get

$$\hat{C} = \sum_{s=1}^{g} \frac{(O_s - n_s \bar{\pi}_s)^2}{n_s \bar{\pi}_s (1 - \bar{\pi}_s)},$$

where, n_s is the total number of values in *s^th^* group, O_s is the number of responses for the number of covariates in the *s^th^* group, defining as

$$O_s = \sum_{i=1}^{n_s} y_i$$

where, $O_s = O_{1s} + O_{0s}$, and $\bar{\pi}_s$ is the average of the estimated probabilities which are defined as:

$$\bar{\pi}_s = \sum_{i=1}^{n_s} \frac{m_i \hat{\pi}_i}{n_s}.$$

Here, the number of observations within covariate pattern *i* is denoted by m_i. Use of an extensive set of simulations proved that when $m_i = 1$, where m_i is the individual binomial denominator and the fitted logistic model is the correct model, then the distribution of $\hat{C}$ is approximated by the χ^2 distribution with $(g - 2)$ degrees of freedom [9].

Hosmer and Lemeshow Test $\hat{H}$

The second grouping strategy was proposed from Hosmer and Lemeshow denoted by $\hat{H}$, this method depends upon grouping the estimated probabilities in groups based on fixed cutpoint, so each group contains all subjects with fitted probability located in specific intervals. For example, the cutpoint of the first group is $0.0 \le \hat{\pi}(x_i) < 0.1$, then this group contains all subjects with estimated probabilities located in this interval; the second group contains all subjects with estimated probabilities located between cutpoint $0.1 \le \hat{\pi}(x_i) < 0.2$ and the last group has interval $0.9 \le \hat{\pi}(x_i) < 1.0$.

The calculation of $\hat{H}$ uses exactly the same formulae used to calculate $\hat{C}$: the only difference between the two approaches is in the construction of the groups. The distribution of $\hat{H}$ is approximated by the χ^2 distribution with $(g-2)$ degrees of freedom.

Although Hosmer and Lemeshow tests are good, it requires grouping, and choice of g is

- g is arbitrary but almost everywhere in the literature and in software a value of 10, or very similar is chosen.
- Smaller values of g might be chosen for smaller n.
- Sparse data causes a problem for H and lead to uneven group widths for C.

3. Goodness-of-Fit Tests without Grouping

Deviance and Pearson Chi-Square Tests

Two of the most commonly used goodness-of-fit measures, are the Pearson's chi-squared χ^2 and the deviance D goodness-of-fit test statistics but the behaviour of these tests are unstable with bernoulli data; see [13]. The general idea of the deviance is make comparison between two models the first model is full model with p parameters and the second model is a model with q parameters, where $(q<p)$. The deviance can write as

$$D=-2\log\left(\frac{\hat{L}_s}{\hat{L}_r}\right)=-2\left(\ell_s-\ell_r\right),$$

where $\hat{L}_r$, $\hat{L}_s$ are the likelihoods for the full and small model and ℓ_r, ℓ_s denoted to the log-likelihood: Asymptotically this is χ^2 in $p-q$ df. The residual deviance is the case when the large model is saturated and has n parameters. In case of the logistic regression model [13] introduced specific form when $m_i=1$; the residual deviance can then be found as

$$D=-2\sum_{i=1}^{n}\left\{\hat{\pi}_i\log\hat{\pi}_i+\left(1-\hat{\pi}_i\right)\log\left(1-\hat{\pi}_i\right)\right\},$$

In this case the deviance is invalid as a goodness-of-fit test, because it is a function of $\hat{\pi}_i$, which does not compare the observed values with fitted values.

Also, [13] discussed that Pearson chi-square goodness of fit statistic when $m_i=1$; can be written:

$$X^2=\sum_{i=1}^{n}\frac{\left(y_i-\hat{\pi}\right)^2}{\hat{\pi}\left(1-\hat{\pi}\right)}=n$$

which is equal to the sample size: this is not a useful goodness-of-fit test.

