The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2006)
6, Number 1, pp. 97–105

Goodness-of-fit test for a logistic regression model fitted using survey sample data

Kellie J. Archer
Department of Biostatistics
Virginia Commonwealth University
Richmond, VA
kjarcher@vcu.edu

Stanley Lemeshow
School of Public Health
Ohio State University
Columbus, OH

Abstract. After a logistic regression model has been fitted, a global test of goodness of fit of the resulting model should be performed. A test that is commonly used to assess model fit is the Hosmer–Lemeshow test, which is available in Stata and most other statistical software programs. However, it is often of interest to fit a logistic regression model to sample survey data, such as data from the National Health Interview Survey or the National Health and Nutrition Examination Survey. Unfortunately, for such situations no goodness-of-fit testing procedures have been developed or implemented in available software. To address this problem, a Stata ado-command, `svylogitgof`, for estimating the F-adjusted mean residual test after `svy: logit` or `svy: logistic` estimation has been developed, and this paper describes its implementation.

Keywords: st0099, svylogitgof, goodness of fit, survey design, svy, logistic regression, logit

1 Introduction

Once a logistic regression model has been fitted to a given set of data, the adequacy of the model is examined by overall goodness-of-fit tests, area under the receiver operating characteristic curve, and examination of influential observations. The purpose of any overall goodness-of-fit test is to determine whether the fitted model adequately describes the observed outcome experience in the data (Hosmer and Lemeshow 2000). One concludes that a model fits if the differences between the observed and fitted values are small and if there is no systematic contribution of the differences to the error structure of the model. Goodness-of-fit tests are usually general tests that assess the fitted model's overall departure from the observed data.

Appropriate estimation methods that take into account the survey sampling design are available in Stata by specifying the sampling design using `svyset` followed by estimation using the `svy:` command syntax. However, the `estat gof, table group(10)` command ordinarily used for estimating the Hosmer–Lemeshow goodness-of-fit test statistic associated with a fitted logistic regression model is not available after `svy` estimation. Because of the lack of goodness-of-fit methods available after survey estimation, it has been suggested that goodness of fit be examined by first fitting the "design-based" model (i.e., one that takes the survey design structure into account),

then estimating the corresponding probabilities, and subsequently using independently and identically distributed (i.i.d.)–based tests and applying any findings to the design-based model (Hosmer and Lemeshow 2000). The statistical properties of this procedure and an alternative goodness-of-fit test for logistic regression when modeling data collected using sample survey data have been studied previously (Archer 2001). Unlike ordinary goodness-of-fit tests, this alternative test takes into account the sampling weights and design.

2 Traditional goodness-of-fit tests

Logistic regression is used to model the relationship between a categorical outcome variable, which is usually dichotomous, such as disease being present versus absent, and a set of predictor variables. Traditionally, logistic regression assumes that the observations are a random sample from a population (i.e., i.i.d.), where the model is expressed as $y_i = \pi(\mathbf{x}_i) + \varepsilon_i$. In this equation, y_i represents the dichotomous dependent or outcome variable; $\pi(\mathbf{x}_i)$ represents the conditional probability of experiencing the event given independent predictor variables, $\mathbf{x}_i$, or $\Pr(Y_i = 1|\mathbf{x}_i)$; and ε_i represents the binomial random error term. More formally, the conditional probability, $\pi(\mathbf{x}_i)$, as a function of the independent covariates, $\mathbf{x}_i$, is expressed as

$$\pi(\mathbf{x}_i) = \Pr(Y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ are the model parameters to be estimated and $p + 1$ is the number of independent terms in the model.

Pearson's chi-squared test is one such goodness-of-fit test that examines the sum of the squared differences between the observed and expected number of cases per covariate pattern divided by its standard error. In traditional logistic regression where n observations are independently sampled (i.e., there are no clusters), a covariate pattern is defined to be a unique set of the $\mathbf{x}_i$'s, where $i = 1, \ldots, n$, and m_k will represent the number of subjects with the same covariate pattern where $k = 1, \ldots, K$. Therefore, K represents the number of unique covariate patterns. Let $\widehat{\pi}(\mathbf{x}_i)$ (or expressed as $\widehat{\pi}_i$ for the ith subject) be the estimated probabilities, which are the same for all m_k subjects in the same covariate pattern. Likewise, y_i represents the outcome for the ith subject, and y_k represents the sum of the observed outcomes in the kth covariate pattern. Then Pearson's chi-squared goodness-of-fit test for logistic regression is expressed as the sum of the squared Pearson's residuals,

$$X^2 = \sum_{k=1}^{K} \frac{(y_k - m_k\widehat{\pi}_k)}{m_k\widehat{\pi}_k(1 - \widehat{\pi}_k)}$$

