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.

Once a logistic regression model has been fit to a given set of data, the adequacy of the model is examined by overall goodness-of-fit tests and examination of influential observations. One concludes 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. A goodness-of-fit test that is commonly used to assess the fit of logistic regression models is the Hosmer–Lemeshow test (Hosmer and Lemeshow, 1980). Other goodness-of-fit tests for logistic regression models have been proposed (Cox, 1958; Tsiatis, 1980; Brown, 1982; Azzalini et al., 1989; le Cessie and van Houwelingen, 1991, 1995; Su and Wei, 1991; Osius and Rojek, 1992; Pigeon and Heyse 1999a,b). These goodness-of-fit tests have been studied under independent and identically distributed random variable assumptions, which we refer to as the 'iid-based' setting.

Although appropriate estimation methods which take into account the sampling design in estimating logistic regression model parameters are available in various statistical packages, there is a corresponding absence of design-based goodness-of-fit testing procedures. Due to this noted absence, it has been suggested that goodness-of-fit be examined by first fitting the design-based model, then estimating the probabilities, and subsequently using iid-based tests for goodness-of-fit and applying any findings to the design-based model (Hosmer and Lemeshow, 2000). Unfortunately, the statistical properties of this method have not been examined. In this article we studied this proposed method and additionally proposed alternative design-based goodness-of-fit tests for logistic regression models. Unlike ordinary goodness-of-fit tests, the proposed tests take the sampling design and weights into account.

2. Goodness-of-fit

Three modifications to existing goodness-of-fit tests for design-based logistic regression models were previously studied. First, Graubard et al. (1997) proposed an alternative grouping strategy for establishing deciles of risk for the Hosmer–Lemeshow goodness-of-fit test. As usual, after fitting the logistic regression model, the probabilities are estimated as $\hat{\pi}\left(\mathbf{x}_{ji}\right) = \Pr\left(Y_{ji} = 1|\mathbf{x}_{ji}\right) = \mathrm{e}^{\mathbf{x}'_{ji}\hat{\boldsymbol{\beta}}} \Big/ \left(1 + \mathrm{e}^{\mathbf{x}'_{ji}\hat{\boldsymbol{\beta}}}\right)$. When the estimated probabilities are sorted in ascending order and subsequently grouped into 10 roughly equally sized groups, we refer to the estimated probabilities as being grouped into "deciles of risk." However, in the procedure outlined by Graubard et al. (1997) the deciles of risk were not formed based strictly on the ordered probabilities. Rather, deciles of risk were constructed by first ordering the probabilities and letting $\tilde{w}_{ji}$ represent the associated sorted weights. Then, deciles are formed using the ordered probabilities so that each decile represents one tenth of the subjects in the total population. Specifically, the first decile contains the smallest $\hat{\pi}_{ji}$ values such that $\sum_j \sum_i \tilde{w}_{ji} = 0.1 \times N_{\text{Population}}$, the second decile contains the next smallest $\hat{\pi}_{ji}$ values such that $\sum_j \sum_i \tilde{w}_{ji} = 0.1 \times N_{\text{Population}}, \ldots,$ and the 10th decile contains the largest $\hat{\pi}_{ji}$ values such that $\sum_j \sum_i \tilde{w}_{ji} = 0.1 \times N_{\text{Population}}$. The test statistic is computed by comparing the weighted number of observed outcomes in the gth decile, $\hat{o}_g = \sum_j \sum_i \tilde{w}_{ji} y_{ji}$, to the weighted number of expected outcomes in the gth decile, $\hat{e}_g = \sum_j \sum_i \tilde{w}_{ji} \hat{\pi}_{ji}$. They studied the modified Hosmer–Lemeshow test without further taking the sampling weights into consideration in the test statistic. Their simulations showed that this modification had a correct Type I error rate under simple random sampling but an inflated Type I error rate when applied to data from a cluster sample (Graubard et al., 1997). Unfortunately, they did not study this modification when incorporating the sampling weights nor did they assess power.