Residual Sum of Squares Test

[14] proposed a method, which used the unweighted residual sum of squares a goodness-of-fit test to assess the model adequacy. The idea of this approach is to keep all the individual values of m_i but to give less weight in cases of m_i are small. The unweighted residual sum of squares statistic considers only the numerator of the Pearson chi-squares statistic, which is the summation again over the individual observations, the statistic can be written:

$$RSS=\sum_{i=1}^{n}\left(y_i-\hat{\pi}_i\right)^2.$$

Of course, the relative weighting for varying m_i is not relevant for our case where $m_i = 1$. [11] discussed how to compute the moments and asymptotic distribution of the RSS statistic. They give useful expressions for the mean and variance which are easier to compute than the expressions given by [14]. The proposed asymptotic mean and variance of RSS are respectively, $E\left[RSS - S(W)\right] \cong 0$ and var $\left[RSS - S(W)\right] \cong d^{\mathrm{T}}(I - M)Wd$, where $M = WX\left(X^{\mathrm{T}}WX\right)^{-1}X^{\mathrm{T}}$, $W = \operatorname{diag}\left[\pi_i(1-\pi_i)\right]$, $S(W) = \sum_{i=1}^{n}\left[\operatorname{diag}\left(\pi_i(1-\pi_i)\right)\right]$ and d is vector with elements $d_i = (1 - 2\pi_i)$. Used the standardized statistic to assess significance by referring the following to the standard normal

$$\frac{\left[RSS - S(W)\right]}{\sqrt{\operatorname{var}\left[RSS - S(W)\right]}}.$$

R^2 Test

Several R^2 type statistics have been used for goodness-of-fit in logistic regression, such as that proposed by [15].

$$R_g^2 = 1 - \left(\frac{\hat{L}_c}{\hat{L}_0}\right)^{n/2}$$

where, $\hat{L}_c$ represents the log-likelihood evaluated at the ML estimation parameters and $\hat{L}_0$ represents the log-likelihood of the model containing only an intercept. Another version due to [16] is

$$\bar{R}_g^2 = \frac{R_g^2}{\max\left(R_g^2\right)}$$

where, $\max\left(R_g^2\right) = 1 - \left(\hat{L}_0\right)^{2/n}$.

Information Matrix tests: IMT and IMT_{DIAG}

The Information Matrix test (*IMT*) is a test for general mis-specification, proposed by [17]. The two well-known expressions for the information matrix coincide only if the correct model has been specified and the *IMT* takes advantage of this fact. The *IMT* avoids the grouping necessary for tests like the Hosmer-Lemeshow test. Many researchers, [18] [19] [20] [21] pointed out the behaviour of the asymptotic distribution of *IMT* statistic and dispersion matrix. [22] discussed the information matrix test and showed that it is useful with binary data models. [8] claimed that, the *IMT* has reasonable power compared with other tests, without information about the behaviour of the asymptomatic distribution of *IMT*. The idea of the information matrix test is to compare $E\left(\frac{-\partial^2 \ell}{\partial\theta\partial\theta^T}\right)$ and $E\left(\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta^{\mathrm{T}}}\right)$, as these differ when the model is mis-specified but not when the model is correct.

Let, consider binary regression, where the outcome for individual *i*, $i = 1, \cdots, n$ is a random variable $Y_i \in \{0,1\}$. Also $\Pr(Y_i \mid x_i) = \pi_i = f\left(x_i^{\mathrm{T}}\beta\right)$ where x_i is a $p \times 1$ dimensional vector of covariates and β is a *p*-dimensional vector of parameters. It will be convenient to write $a_i = x_i^{\mathrm{T}}\beta$ and ℓ_i to be the contribution to the log-likelihood ℓ from unit *i*.