This test statistic is distributed approximately as χ^2 with $K-(p+1)$ degrees of freedom when $m_k\widehat{\pi}_k$ is large for every k, where K is the number of covariate patterns and p is the number of independent covariates in the model. Obviously, once some continuous variables are incorporated into a logistic regression model, the number of distinct covariate

patterns can be approximately n, the total sample size. Hence, Pearson's chi-squared test is not effective since $m_k\widehat{\pi}_k$ may be small for every k when $K \sim n$.

To avoid problems associated with the asymptotic distribution of the chi-squared test when $K \sim n$, Hosmer and Lemeshow (1980) developed a set of formal goodness-of-fit tests whereby subjects are grouped into g groups and a chi-squared test is then estimated using the amalgamated cells. The properties of this test statistic have been studied using extensive simulations under i.i.d. data assumptions (Hosmer and Lemeshow 1980). Specifically, the Hosmer–Lemeshow goodness-of-fit test statistic is estimated by grouping observations into "deciles of risk", in which observations are partitioned into $g = 10$ equal-sized groups based on their ordered estimated probabilities, $\widehat{\pi}_i$. A chi-squared test is then calculated using these deciles of risk as follows. The observed number of cases in the dth decile is given by $O_{1d} = \sum_{i \in \text{decile}_d} y_i$, whereas the observed number of noncases in the dth decile is given by $O_{0d} = \sum_{i \in \text{decile}_d} (1 - y_i)$. The expected number of cases in the dth decile is given by $E_{1d} = \sum_{i \in \text{decile}_d} \widehat{\pi}_i$, whereas the expected number of noncases in the dth decile is given by $E_{0d} = \sum_{i \in \text{decile}_d} (1 - \widehat{\pi}_i)$. The Hosmer–Lemeshow test is then calculated as $\widehat{C}_g = \sum_{h=0}^{1} \sum_{d=1}^{g} \{(O_{hd} - E_{hd})^2 / E_{hd}\}$. This quantity is distributed as χ^2_{g-2} when $K = n$ (Hosmer and Lemeshow 1980).

When fitting logistic regression models using survey data, the sampling weight, w_{ji}, calculated as the inverse of the product of the conditional inclusion probabilities at each stage of sampling, represents the number of units that the given sampled observation represents in the total population. Expanding each observation by its sampling weight produces a dataset for the N units in the total population. Hence, a logistic regression model fitted using sampling weights is essentially a fit to the "census" data. Therefore, the number of observed and expected cell counts for the Hosmer–Lemeshow goodness-of-fit test will total the population size.

Pearson's chi-squared test is heavily influenced by the sample size, even when the relationship between two variables is preserved. For example, a chi-squared test of homogeneity can be used for examining the relationship between age category and birthweight (categorized as low versus normal). For the dataset consisting of $N = 189$ observations displayed below, there is no apparent relationship (Pr = .309).

```
. use smallset
. tab birthwgt agecat, chi2
```

birthwgt	agecat 0	1	Total
0	115	15	130
1	49	10	59
Total	164	25	189

Pearson chi2(1) = 1.0351 Pr = 0.309

However, suppose that this dataset is inflated by a factor of 5. That is, the column percentages are identical but the total number of observations is now $N = 945$. The chi-squared test of homogeneity now suggests that there is a high degree of association between age and birthweight (Pr = .023).

```
. use largeset
. tab  birthwgt agecat, chi2
```

birthwgt	agecat 0	1	Total
0	575	75	650
1	245	50	295
Total	820	125	945

Pearson chi2(1) = 5.1755 Pr = 0.023

Survey samples often include many sampled subjects. Use of the sampling weights in estimating the chi-squared statistic would further inflate the cells representing the observed and expected number of observations. Since the chi-squared test is heavily influenced by the sample size, the properties of an alternative goodness-of-fit test for survey sample data were studied, which was found to have a Type I error rate close to the nominal rate, especially as the number of sampled clusters increased (Archer 2001). Also the test was found to have reasonable power.

