Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4450–4464
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design

Kellie J. Archer[a,*], Stanley Lemeshow[b], David W. Hosmer[c]

[a] *Department of Biostatistics, Virginia Commonwealth University, 1101 East Marshall St., B1-066, Richmond, VA 23298-0032, USA*
[b] *Division of Epidemiology and Biostatistics and Department of Statistics, School of Public Health, The Ohio State University, 320 West Tenth Ave., M200 Starling-Loving Hall, Columbus, OH 43210, USA*
[c] *University of Massachusetts, 128 Worcester Road, Stowe, VT 05672-4320, USA*

Received 15 May 2006; received in revised form 30 June 2006; accepted 2 July 2006
Available online 28 July 2006

Abstract

Logistic regression models are frequently used in epidemiological studies for estimating associations that demographic, behavioral, and risk factor variables have on a dichotomous outcome, such as disease being present versus absent. After the coefficients in a logistic regression model have been estimated, goodness-of-fit of the resulting model should be examined, particularly if the purpose of the model is to estimate probabilities of event occurrences. While various goodness-of-fit tests have been proposed, the properties of these tests have been studied under the assumption that observations selected were independent and identically distributed. Increasingly, epidemiologists are using large-scale sample survey data when fitting logistic regression models, such as 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, goodness-of-fit tests for logistic regression models when data are collected using complex sampling designs are proposed. Properties of the proposed tests were examined using extensive simulation studies and results were compared to traditional goodness-of-fit tests. A Stata ado function `svylogitgof` for estimating the F-adjusted mean residual test after `svylogit` fit is available at the author's website http://www.people.vcu.edu/ kjarcher/Research/Data.htm.

Keywords: Goodness-of-fit; Logistic regression; Survey sampling; Design-based estimation

1. Introduction

Logistic regression is frequently used in epidemiological studies to model the relationship between a categorical outcome variable and a set of predictor variables. Traditionally, logistic regression assumes that the observations represent a random sample from a population (i.e., independent and identically distributed (iid)), where the model is expressed as

$$y_i = \pi(\mathbf{x}_i) + \varepsilon_i. \tag{1}$$

* Corresponding author. Tel.: +1 804 827 2039; fax: +1 804 828 8900.
E-mail addresses: kjarcher@vcu.edu (K.J. Archer), lemeshow.1@osu.edu (S. Lemeshow), hosmer@schoolph.umass.edu (D.W. Hosmer).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.07.006

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

$$\pi(\mathbf{x}_i) = \Pr(Y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}, \tag{2}$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ are the model parameters to be estimated and p is the number of independent terms in the model.

Under iid-based sampling, elements are selected independently; therefore, the covariance between elements is zero. Under complex sampling, there may be a number of 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 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 results in a positive covariance between elements within a cluster. Therefore, the intra-class 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, which involves both stratification and possibly several stages of cluster sampling, pseudo-maximum likelihood is used (Skinner et al., 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, pseudo-maximum likelihood estimation is like obtaining the maximum likelihood estimates for the expanded dataset. In other words, the logistic regression model is being fit to the 'census' data. The model parameters $\boldsymbol{\beta}$ for logistic regression models built from complex survey data are found by using pseudo-maximum likelihood. The contribution of a single observation using pseudo-maximum likelihood is

$$\pi(\mathbf{x}_{ji})^{w_{ji}\times y_{ji}}\left[1 - \pi(\mathbf{x}_{ji})\right]^{w_{ji}\times(1-y_{ji})}. \tag{3}$$

The pseudo-maximum 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

$$l_p(\beta) = \prod_{j=1}^{m}\prod_{i=1}^{n_j} \pi(x_{ji})^{w_{ji}\times y_{ji}}\left[1 - \pi(x_{ji})\right]^{w_{ji}\times(1-y_{ji})}. \tag{4}$$

Given the pseudo-likelihood equation we find the PMLE (pseudo-maximum likelihood estimator) is that value that maximizes the pseudo log-likelihood function

$$\ln\{L_p(\beta)\} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left[w_{ji} \times y_{ji}\right] \times \ln\left[\pi(x_{ji})\right] + \left[w_{ji} \times (1 - y_{ji})\right] \times \ln\left[1 - \pi(x_{ji})\right]. \tag{5}$$

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, it is important to account for the sampling design.

The need to account for the sampling design in the statistical analysis of survey data has been widely reported in the literature. A brief tutorial regarding the importance of accounting for clustering and sampling weights, accompanied by an illustration using the National Health and Nutrition Examination Survey I data has previously been reported (Korn and Graubard, 1991). A more comprehensive review was subsequently provided by Korn and Graubard (1995). In another example, the difference between "model-based" (assuming the observations are from a random sample) and "design-based" analyses (an analysis which accounts for the survey design) was illustrated using the Personnes Ages Quid study, a stratified cluster sample (Lemeshow et al., 1998). It is of particular importance to model the survey design when estimating standard errors associated with model parameters or odds ratios.