Graubard et al. (1997) also studied the use of the Wald statistic $\left(\hat{\mathbf{O}} - \hat{\mathbf{E}}\right) \hat{\mathbf{S}}_d^{-1} \left(\hat{\mathbf{O}} - \hat{\mathbf{E}}\right)$, where $\hat{\mathbf{O}}$ is the vector of the total weighted observed outcomes by decile, $\hat{\mathbf{E}}$ is the vector of the total weighted expected outcomes by decile, and $\hat{\mathbf{S}}_d$ is the variance–covariance matrix of $\left(\hat{\mathbf{O}} - \hat{\mathbf{E}}\right)$ with design taken into account. They found that this test had an inflated Type I error rate, even under simple random sampling. In their more recent work, Korn and Graubard (1999) proposed comparing the weighted means of the observed proportions to the weighted means of the probabilities, using a Wald statistic to test the goodness-of-fit. The weighted mean of the observed proportions in the $i = 1, \ldots, g$ deciles of risk is expressed as $\bar{o}_i = \sum_{i=1}^{g} w_i y_i \Big/ \sum_{i=1}^{g} w_i$, whereas the weighted mean of the probabilities is expressed as $\bar{e}_i = \sum_{i=1}^{g} w_i \hat{\pi}_i \Big/ \sum_{i=1}^{g} w_i$. The Wald test is then conducted using, in the denominator, the standard error of the differences between the weighted means of the observed proportions and the weighted means of the probabilities. In their simulation studies the authors found that the Wald test had an inflated Type I error even under simple random sampling, and concluded that the χ^2 distribution may not be a good reference distribution (Graubard et al., 1997).

In an effort to compensate for lack of a seemingly good reference distribution for the Wald test, Graubard et al. (1997) used a simulation-based method for estimating the p-value for an observed Wald statistic $\hat{W}$. First, the probabilities from the fitted logistic regression model were estimated and used as the empirical distribution from which the binary outcomes were randomly sampled with replacement, resulting in a simulated set of responses $\mathbf{y}^*$, where each $y^* = I\left\{\hat{\pi}^* > 0.5\right\}$. Then, the Wald test statistic was estimated using $\mathbf{y}^*$, resulting in $\hat{W}^*$. The process of generating a simulated dataset and estimating the associated Wald test was repeated $R = 1000$ times. The p-value was then reported as the proportion of $\hat{W}^*$'s at least as great as $\hat{W}$, or $\left\{\#\left(\hat{W}^* \geqslant \hat{W}\right)/R\right\}$. Although the Type I error rate was accurate for the simulated Wald test, it makes the assumption that the fitted logistic regression model is correct. This approach will not be considered further at this time.

Other researchers have proposed corrections to the chi-square statistic to account for the clustering effect, however, these corrections were not implemented in the three previously proposed design-based goodness-of-fit tests. In the complex sampling setting, the Wald test should work well provided the number of sampled clusters is large since the Wald test is asymptotically correct. However, when the number of sampled clusters is small, the variance estimates of $\hat{\beta}$ may become unreliable because there are few degrees of freedom for the estimate (Thomas and Rao, 1987). This usually results in an overly liberal test statistic. The effects of the instability of the Wald test can be adjusted by using the F-corrected Wald statistic, which is

$$F = \frac{f - u + 2}{fu} W \tag{6}$$

which is approximately F-distributed with $u - 1$ numerator degrees of freedom and $f - u + 2$ denominator degrees of freedom, where f is the number of sampled clusters minus the number of strata and u is the number of slope coefficients hypothesized to be zero (Skinner et al., 1989; Thomas and Rao, 1987). Ignoring these corrections may have at least partially accounted for the overly liberal Type I error rates of the modified Hosmer–Lemeshow and Wald goodness-of-fit tests under complex sampling previously reported (Graubard et al., 1997).

