Hosmer-Lemeshow goodness of fit test for Survey data

Babubhai V. Shah, Safal Institute, and Beth G. Barnwell, Research Triangle Institute

Babubhai Shah, Safal Institute, 22 Autumn Woods Drive, Durham, NC,

Keywords: Hosmer-Lemeshow test; Goodness of fit test ; Sample surveys.

1. Introduction

The Hosmer-Lemeshow goodness of fit test is well known when data are obtained from a simple random survey. The procedure involves grouping of the observations based on the expected probabilities and then testing the hypothesis that the difference between observed and expected events is simultaneously zero for all the groups. We consider the weighted analog of the hypothesis and propose a test that accounts for the sample design. Some simulation results are also presented.

2. Test for simple random sample

Most of the tests for goodness of fit of a model are carried out by analyzing residuals, however, such an approach is not feasible for a binary outcome variable. Hosmer and Lemeshow (1989) proposed a statistic that they show, through simulation, is distributed as chi-square when there is no replication in any of the subpopulations. This test is only available for binary response models.

First, the observations are sorted in increasing order of their estimated event probability. The observations are then divided into G groups. The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the 2×G table of observed and expected frequencies, for the G groups. The statistic for the case of a simple random sample is defined as

$$H_L = \sum_{g=1}^{G} \frac{(O_g - N_g\bar{\pi}_g)^2}{N_g\bar{\pi}_g(1 - \bar{\pi}_g)} \qquad (1)$$

where N_g is the total frequency of subjects in the g-th group, O_g is the total frequency of event outcomes in the g-th group, and $\bar{\pi}_g$ is the average estimated probability of an event outcome for the g-th group. The distribution of the statistic H_L is approximated by a chi-square with (G-2) degrees of freedom.

3. Test for complex survey data

The chi-square test proposed by Hosmer-Lemeshow is equivalent to testing the hypothesis that the observed number of events in each of the groups is equal to the expected number of events based on the fitted model. This equivalent to testing the hypothesis that all statistics in the vector $\hat{\theta}$ are all zero, where

$$\hat{\theta} = (O_1 - E_1, O_2 - E_2, ..., O_G - E_G),$$

and the estimates O_g and N_g are the weighted estimates:

$$O_g = \Sigma w_i \delta_{gi} y_i \ ; \ E_g = \Sigma w_i \delta_{gi} \hat{y}_i \qquad (2)$$

We propose that the statistic equivalent to the Hosmer Lemeshow test for complex survey data is an F test with numerator degrees of freedom equal to (G-2) and denominator degrees of freedom equal to (Number of primary sampling units “PSUs” - number of strata).

$$F = \hat{\theta}' \, [\hat{V}(\hat{\theta})]^{-1} \, \hat{\theta} \, / \, (G - 2) \, . \qquad (3)$$

The variance covariance matrix $\hat{V}$ of the vector $\hat{\theta}$ is obtained by using the Taylor deviation method. The F-statistic defined in Equation (3) is the complex sample survey equivalent to the Hosmer-Lemeshow test of Equation (2).

4. Taylor deviations

The g-th element of the vector $\hat{\theta}$ is

$$\hat{\theta}_g = \Sigma w_i \delta_{gi}(y_i - \hat{y}_i), \tag{4}$$

where the summation is over all observations, w_i, y_i, and $\hat{y}_i$ are weight, observed response and expected probability respectively, and δ_{ji} is equal to 1 if the i-th observation belongs to j-th group and is 0 otherwise. We compute the taylor deviation of each element of the vector $\hat{\theta}$ by applying the method described in Shah(2002):

$$\Delta_r \hat{\theta}_g = w_r \frac{\partial}{\partial w_r} \hat{\theta}_g = w_r \left[\delta_{gr} y_r - \Sigma w_i \delta_{gi} \frac{\partial}{\partial w_r} \hat{y}_i \right]. \tag{5}$$

The detailed algebra for the Taylor deviation of $(\partial \hat{y}_i / \partial w_r)$ is presented in the appendix.

