TESTING GOODNESS-OF-FIT FOR LOGISTIC REGRESSION WITH SURVEY DATA

Barry I Graubard, Edward L Korn, National Cancer Institute;
Douglas Midthune, Information Management Services
Barry I Graubard, Biometry Branch, EPN-344, National Cancer Institute, Bethesda, MD 20892

Key Words: Wald test, Simulation, Sample weights, Cluster sampling

1. Introduction

Logistic regression is used to model the probability of a positive outcome for a binary 0-1 outcome variable as a function of covariates. An example, described latter in more detail, is a logistic regression model predicting snuff use as a function of demographic variables--age, income, education, etc. After choosing a logistic regression model and estimating its parameters from the data, we are interested assessing how consistent the model is with the data. This is referred to as goodness-of-fit (Korn and Simon, 1991). An early qualitative approach for examining goodness-of-fit of logistic regression models was described by Truett, Cornfield and Kannel (1967). They divided the data into cells of deciles of risk and compared the expected cell counts of the outcome variable with the observed counts without a formal statistical test. Hosmer and Lemeshow (1980) proposed a Pearson chi-squared statistic to test globally whether the expected cell counts were different from the observed; which is referred to as the Hosmer-Lemeshow (HL) test. Simulations were used to show that the HL test appeared to have approximately correct type I error (Hosmer and Lemeshow, 1980). An important assumption when using the HL test is that the data come from a simple random sample.

Data from sample surveys are typically not from simple random samples but are from complex samples with sample designs involving cluster sampling and differential probabilities of selection. The cluster sampling can induce correlation among observations from the same sampled cluster and the differential probabilities of selection can require sample weighting for unbiased estimators. Ignoring these aspects of complex samples can give invalid statistical tests (Skinner, Holt and Smith, 1989). In this paper, we examine the properties of goodness-of-fit tests for logistic regression in complex samples. Section 2 describes procedures for testing for goodness-of-ft in simple random samples, and section 3 compares the type I errors of these procedures using simulations. Procedures for testing goodness-of-fit in complex samples are described in section 4, with simulations comparing the type I errors of these procedures given in section 5. In section 6, we give an example using data from the 1987 National Health Interview Survey (NHIS) of testing the goodness-of-fit for a logistic regression analysis predicting snuff use. A brief discussion is given in section 7.

2. Testing Goodness-of-Fit for Logistic Regression in Simple Random Samples

Let y be a binary outcome variable and $\mathbf{x} = (x_1, ..., x_p)'$ be a vector of covariates. Under the logistic regression model

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \qquad (1)$$

where $\pi = \Pr(y=1|\ \mathbf{x}\ )$ the β's can be estimated using maximum likelihood theory and substituted into (1) to obtain the predicted probability $\hat{\pi}$ for each observation. We focus on testing for goodness-of-fit when the number of different $\mathbf{x}$ configurations is approximately the sample size n. Other approaches such as likelihood ratio tests comparing (1) to a fully saturated model can be used when the number of different $\mathbf{x}$'s configurations is small compared to the sample size n (Bishop, Fienberg and Holland, 1975, pp 524-526).

The approach used for testing good-of-fit is to compare the observed to expected (from the estimated model) number of outcomes for values of $\mathbf{x}$. Since there are too few observations for each $\mathbf{x}$ configuration, Hosmer and Lemeshow (1980) followed Truett, Cornfield and Kannel (1967) and divided the data into g=10 deciles of risk groups to do the comparison. These groups are formed by dividing up the observations so that $n_1 \simeq n/10$ observations with the smallest estimated probabilities are in the first group, $n_2 \simeq n/10$ observations with next smallest estimated probabilities are in the second group and so forth until the last group is formed with $n_{10} \simeq n/10$ observations with the largest estimated probabilities. The observed number of outcomes in the kth decile is $o_k = \sum_{j=1}^{n_k} y_{kj}$ and the expected number of outcomes is $e_k = \sum_{j=1}^{n_k} \hat{\pi}_{kj}$, where y_{kj} and $\hat{\pi}_{kj}$ are the outcome and predicted probability for observation j in decile-of-risk group k. The HL test uses a Pearson chi-square statistic $C = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{e_k(1-e_k/n_k)}$, where n_k is the number of observations in the kth decile of risk group. The distribution of C was determined by simulation studies to be approximately a χ^2 with 8 degrees of freedom (Hosmer and Lemeshow, 1980).