We have

$$\ell(\beta)=\sum_{i=1}^{n}\ell_i(\beta)=\sum_{i=1}^{n}Y_i\log\pi_i+(1-Y_i)\log(1-\pi_i)$$

The p-dimensional likelihood equations $\partial\ell/\partial\beta=0$ can be written:

$$\frac{\partial\ell}{\partial\beta}=\sum_{i=1}^{n}\left[\frac{(Y_i-\pi_i)}{\pi_i(1-\pi_i)}\right]\frac{\partial\pi_i}{\partial a_i}x_i=0 \tag{1}$$

We can also derive the $p\times p$ matrix $\partial^2\ell/\partial\beta\partial\beta^{\mathrm{T}}$ as:

$$\sum_{i=1}^{n}\left[\frac{(Y_1-\pi_i)}{\pi_i(1-\pi_i)}\frac{\partial^2\pi_i}{\partial a_i^2}-\frac{(Y_1-\pi_i)^2}{\pi_i^2(1-\pi_i)^2}\left(\frac{\partial\pi_i}{\partial a_i}\right)^2\right]x_ix_i^{\mathrm{T}} \tag{2}$$

The idea behind the information matrix test is that if the model is correctly specified then the quantity:

$$IM=\sum_{i=1}^{n}\left(\left.\frac{\partial\ell_i}{\partial\beta}\frac{\partial\ell_i}{\partial\beta^{\mathrm{T}}}\right|_{\hat\beta}+\left.\frac{\partial^2\ell_i}{\partial\beta\partial\beta^{\mathrm{T}}}\right|_{\hat\beta}\right)$$

has zero mean. By comparing (1) and (2) we can compute this quantity, for a general value of β, as the sum of:

$$\frac{\partial\ell_i}{\partial\beta}\frac{\partial\ell_i}{\partial\beta^{\mathrm{T}}}+\frac{\partial^2\ell_i}{\partial\beta\partial\beta^{\mathrm{T}}}=\frac{(Y_i-\pi_i)}{\pi_i(1-\pi_i)}\frac{\partial^2\pi_i}{\partial a_i^2}x_ix_i^{\mathrm{T}} \tag{3}$$

We can test the null hypothesis that *IM* has zero mean by computing the variance of *IM* and then constructing a standard χ^2 statistic. The first step is to compute the variance of $n^{-\frac{1}{2}}\sum d_i$ where we write d_i for essentially the right hand side of (3):

$$\frac{(Y_i-\pi_i)}{\pi_i(1-\pi_i)}\frac{\partial^2\pi_i}{\partial a_i^2}z_i$$

where we have changed the $p\times p$ symmetric matrix into a vector z_i in order to be able to use standard methods. As $x_ix_i^{\mathrm{T}}$ is symmetric we do not wish to duplicate entries, so z_i is the $\frac{1}{2}p(p+1)$-dimensional vector:

$$z_i^{\mathrm{T}}=\left(\left[x_{11},x_{21},\cdots,x_{p1}\right],\left[x_{22},x_{32},\cdots,x_{p2}\right],\cdots,\left[x_{(p-1),(p-1)},x_{p,(p-1)}\right],\left[x_{pp}\right]\right)$$

where x_{st} is the $(s,t)^{th}$ element of $x_ix_i^{\mathrm{T}}$. If we write:

$$\begin{aligned}A=n^{-\frac{1}{2}}\sum d_i&=n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\pi_i)}{\pi_i(1-\pi_i)}\frac{\partial^2\pi_i}{\partial a_i^2}z_i\\&=n^{-\frac{1}{2}}\sum_{i=1}^{n}(Y_i-\pi_i)(1-2\pi)z_i\end{aligned}$$

then because the different terms are independent we obtain:

$$\Psi=\operatorname{var}(A)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i(1-\pi_i)}\left(\frac{\partial^2\pi_i}{\partial a_i^2}\right)^2z_iz_i^{\mathrm{T}}.$$

which is a $q\times q$ dimensional matrix where $q=\frac{1}{2}p(p+1)$.