5. Simulation Results.

It is not possible to evaluate the methods analytically, so we have used simulation. The data were derived from large national survey with 48 strata with four PSUs in each stratum. Three independent variables were selected from a large national survey. For each observation, the value for the binary dependent variable was randomly generated with probability based on the logistic model:

$$E(y_i) = p_i = \frac{\exp[f(x_i)]}{1 + \exp[f(x_i)]}$$

where the linear function f was:

$$f(X_i) = -0.31 - 0.5 x_{i1} + 0.66 x_{i2} + 0.066 x_{i3}$$

For the generated dependent variable, the logistic model is known to be a good fit, that is, the null hypothesis is true. Hence, the percentiles of the computed P=values for the test of goodness of fit should be close to the percentile values. Since, two of the dependent variables had only a few distinct vales, they may be treated as categorical. We fitted the model two ways:

- By treating two of the independent variables as categorical in the first model
- By treating all independent variables as continuous in the second model

We drew one hundred thousand samples as simple random samples, and applied the methods for a simple random sample. The results for both models are presented in Tables I and II.

We also selected one hundred thousand samples, after selecting two PSU's from each stratum with probability proportional to size, and then selected a varying number of units with equal probability within a PSU. The results for these samples are presented in Tables III and IV..

For each of the generated samples, we computed a P-value by the each of the methods and the rank of the model. The table presents the percentile for the P-values. We also computed P-values using Wald F and the Satterthwaite adjusted F statistic for the stratified clustered samples (Table III and IV).

It should be noted that the Wald F and Satterthwaite adjusted F are identical for the case of a simple random sample and hence only one of them is presented in Tables I and II..

Table I. Percentiles for P-values of 100,000 simple random samples with two categorical independent variables

Percentile	HL Original	HL Taylor
Test, DF	Chi square, 8	Wald F-test, 9
1	0.0166	0.0095
5	0.0782	0.0491
10	0.1471	0.09937
20	0.2697	0.1995
30	0.3822	0.2987
40	0.4867	0.3992
50	0.5856	0.4983
60	0.6783	0.5978
70	0.7670	0.6987
80	0.8508	0.7996
90	0.9290	0.8999

Table II. Percentiles for P-values of 100,000 simple random samples with no categorical independent variables

Percentile	HL Original	HL Taylor
Test, DF	Chi square, 8	Wald F-test, 9
1	0.0156	0.0103
5	0.0677	0.0514
10	0.1278	0.1036
20	0.2389	0.2055
30	0.3458	0.3091
40	0.4466	0.4104
50	0.5442	0.5112
60	0.6406	0.6102
70	0.7337	0.7084
80	0.8240	0.8064
90	0.9132	0.9039

Table III. Percentiles for P-values of 100,000 stratified clustered samples with two categorical independent variables

Percentile	HL Original	HL Taylor	HL Taylor
Test, DF	Chi square, 8	Wald F-test, 9	Satterthwaite F
1	0.0019	0.0000	0.0350
5	0.0152	0.0028	0.0845
10	0.03512	0.0097	0.1322
20	0.08630	0.0317	0.2051
30	0.14900	0.0680	0.2794
40	0.22610	0.1167	0.3465
50	0.31340	0.1749	0.4203
60	0.41580	0.2609	0.4955
70	0.53400	0.3635	0.5761
80	0.66070	0.4955	0.6704
90	0.80960	0.6812	0.7689

6. Conclusions.

From Table I, For the case of the model with two categorical variables and simple random samples, results obtained by the method based on Taylor deviations is better than those based on the original Hosmer Lemeshow method. The results in Table II for the model with all continuous variables are similar.

For the case of a stratified clustered sample with unequal probabilities, the tests based on Wald F and. Satterthwaite adjusted F statistics seem to provide lower and upper bounds for the “true” confidence level. The Homer Lemeshow produces results that are poor in the tail of the distribution, which is critical for a test of hypothesis.

The results are preliminary, because they are based on one data set, and only two models. Further simulations are needed to confirm the finding that Taylor linearization based tests are appropriate for a variety of sample designs and different models.

References

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys," International Statistical Review, 51, 279-292.

Hosmer, D. W. and Lemeshow S. (1989). Applied Logistic Regression. New York: John Wiley & Sons, Inc.