The HL test rejects the fit of the logistic model at the α level when $C > \chi^2_{8,1-\alpha}$, where $\chi^2_{t,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2distribution with t degrees of freedom.

Another method to test for goodness-of-fit is based on the Wald statistic $W = (\mathbf{o}-\mathbf{e})'\mathbf{S}^{-1}(\mathbf{o}-\mathbf{e})$, where $\mathbf{o} = (o_1, ..., o_{10})'$, $\mathbf{e} = (e_1, ..., e_{10})'$ and $\mathbf{S}$ is a consistent estimator of the covariance matrix of $\mathbf{o}-\mathbf{e}$; for example, one obtained using the Taylor series (linearization) approximation. The Wald test rejects the fit of the logistic model at the α level when $W > \chi^2_{9,1-\alpha}$. (Note that $\sum_{k=1}^{10} o_k = \sum_{k=1}^{10} e_k$.)

We find in a simulation study presented in the next section that the Wald test has an inflated type I error even in the case of simple random sampling. This brings into question the accuracy of using the χ^2 as a reference distribution for W for more complex sampling.

A third approach is to use a simulation method to compute the p-value for W. The p-value for this simulated Wald test is computed as follows: (1) the dataset is fit to a logistic regression model and its W is computed; (2) under this logistic model using the observed covariates, n binary outcomes corresponding to the observations in the dataset are repeatedly generated to create 999 simulated datasets; (3) the p-value for the simulated Wald test is computed as one plus the number of W's for the simulated datasets that are greater than or equal to the W for the original dataset divided by 1000.

3. Simulations for Simple Random Samples

A limited simulation study was conducted to investigate the type I errors for the HL, Wald and simulated Wald tests for simple random samples. The simulation study consisted of 10,000 simulations where in each simulation $n = 2{,}000$ independent observations was generated as follows: Each observation consisted of a vector of three covariates $\mathbf{x} = (x_1, x_2, x_3)'$ that was generated from independent standard normal distributions; and an outcome y that was generated as a Bernoulli variable with logit $[\Pr(y = 1 \mid \mathbf{x})] = -1.4 + x_1$. The logistic regression model

$$\text{logit}[\Pr(y = 1 \mid \mathbf{x})] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (2)$$

was used to fit each simulated dataset and the three tests were computed. The simulation results for the type I error of simulated Wald were computationally intensive because it involved doing 999 simulations within each of the 10,000 simulations. In this simulation study, we are under the null hypothesis for testing goodness-of-fit because the logistic regression model used to generate the data was contained in the model used to fit the data. Thus, we were able to estimate the type I errors for the three test procedures (Table 1).

Table 1. Type I Error of Goodness-of-Fit Tests for Simple Random Samples

Test	Nominal Type I Error	
	.05	.10
Hosmer-Lemeshow	.052	.10
Wald	.11	.17
Simulated Wald	.050	.10

The Wald test has an inflated type I error while the type I error for the HL and simulated Wald tests are at the nominal level.

4. Testing Goodness-of-Fit for Logistic Regression in Complex Samples

Now we consider the logistic regression model (1) for complex samples. The complex sample can involve multistage stratified cluster sampling with differential sample weighting due to selection probabilities and adjustments from nonresponse or poststratification. Because of the complex sampling, maximum likelihood estimation would not, in general, be valid. Instead, we use pseudo-maximum likelihood estimation to estimate the β's (Skinner, Holt and Smith, 1989, pp 80-84). The pseudo-likelihood is given by

$$\prod_{i=1}^{n} \pi_i^{w_i y_i}[1-\pi_i]^{w_i(1-y_i)}$$

where the w_i are the sample weights. The weighted estimates $\hat{\beta}$'s are the β's that maximize the pseudo-likelihood. Substituting these $\hat{\beta}$'s into (1), we obtain the weighted predicted probabilities $\hat{\pi}_i$'s. Consistent standard errors for the $\hat{\beta}$'s and $\hat{\pi}_i$'s can be estimated using a linearization approximation (Binder, 1983). For sample designs where the all sample weights are the same, the $\hat{\beta}$'s are exactly the maximum likelihood estimates. However, even in this case the standard error estimates would not be the same as those from maximum likelihood theory because they would have to reflect the variability of the estimator due to the complex sampling e.g., cluster sampling.