We should also note that if B is defined as essentially the log-likelihood, *i.e.*

$$B=n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{(Y_i-\pi_i)}{\pi_i(1-\pi_i)}\frac{\partial\pi_i}{\partial a_i}x_i=n^{-\frac{1}{2}}\sum_{i=1}^{n}(Y_i-\pi_i)x_i$$

then the variance of B is the $p\times p$ matrix Ω:

$$\Omega=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i(1-\pi_i)}\left(\frac{\partial\pi_i}{\partial a_i}\right)^2 x_i x_i^{\mathrm{T}}$$

Before compute the covariance of A and B, we get, using

$$(y_i-\pi_i)^2\pi_i(1-\pi_i)=(y_i-\pi_i)(1-2\pi_i)$$

Now,

$$\operatorname{cov}(A,B)=E(AB)-E(A)E(B)$$

For independently and identically random variables and under the H_0 the second term of the $\operatorname{cov}(A,B)$ is zero, and covariance of A and B in this case is the $q\times p$ matrix, and so

$$\Delta=\operatorname{cov}(A,B)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i(1-\pi_i)}\left(\frac{\partial\pi_i}{\partial a_i}\right)\left(\frac{\partial^2\pi_i}{\partial a_i^2}\right)z_i x_i^{\mathrm{T}}$$

Central limit arguments suggest that asymptotically $(A^{\mathrm{T}},B^{\mathrm{T}})$ is a $q+p$ dimensional normal variable. However, the *IM*-test requires A to be evaluated at $\hat{\beta}$, $\hat{A}$, say, and at this value we know that $B = 0$. Consequently the variance of $\hat{A}$ is the variance of A conditional on $B = 0$ which is $\Psi-\Delta\Omega^{-1}\Delta^{\mathrm{T}}$.

Assuming a logistic regression we have $\partial\pi_i/\partial a_i=\pi_i(1-\pi_i)$ and $\partial^2\pi_i/\partial a_i^2 =\pi_i(1-\pi_i)(1-2\pi_i)$ so we can evaluate the dispersion matrices at the MLEs as:

$$\hat{\Omega}=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)x_i x_i^{\mathrm{T}}$$

$$\hat{\Psi}=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)^2 z_i z_i^{\mathrm{T}}$$

$$\hat{\Delta}=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)(1-2\hat{\pi}_i)z_i x_i^{\mathrm{T}}$$

If we write $\hat{V}=\hat{\Psi}-\hat{\Delta}\hat{\Omega}^{-1}\hat{\Delta}^{\mathrm{T}}$ then one version of the *IM* test is found by referring $\hat{A}^{\mathrm{T}}\hat{V}^{-1}\hat{A}$ to a χ^2 variable with degrees of freedom equal to the rank of $\hat{V}$.

The idea of the *IM*$_{DIAG}$ test and *IM* test are the same, the only difference is that for the former the elements of z_i are just the diagonal elements of $x_i x_i^{\mathrm{T}}$, so z_i is the p dimensional vector:

$$z_i^{\mathrm{T}}=\left(x_{i1}^2,x_{i2}^2,\cdots,x_{ip}^2\right).$$

To explain the difference in size of vector z_i in the two cases of *IM* test and *IM*$_{DIAG}$ test, let us consider a simple example. Suppose we have a symmetric matrix with elements $x_i x_i^{\mathrm{T}}$ and 3×3 dimension as:

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix},$$

where, $x_{rs} = x_{ri}x_{si}$. Then in the case of the *IM* test, the dimension of vector z_i^{T} is 1×6 and elements are:

$$z_i^{\mathrm{T}} = \left[x_{11}, x_{12}, x_{13}, x_{22}, x_{23}, x_{33}\right],$$

whereas in the case of IM_{DIAG} test, z_i is the 1×3 dimensional vector:

$$z_i^{\mathrm{T}} = \left[x_{11}, x_{22}, x_{33}\right].$$

4. Simulation Study

Our work, focus on behaviour of goodness of fit tests under alternative hypotheses in case of missing covariate model and in case of the wrong model, because these cases we could not reproduce Kuss's work in. We will focus on four goodness-of-fit tests $\left(\hat{C}_g, RSS, IM, IM_{DIAG}\right)$. Therefore, we examine in more depth the behaviour of the tests and determine more information about asymptotic MLE distribution in case of the wrong model

$$\pi_i = \operatorname{expit}\left(0.405 x_i^2\right),$$

or in the case of the missing covariate,

$$\pi_i = \operatorname{expit}\left(0.405 x_i + 0.223 u_i\right),$$

where $X, U \sim U\left(-6, 6\right)$, X and U independent.