Shah, B. V. (2002) “Calculus of Taylor Deviations” paper presented at the Joint Statistical Meetings.

Table IV. Percentiles for P-values of 100,000 stratified clustered samples with no categorical independent variables

Percentile	HL Original	HL Taylor	HL Taylor
Test, DF	Chi square, 8	Wald F-test, 9	Satterthwaite F
1	0.0009	0.0002	0.0334
5	0.0084	0.0037	0.0895
5	0.0084	0.0037	0.0895
10	0.0223	0.0106	0.1373
20	0.0615	0.0347	0.2139
30	0.1127	0.0698	0.2877
40	0.1792	0.1170	0.3581
50	0.2612	0.1824	0.4280
60	0.3552	0.2656	0.5017
70	0.4693	0.3655	0.5819
80	0.5963	0.5017	0.6643
90	0.76270	0.68680	0.77900

Appendix: Taylor deviations for Logistic Regression

For logistic regression, the assumptions are:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right),\ g^{-}(L) = \frac{\exp(L)}{1+\exp(L)},$$

and

$$V(\mu) = \mu(1-\mu),$$

Hence

$$E(y_{hijk}) = \mu_{hijk} = g^{-}(L_{hijk}) = \frac{\exp(\boldsymbol{x}'_{hijk}\beta)}{1+\exp(\boldsymbol{x}'_{hijk}\beta)},$$

and

$$\frac{\partial g^{-}(L)}{\partial L} = \mu(1-\mu).$$

The corresponding score functions for the parameter β_r are:

$$S(\beta) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} w_{hijk}\,(y_{hijk} - \mu_{hijk})\,\boldsymbol{x}_{hijk} \tag{6}$$

The matrix $\boldsymbol{J_0}$ for Binder (1982) method is

$$\boldsymbol{J_0} = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{J_{hi}}\sum_{k=1}^{K_{hij}} w_{hijk}\mu_{hijk}(1-\mu_{hijk})\boldsymbol{x}_{hijk}\boldsymbol{x}'_{hijk} \tag{7}$$

As shown by Shah(2002), the Taylor deviation of the estimate $\hat{\beta}$ is

$$\Delta_{rstu}(\hat{\beta}) = w_{rstu}\frac{\partial}{\partial w_{rstu}}\hat{\beta} = w_{rstu}[\boldsymbol{J_0}]^{-1}(y_{rstu} - \mu_{rstu})\boldsymbol{x}_{rstu} \tag{8}$$

The estimated Value of $\hat{y}$ is:

$$\hat{y}_{hijk} = \exp(\boldsymbol{x}'_{hijk}\hat{\beta}) \,/\, [1+\exp(\boldsymbol{x}'_{hijk}\hat{\beta})], \tag{9}$$

The Taylorized deviation for $\hat{y}_{hijk}$ with respect to the observation (rtsu) is

$$\Delta_{rstu}(\hat{y}_{hijk}) = \frac{\partial}{\partial w_{rstu}}\hat{y}_{hijk} = \frac{\partial}{\partial w_{rstu}}\left[\frac{\exp(\boldsymbol{x}'_{hijk}\hat{\beta})}{1+\exp(\boldsymbol{x}'_{hijk}\hat{\beta})}\right], \tag{10}$$

$$\Delta_{rstu}(\hat{y}_{hijk}) = w_{rstu}\hat{y}_{hijk}(1-\hat{y}_{hijk})\boldsymbol{x}'_{hijk}\frac{\partial}{\partial w_{rstu}}\hat{\beta}. \tag{11}$$

On substituting the partial derivative of beta from Equation (8), in Equation (11) the result is:

$$\Delta_{rstu}(\hat{y}_{hijk}) = w_{rstu}\,\hat{y}_{hijk}\,(1 - \hat{y}_{hijk})\,(y_{rstu} - \hat{y}_{rstu})\,\boldsymbol{x}'_{hijk}\,[\boldsymbol{J_0}]^{-1}\,\boldsymbol{x}_{rstu} \tag{12}$$

Equation (12) provides the Taylor deviation needed for calculation of Taylor deviations of θ for computing variance covariance matrix required in Equation (3).