To assess goodness-of-fit, the data are divided into weighted decile groups which have a weighted one-tenth of the n observations in the dataset in each group: n_1 observations with the smallest predicted probabilities are in the first group where n_1 is chosen so that $\sum_{i=1}^{n_1} w_{1i} \Big/ \sum_{i=1}^{n} w_i \simeq .1$; n_2 observations with the next smallest predicted probabilities are in the second group where n_2 is chosen so that $\sum_{i=1}^{n_2} w_{2i} \Big/ \sum_{i=1}^{n} w_i \simeq .1$; and so

forth until the tenth group is formed with n_{10} observations with the largest predicted probabilities in the tenth group where n_{10} is chosen so that $\sum_{i=1}^{n_{10}} w_{10i} \Big/ \sum_{i=1}^{n} w_i \simeq .1$. Here a w_{ki} is the sample weight for the ith observation in the kth (weighted) decile of risk. In the kth decile, the weighted number of observed outcomes is $o_k = \sum_{i=1}^{n_k} w_{ki} y_{ki}$ and the weighted number of expected outcomes is $e_k = \sum_{i=1}^{n_k} w_{ki} \hat{\pi}_{ki}$.

A Wald test statistic for complex samples is based on $W_d = (\mathbf{o} - \mathbf{e})' \mathbf{S}_d^{-1} (\mathbf{o} - \mathbf{e})$ where $\mathbf{o}$ is vector of the weighted number of outcomes, $\mathbf{e}$ is the vector of the weighted number of expected outcomes and $\mathbf{S_d}$ is a (design) consistent estimator for the covariance matrix of $\mathbf{o} - \mathbf{e}$; for example, one obtained by linearization. The Wald test rejects the fit of the logistic model when $W_d > \chi^2_{9,1-\alpha}$.

For the moderately large samples sizes in our simulation study for simple random samples, the Wald test statistic did not appear to be close enough to its chi-squared asymptotic distribution to have nominal type I error. Since we expected to see a similar finding for complex samples, we considered a simulated Wald test for complex samples. The p-value for the simulated Wald test is computed using the following steps: (1) a logistic regression model is estimated for a dataset using the psuedo-likelihood and the W_d is computed for this estimated model; (2) 999 simulated datasets are created by repeatedly generating independent binary outcomes for each observed covariate vector according to the estimated model in step (1); (3) the logistic regression model used in step (1) was reestimated, using the psuedo-likelihood, and the W_d recomputed for each of the 999 simulated datasets; and (4) the p-value for the simulated Wald is computed as one plus the number of W_d's for the simulated datasets that were greater than or equal to the W_d for the original dataset divided by 1000. The original sample design characteristics such as the clustering and sample weights are carried with the observations for the recomputations in step (3).

5. Simulations for Complex Samples

Two limited simulations studies were conducted to investigate the type I error of the goodness-of-fit tests for complex samples. In the first simulation study, the HL, Wald and simulated Wald tests were studied for cluster samples without sample weighting. This simulation study consisted of 10,000 simulations in which 2,000 observations were generated for each simulation as 100 clusters containing 20 observations. Each observation in a cluster consisted of a binary outcome variable y and a single covariate x. The y's within a cluster were generated as independent Bernoulli's with Pr(y=1) = p where the p was generated once for each cluster from a uniform on [0, 1] The covariate x_{ij}, for observation j from cluster i, were generated as a sum of two independent normals $z_i + e_{ij}$ where the z_i is distributed as a N(0,1) and e_{ij} is distributed as a N(0,2). According to this data generation scheme, the y's are not related to the x's and the y's and x's have an intracluster correlation of 1/2 and 1/3, respectively. The logistic regression model

$$\text{logit}\,[\Pr(y = 1 \mid x)] = \beta_0 + \beta_1 x$$

was used to fit the data from each simulation and the HL, Wald and simulated Wald test were computed for each simulation. In this simulation study, we are under the null hypothesis for testing goodness-of-fit because the logistic regression model used to generate the data was a model with only an intercept which is contained in the model used to fit the data. Thus, we were able to estimate the type I errors for the three test procedures (Table 2).

Table 2. Type I Error of Goodness-of-Fit Tests for Cluster Samples

Test	Nominal Type I Error	
	.05	.10
Hosmer-Lemeshow	.21	.30
Wald	.096	.17
Simulated Wald	.053	.10

The HL and Wald tests have inflated type I errors while the type I error for the simulated Wald is close to nominal.