Simulation study designed as Kuss's work:

- The sample sizes are $n = 100$ and $n = 500$.
- Applied only on extreme sparseness when $m_i = 1$.
- number of simulation is 1000.
- distribution of the predictor variables X, U is $U\left(-6, 6\right)$, X and U independent chosen to confirm with Kuss's work.
- Use four of goodness-of-fit tests from the simulation study under three different alternative hypotheses:

  (a) True covariate.

  (b) Missing covariate.

  (c) Wrong functional form of the covariate.
- Fitted model in all cases is a standard logistic model with an intercept and one covariate.
- All the tests on the null hypothesis under $\alpha = 0.05$.

Results and Discussion of Tests under Correct Model

In Table 2, reported some results, the mean, variance and the empirical power of four goodness-of-fit tests from simulation study under correct model, namely

$$\pi_i = \operatorname{expit}\left(0.693 x_i\right).$$

Statistics used in the simulation as goodness-of fit tests are: Hosmer-Lemeshow $\left(\hat{C}_g\right)$, Information matrix $\left(IM\right)$, Information matrix Diagonal

(IM_{DIAG}) and residual sum of squares (*RSS*). The asymptotic distribution of statistics is χ^2_{df} distribution, where the mean and variance equal *df* and 2*df* respectively. In case of $(\hat{C}_g)$ statistic we chosen the number of group is *g* = 10 so, degree of freedom is $df = g - 2$. The results shown in **Table 1**, the mean and variance of all statistics appeared close to *df* and 2*df*. Moreover, the simulation study appeared reasonable results when fit the model with sample size *n* = 500. However, there is slightly large variance of $(\hat{C}_g)$ in case of sample size *n* = 100. Overall, the empirical power and type I error looks good.

In the second case, the results reported the mean, variance and the power to detect a mis-specified model for same goodness-of-fit tests under missing covariate model, when the model is:

$$\text{logit}(\pi_i) = \text{expit}(0.405x_i + 0.223u_i),$$

and fit standard logistic regression model with x_i.

Table 2, showed results from simulation study under alternative hypotheses missing covariate model. The mean and variance of all statistics close to *df* and 2*df*, but we have slightly smaller variance in case of $\hat{C}_g$. However, we have low power when used *IM* statistics in case of sample size *n* = 500, IM_{DIAG} statistic and *RSS* in case of sample size *n* = 100 and $\hat{C}_g$ statistic in both cases of sample size.

The final case we will show the results of power to detect a mis-specified model for four goodness-of-fit tests under the wrong functional form of the covariate model

$$\text{logit}(\pi_i) = \text{expit}(0.405x_i^2)$$

and fit the model as previous cases.

Table 1. Results of *N* = 1000 simulation with sample size *n* = 100 and *n* = 500 under correct model.

		n = 100			*n* = 500		
-	df	Mean	Var	%Rej	Mean	Var	%Rej
$\hat{C}_g$	8	8.06	20.47	4.6	7.96	17.12	5.70
IM	3	3.06	7.23	5.10	3.00	6.33	4.70
IM_{DIAG}	2	2.04	3.97	5.50	1.94	3.63	4.20
RSS	1	0.98	1.81	4.60	0.99	1.83	4.10

Table 2. Results of *N* = 1000 simulation with sample size *n* = 100 and *n* = 500 under missing covariate model.