3 Goodness of fit for survey samples

Under i.i.d.-based sampling, elements are selected independently; therefore, the covariance between elements is zero. Under complex sampling, there may be several primary sampling units (PSUs); that is, there are $j = 1, \ldots, M$ PSUs (or "clusters") from which m PSUs are sampled. Furthermore, within each sampled PSU, there are $i = 1, \ldots, N_j$ units from which n_m observations are sampled. A disadvantage generally associated with cluster sampling is that elements from the same cluster are often more homogeneous than elements from different clusters. This state results in a positive covariance between elements within a cluster. Therefore, the intraclass correlation, which measures the homogeneity within clusters, is generally positive for cluster sample designs, and as a result, traditional maximum likelihood methods for estimation cannot be used. Rather, under complex sampling, pseudomaximum likelihood is used (Skinner, Holt, and Smith 1989). The sampling weight, w_{ji}, calculated as the inverse of the product of the conditional inclusion probabilities at each stage of sampling, represents the number of units that the given sampled observation represents in the total population. Expanding each observation by its sampling weight will produce a dataset for the N units in the total population. Conceptually, pseudomaximum likelihood estimation is like obtaining the maximum likelihood estimates for the expanded dataset: the logistic regression model is being fitted to the "census" data. The model parameters, $\boldsymbol{\beta}$, for logistic regression models built from complex survey data are found by using pseudomaximum likelihood. The contribution of a single observation using pseudomaximum likelihood is

$$\pi(\mathbf{x}_{ji})^{w_{ji}\times y_{ji}}\{1-\pi(\mathbf{x}_{ji})\}^{w_{ji}\times(1-y_{ji})}$$

The pseudomaximum likelihood function is still constructed as the product of the individual contributions to the likelihood, but now it is the product over the m clusters

sampled and n_m observations within the given cluster, expressed as

$$l_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \pi(\mathbf{x}_{ji})^{w_{ji} \times y_{ji}} \{1 - \pi(\mathbf{x}_{ji})\}^{w_{ji} \times (1 - y_{ji})} \tag{2}$$

Given the pseudolikelihood in (2), we find that the PMLE (pseudomaximum likelihood estimator) is the value that maximizes the pseudolog-likelihood function.

The survey sampling design may induce correlation among observations, particularly when cluster samples are drawn. To appropriately estimate standard errors associated with model parameters and estimated odds ratios, one must account for the sampling design. Likewise, the proposed goodness-of-fit test, called the F-adjusted mean residual test, is estimated as follows. First, after the logistic regression model is fitted, the residuals, $\widehat{r}_{ji} = y_{ji} - \widehat{\pi}(\mathbf{x}_{ji})$, are obtained. The goodness-of-fit test is based on the residuals since large departures between observed and predicted values, taking variability into account, would seemingly indicate lack of fit. Then using a previously proposed grouping strategy (Graubard, Korn, and Midthune 1997), observations are sorted into deciles based on their estimated probabilities, and each decile of risk includes approximately equivalent total sampling weights. That is, survey estimates of the mean residuals by decile of risk, $\widehat{M} = \left(\widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_{10}\right)$, are obtained such that $\widehat{M}_1 = \sum_j \sum_i w_{ji}\widehat{r}_{ji} / \sum_j \sum_i w_{ji}$ for the smallest 10% of the $\widehat{r}_{ji}$ values, $\widehat{M}_2 = \sum_j \sum_i w_{ji}\widehat{r}_{ji} / \sum_j \sum_i w_{ji}$ for the next smallest 10% of the $\widehat{r}_{ji}$ values, ..., and $\widehat{M}_{10} = \sum_j \sum_i w_{ji}\widehat{r}_{ji} / \sum_j \sum_i w_{ji}$ for the largest 10% of the $\widehat{r}_{ji}$ values. Here w_{ji} represents the sampling weights associated with the ordered residuals grouped into the indicated decile of risk. The associated estimated variance–covariance matrix, $\widehat{V}(\widehat{M})$, is obtained using linearization, which is based on a first-order Taylor series approximation. The Wald test statistic for testing the g categories is then $\widehat{W} = \widehat{M}^T \left\{\widehat{V}(\widehat{M})^{-1}_{g \times g}\right\} \widehat{M}$. However, the chi-squared distribution has been found to not be an appropriate reference distribution, so the F-corrected Wald statistic (Thomas and Rao 1987), which is $F = (f - g + 2)/(fg)W$ is approximately F-distributed with $g-1$ numerator degrees of freedom and $f-g+2$ denominator degrees of freedom, where f is the number of sampled clusters minus the number of strata and g is the number of categories included in the hypothesis test (here, $g = 10$ corresponding to deciles of risk). Therefore, the goodness-of-fit test implemented in `svylogitgof` is of the form $\widehat{Q}_M = (f - g + 2)/(fg)\widehat{M}^t \widehat{V}(\widehat{M})^{-1} \widehat{M}$. The properties of this test have been studied using extensive simulations and are reported elsewhere (Archer 2001).