The second simulation was used to study the effect of sample weighting on the Wald and simulated Wald tests. (The HL test was not considered because it could not be modified to incorporate the sample weights). The simulation study consisted of 10,000 simulations where in each simulation 2,000 independent observations was generated as follows: Each observation consisted of a vector of three covariates $\mathbf{x} = (x_1, x_2, x_3)'$, that were generated from independent N(0,1)'s, and a 0-1 binary outcome y that was generated as a Bernoulli variable with logit $[\Pr(y = 1 \mid \mathbf{x})] = -1.4 + x_1$. A sample weight of either 1 or 10 was randomly generated for each observation. This sample weighting was non-informative because the sample weights were not related to either the y or x. The logistic regression model (2) was used to fit the simulated datasets and the Wald and simulated Wald tests were computed for the estimated models. In this simulation study, we are under the null hypothesis for testing goodness-of-fit because the logistic regression model used to generate the data was contained in the model used to fit the data

and that the sample weights are non-informative. Thus, we were able to estimate the type I errors for the Wald and simulated Wald tests (Table 3).

Table 3. Type I Error of Goodness-of-Fit Tests for Samples with Unequal Sample Weights

Test	Nominal Type I Error	
	.05	.10
Wald	.11	.18
Simulated Wald	.049	.095

The Wald test has an inflated type I error while the type I error for the simulated Wald test is nearly nominal.

6. Example from the National Health Interview Survey

Data on 16,008 individuals with interviews from the Cancer Control and Epidemiology Supplements to the 1987 National Health Interview Survey (NHIS) were used to fit a logistic regression model to predict snuff use from a set of demographic covariates. The NHIS is a multistage stratified cluster sample of the non institutionalized population of the US in which 198 primary sampling units (PSU) were selected from 125 strata at the first stage of sampling. Further stages of sampling within the PSU's were conducted until a sample of households was selected. One sampled adult per sample household was randomly given either the Cancer Control or Epidemiology Supplement. The sample weights, which reflected differential sampling rates and adjustments for nonresponse and postratification, ranged from 252 to 36841 with a cv of 52.6. There were 495 snuff users reported in the survey. After stepwise logistic regression which utilized the sample design, we settled on a model with main effect terms for age, age^2, income, education, race, occupation, region, population size, and marital status, and interactions of age by education and region by population size. The observed and expected sample weighted counts (the sample weights were adjusted to add to the sample size) are given in the Table 4.

The goodness-of-fit was tested using the simulated Wald test. The linearization method was used to estimate the covariance matrices $\mathbf{S_d}$. The p-value for the simulated Wald was .30 suggesting that the goodness-of-fit of the logistic regression model was acceptable. As an indication of the power of the simulated Wald, we also fit a logistic regression model with only the main effects, and found a p-value of .074. This suggests that the fit of this logistic regression model may not be as good as the other one.

Table 4. Weighted Counts of Observed and Expected Numbers of Snuff Users from a Logistic Regression Analysis of the 1987 NHIS

Deciles of Risk	Snuff Users	
	Observed	Expected
1	1.7	2.3
2	7.0	7.3
3	13.8	12.5
4	17.7	18.4
5	20.0	25.7
6	33.4	35.0
7	40.9	48.2
8	76.4	66.6
9	100.1	96.1
10	186.7	185.5
Total	495	495

7. Discussion

In our simulation studies, we have shown that the Hosmer-Lemeshow test for goodness-of-fit for logistic regression models can be inappropriate in complex samples. It can have inflated type I error in cluster samples. The simulated Wald test was shown to be better than the Wald test at maintaining the nominal level for both simple random samples and complex samples. Further research is needed to study the simulated Wald test under other complex sampling designs such as when the sample weights are informative. Also, we would like to study the power of the simulated Wald procedure.

Another way to use the Wald test and simulated Wald test to examine the goodness-of-fit when there are categorical covariates such as race. In this case, it is natural to group the observations within the levels of the categorical variable and compare the number of observed outcomes to the number expected within these levels. This may have more appeal than using deciles-of-risk because by using this approach we can identify meaningful groups of the population where the model might not fit well.

REFERENCES

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

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

Korn, E. L. and Simon, R. (1991). Explained residual variation, explained risk and goodness-of-fit. American Statistician, 45, 201-206.

Skinner, C. J., Holt, D. and Smith, T. M. F. (eds.) (1989). Analysis of Complex Surveys. New York: John Wiley.

Truett, J., Cornfield, J. and Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. Journal of Chronic Diseases, 20, 511-524.