		n = 100			*n* = 500		
-	df	Mean	Var	%Rej	Mean	Var	%Rej
$\hat{C}_g$	8	7.44	11.13	1.50	7.35	12.62	3.20
IM	3	3.01	6.05	5.50	2.38	4.15	1.90
IM_{DIAG}	2	1.82	3.06	3.3	2.05	3.46	4.80
RSS	1	0.92	1.51	4.10	0.99	1.73	4.50

In Table 3, reported results for goodness-of-fit tests from simulation study under wrong model. The mean and variance of all statistics appeared very larger in two cases of sample size comparing with degree of freedom of statistics. However, high power in all goodness-of-fit tests in both sample size were found, that is meaning this tests have rejected all the null hypothesis. On the other hand, Kuss's results appeared low power in case of sample size n = 100 compared with our results.

In Figure 1, we plot π vs x and we show the true model (continues line). If we fit $\pi = \text{expit}(\alpha + \beta x)$, these putative approximation are shown for $\beta < 0$, $\beta > 0$ and $\beta = 0$ (dot and dash, dash and dot) line respectively.

Table 3. Results of N = 1000 simulation with sample size n = 100 and n = 500 under wrong model.

		n = 100			n = 500		
-	df	Mean	Var	%Rej	Mean	Var	%Rej
$\hat{C}_g$	8	31.50	75.73	98.8	133.73	382.62	100
IM	3	17.33	17.97	100	75.57	72.70	100
IM_D	2	16.85	16.64	100	76.28	71.82	100
RSS	1	17.07	17.16	100	76.17	163.84	99.5

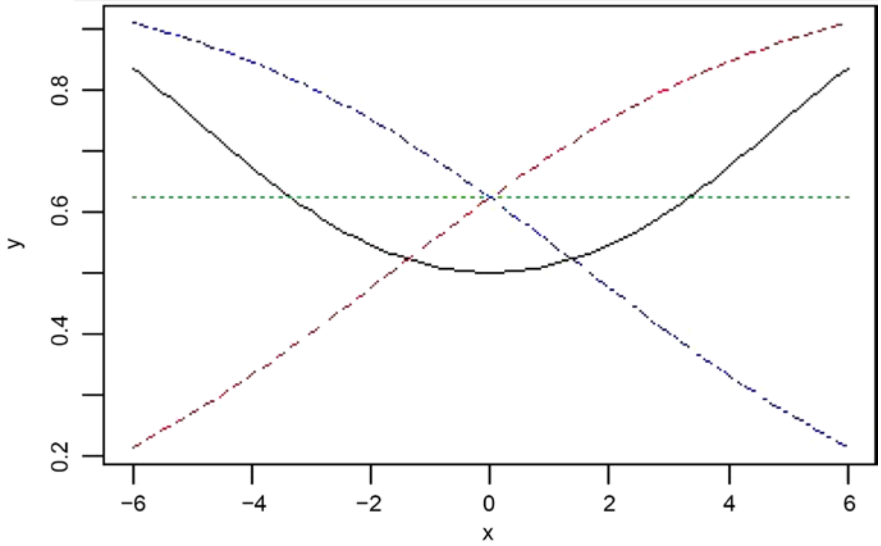

Figure 1. Plots of the different logistic model π_i given $X \sim U(-6,6)$.