The implementation of this test statistic is as follows: after a logistic regression model has been fitted using either `svy: logit` (see [SVY] **svy: logit**) or `svy: logistic` (see [SVY] **svy: logistic**), the command `svylogitgof` is issued for estimating goodness of fit.

4 NHIS illustration

The 2004 National Health Interview Survey (NHIS) sample adult core survey collected disease, health status, health behavior, health care utilization, demographic, and AIDS-related data on one sampled adult within each interviewed household (National Center for Health Statistics 1999). This release includes 31,326 observations from a multistage sampling design. For the example in this paper, we fitted a logistic regression model predicting hypertension (`hypev`), where the independent covariates that were examined for incorporation into the final model included gender (`sex`), race (`racereci2`), age (`agep`), having ever smoked 100 cigarettes (`smkev`), frequency of vigorous activity (`vigno`), and having ever had 12 or more drinks in any one year (`alc1yr`).

The following rules were followed in recoding the variables in the NHIS adult core dataset. For `hypev`, `smkev`, `alc1yr`, and `vigno`, responses that were coded as either {7 or 997} = "Refused", {8 or 998} = "Not ascertained", or {9 or 999} = "Don't know" were changed to missing values. The `racreci2` included the categories "White" and "Black" with "All other race groups" serving as the referent. Frequency of vigorous activity was recoded as "Never" and "Some (1–500 units)", with "Unable to do this type of activity" as the referent. Fractional polynomials were used to assess the most appropriate form of the continuous covariate `age`. The best-fitting second-order model ($\texttt{age}^2 + \texttt{age}^3$) was significantly better than any other simpler model; however, it was not substantially different from the more readily interpretable full polynomial $\texttt{age} + \texttt{age}^2 + \texttt{age}^3$. Therefore, the full polynomial was carried forward in all subsequent models. Prior to fitting the logistic regression model, we redistributed the weights of observations with missing values for any of the dependent or independent covariates to the remaining observations equally.

For this dataset, we fitted the logistic regression model ignoring the survey sampling design and then estimated the Hosmer–Lemeshow goodness-of-fit test.

```
. use nhis2004
. xi: logistic hypev age_p age2 age3 smkev alc1yr sex i.vigno i.racreci2, or
i.vigno           _Ivigno_1-3         (_Ivigno_1 for vigno==-1 omitted)
i.racreci2        _Iracreci2_1-3      (naturally coded; _Iracreci2_1 omitted)
Logistic regression                               Number of obs   =      30477
                                                  LR chi2(10)     =    6261.54
                                                  Prob > chi2     =     0.0000
Log likelihood = -14815.969                       Pseudo R2       =     0.1744
```

hypev	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age_p	.9674684	.0215879	-1.48	0.138	.9260688	1.010719
age2	1.002591	.0004399	5.90	0.000	1.001729	1.003453
age3	.9999797	2.72e-06	-7.46	0.000	.9999744	.999985
smkev	1.1304	.0342743	4.04	0.000	1.065181	1.199613
alc1yr	.9102218	.0285273	-3.00	0.003	.855992	.9678873
sex	.9839784	.0295462	-0.54	0.591	.92774	1.043626
_Ivigno_2	.527136	.0395233	-8.54	0.000	.4550947	.6105814
_Ivigno_3	.4191798	.0329882	-11.05	0.000	.3592637	.4890883
_Iracreci2_2	1.82244	.0733332	14.92	0.000	1.684231	1.97199
_Iracreci2_3	.9658363	.075581	-0.44	0.657	.8285012	1.125936

```
. estat gof, table group(10)
```