5. Conclusion and Further Work

The work considered in this paper was centered on the asymptotic distribution of goodness-of-fit tests in logistic regression model. We also consider the comparison between some global goodness-of-fit tests, which compared with Kuss's results. Application of simulation apply in two types of goodness-of-fit tests, those based a test which groups the observation and those which do not group observation. Our results of study confirm the work of Kuss's regarding the power of goodness-of-fit tests, which related the Rss , Hosmer-Lemeshow, IM and IM_{DIAG} tests under correct and missing model. However, our results about the asymptomatic distribution of goodness-of-fit tests show, various combinations of behavior on the mean and variance of statistics, which, the asymptotic distribution of statistics is Chi-square χ^2_{df} . The results under correct model show reasonable power for all methods, slightly larger variance found in case of Hosmer-Lemeshow test, and smaller variance under missing covariate model. As we know the goodness-of-fit statistics are distributed asymptotically as central χ^2 distribution under H_0 when the model is correctly specified, and is non-central χ^2 under H_1 when the model mis-specificed. However, under wrong model the results show strange behavior, which all the means and variances are not satisfy the assumption on asymptotic distribution χ^2_{df} with men df and variance $2df$, also, it is appeared with high power. The problem means that in some circumstances properties of the distribution of the statistics of tests (e.g mean and variance) are far away from the properties of χ^2 distribution. In fact, the interesting point here, some of goodness-of-fit tests seem affected by assumption on covariance matrix. So, many issues about the mean and variance of the asymptotic distribution of goodness-of-fit statistic should also be examined.

References

[1] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384. https://doi.org/10.2307/2344614

[2] Dobson, A. (1990) An Introduction to Generalized Linear Models. Chapman and Hall, London.

[3] Kleinbaum, D.G. (1994) Logistic Regression A Self-Learning Text. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4757-4108-7

[4] Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression. Wily, Chichester. https://doi.org/10.1002/0471722146

[5] Hosmer, D., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd Edition, Wily, Chichester. https://doi.org/10.1002/9781118548387

[6] Hilbe, J.M. (2009) Logistic Regression Model. Chapman and Hall, New York.

[7] Dobson, A.J. and Barnett, A.G. (2008) An Introduction to Generalized Linear Models. 3rd Edition, Chapman and Hall, New York.

[8] Kuss, O. (2002) Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data. *Statistics in Medicine*, **21**, 3789-3801. https://doi.org/10.1002/sim.1421

[9] Hosmer, D.W., Hosmer, T. and Lemeshow, S. (1980) A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics, **10**,* 1043-1069. https://doi.org/10.1080/03610928008827941

[10] Lemeshow, S. and Hosmer, D.W. (1982).A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models. *American Journal of Epidemiology*, **115**, 92-106. https://doi.org/10.1093/oxfordjournals.aje.a113284

[11] Hosmer, D.W., Hosmer, T., Le Cessie, S. and Lemeshow, S. (1997) A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine*, **16**, 965-980.

https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O

[12] Brown, C.C. (1982) On A Goodness of Fit Test for the Logistic Model Based on Score Statistics. *Communications in Statistics Theory and Methods*, **10**, 1097-1105. https://doi.org/10.1080/03610928208828295

[13] McCullagh, P. and Nelder, J.A. (1989) Linear Models. 2nd Edition, Chapman and Hall, London.

[14] Copas, J.B. (1989) Testing for Neglected Heterogeneity. *Econometrica*, **52**, 865-872.

[15] Cox, D.R. and Snell, E.J. (1989) Analysis of Binary Data. 2nd Edition, Chapman and Hall/CRC, London.

[16] Nagelkerke, N.D. (1991) A Note on a General Definition of the Coefficient of Determination. *Biometrika*, **3**, 691-692. https://doi.org/10.1093/biomet/78.3.691

[17] White, H. (1982) Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50**, 1-25. https://doi.org/10.2307/1912526

[18] Lancaster, T. (1984) Covariance Matrix of the Information Matrix Test. *Econometrica*, **4**, 1051-1053. https://doi.org/10.2307/1911198

[19] Newey, W.K. (1984) Maximum Likelihood Specification Testing and Conditional Moment Test. *Econometrica*, **53**, 1047-1070.

[20] Davidson, R. and Mackinnon, J.G. (1984) Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics*, **25**, 241-262.

[21] Orme, C. (1988) The Calculation of the Information Matrix Test for Binary Data Models. *EconPapers*, **56**, 370-376. https://doi.org/10.1111/j.1467-9957.1988.tb01339.x

[22] Chesher, A. (1984) Unweighted Sum of Squares Test for Proportions. *Econometrica*, **38**, 71-80.

Scientific Research Publishing

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact ojs@scirp.org