Logistic model for hypev, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0555	128	138.0	2920	2910.0	3048
2	0.0816	240	209.1	2818	2848.9	3058
3	0.1144	315	296.3	2730	2748.7	3045
4	0.1590	381	414.9	2666	2632.1	3047
5	0.2194	585	575.1	2476	2485.9	3061
6	0.2952	762	781.9	2290	2270.1	3052
7	0.3927	998	1045.0	2051	2004.0	3049
8	0.5026	1391	1358.9	1631	1663.1	3022
9	0.5736	1682	1659.1	1375	1397.9	3057
10	0.8408	1921	1924.7	1117	1113.3	3038

```
     number of observations =     30477
           number of groups =        10
Hosmer-Lemeshow chi2(8) =        16.38
            Prob > chi2 =         0.0372
```

The test suggests that the model is not a good fit even though the observed and expected cell frequencies are generally in good agreement. However, after fitting the logistic regression model taking the survey sampling design into account, the F-adjusted mean residual goodness-of-fit test was applied and suggested no evidence of lack of fit.

```
. svyset psu [pweight=new_wgt], strata(stratum)
      pweight: new_wgt
          VCE: linearized
     Strata 1: stratum
         SU 1: psu
        FPC 1: <zero>
. xi: svy: logistic hypev age_p age2 age3 smkev alc1yr sex i.vigno i.racreci2, or
i.vigno           _Ivigno_1-3         (_Ivigno_1 for vigno==-1 omitted)
i.racreci2        _Iracreci2_1-3      (naturally coded; _Iracreci2_1 omitted)
(running logistic on estimation sample)
Survey: Logistic regression
Number of strata   =       339                 Number of obs      =     30477
Number of PSUs     =       678                 Population size    = 2.152e+08
                                               Design df          =       339
                                               F(  10,    330)    =    346.84
                                               Prob > F           =    0.0000
```

(*Continued on next page*)

hypev	Odds Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_p	.9784137	.0263086	-0.81	0.418	.9280098	1.031555
age2	1.00239	.0005309	4.51	0.000	1.001346	1.003435
age3	.9999808	3.28e-06	-5.84	0.000	.9999743	.9999873
smkev	1.155569	.0399256	4.18	0.000	1.079645	1.236832
alc1yr	.8873859	.0343981	-3.08	0.002	.8222404	.9576928
sex	.924363	.0325434	-2.23	0.026	.8625167	.9906439
_Ivigno_2	.5686226	.0507514	-6.33	0.000	.4770671	.6777488
_Ivigno_3	.450939	.042039	-8.54	0.000	.3753875	.5416963
_Iracreci2_2	1.798082	.0825642	12.78	0.000	1.642798	1.968045
_Iracreci2_3	.8990461	.0792883	-1.21	0.228	.755865	1.06935

```
. svylogitgof
F-adjusted test statistic = 1.1811056
p-value                   = .30612561
```

5 Conclusion

The test statistic proposed and now available in Stata provides a method for investigators to assess model fit after fitting a logistic regression model taking survey design into account.

6 References

Archer, K. J. 2001. Goodness-of-fit tests for logisitic regression models developed using data collected from a complex sampling design. Ph.D. thesis, Ohio State University.

Graubard, B. I., E. L. Korn, and D. Midthune. 1997. Testing goodness-of-fit for logistic regression with survey data. In *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings*, 170–174. Alexandria, VA: American Statistical Association.

Hosmer, D. W., Jr., and S. Lemeshow. 1980. Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods, Part A* 9: 1043–1069.

———. 2000. *Applied Logistic Regression.* 2nd ed. New York: Wiley.

National Center for Health Statistics. 1999. *National Health Interview Survey: Research for the 1995–2004 redesign.* Hyattsville, MD: Vital and Health Statistics.

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

Thomas, D. R., and J. N. K. Rao. 1987. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association* 82: 630–636.

About the authors

Kellie J. Archer (kjarcher@vcu.edu) is an assistant professor in the Department of Biostatistics and a fellow in the Center for the Study of Biological Complexity, Virginia Commonwealth University, 1101 E Marshall St. B1-069, Richmond, VA 23298.

Stanley Lemeshow (lemeshow.1@osu.edu) is dean of the School of Public Health and a professor in the Division of Epidemiology and Biostatistics and Department of Statistics, Ohio State University, 320 W 10th Ave., M200 Starling-Loving Hall, Columbus, OH 43210.