3. Material and methods

3.1. Proposed goodness-of-fit tests for complex sampling

Due to the inflated Type I error rates for the previously proposed tests, we proposed several alternative design-based goodness-of-fit tests and examined their properties empirically. The proposed goodness-of-fit tests for logistic regression applied to complex survey data are calculated in the following manner: after the logistic regression model is fit, the residuals $\hat{r}_{ji} = y_{ji} - \hat{\pi}\left(x_{ji}\right)$ are obtained. These goodness-of-fit tests are based on the residuals since large departures between observed and estimated values would seemingly indicate lack of fit. Then, using the same grouping strategy previously proposed (Graubard et al., 1997), observations are sorted into deciles based on their weights and estimated residuals. Specifically, survey estimates of the sum of the residuals by decile of risk $\hat{\mathbf{T}}' = \left(\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_{10}\right)$ are obtained such that $\hat{T}_g = \sum_j \sum_i \tilde{w}_{ji}\hat{r}_{ji}$ for the gth percentile of the weighted $\hat{r}_{ji}$ values for $g = 1, \ldots, 10$. The associated estimated variance–covariance matrix $\hat{\mathbf{V}}\left(\hat{\mathbf{T}}\right)$ is obtained using linearization. Briefly, the technique of linearization can be used to construct an approximation to the functional form of the estimated population characteristic (Levy and Lemeshow, 1999). As the first step, the functional form of the estimated population characteristic is approximated by a first-order Taylor series, resulting in an approximation that is linear in the sample observations. Afterwards, design-based methods are used to estimate its variance. Using this method, a Wald test can be estimated as

$$Q_T = \hat{\mathbf{T}}'\hat{\mathbf{V}}\left(\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}. \tag{7}$$

which we refer to as the 'total residual test.' This test statistic will be distributed approximately as a χ^2_{g-1} under the null hypothesis, where $g = 10$ (Horton et al., 1999). Variations of this test were also examined. A modification to Eq. (7) was to examine the F-adjusted version, denoted as

$$Q_{T_F} = \frac{f - g + 2}{fg}\hat{\mathbf{T}}'\hat{\mathbf{V}}\left(\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}. \tag{8}$$

Again, f is the number of sampled clusters minus the number of strata and g is the number of groups. Also, the test statistic was constructed assuming the covariances are zero was examined, where $\hat{\mathbf{V}}_0\left(\hat{\mathbf{T}}\right)$ denotes the variance–covariance matrix of $\hat{\mathbf{T}}$ where $\sigma_{ij} = 0$ for every $i \neq j$, and this statistic is denoted as

$$Q_{T_0} = \hat{\mathbf{T}}'\hat{\mathbf{V}}_0\left(\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}. \tag{9}$$

This is referred to as the 'total residual test assuming covariances are zero.'

These tests were re-examined after replacing the total residual and its associated variance–covariance matrix with the mean residuals. That is, survey estimates of the mean residuals by decile of risk $\hat{\mathbf{M}}' = \left(\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_{10}\right)$ are obtained such that $\hat{M}_g = \sum_j \sum_i \tilde{w}_{ji}\hat{r}_{ji} \Big/ \sum_j \sum_i \tilde{w}_{ji}$ for the gth percentile of the $\hat{r}_{ji}$ values for $g = 1, \ldots, 10$. Specifically, the mean residual test, the F-adjusted mean residual test, and the mean residual test assuming the covariances are zero were examined under this modification. These tests are expressed as

$$Q_M = \hat{\mathbf{M}}'\hat{\mathbf{V}}\left(\hat{\mathbf{M}}\right)^{-1}\hat{\mathbf{M}}, \tag{10}$$

$$Q_{M_F} = \frac{f-g+2}{fg}\,\hat{\mathbf{M}}'\hat{\mathbf{V}}\left(\hat{\mathbf{M}}\right)^{-1}\hat{\mathbf{M}}, \tag{11}$$

and

$$Q_{M_0} = \hat{\mathbf{M}}'\hat{\mathbf{V}}_0\left(\hat{\mathbf{M}}\right)^{-1}\hat{\mathbf{M}}. \tag{12}$$

Another proposed goodness-of-fit test for logistic regression models for complex sampling data is based on Tsiatis' approach (Tsiatis, 1980). That is, after the parameter estimates have been obtained a new logistic regression model was fit using the original data and a set of dummy variables that indicate to which of the deciles of risk the subject belongs. This test statistic was investigated due to the perceived ease with which the general user would be able to calculate and apply this goodness-of-fit test. To assess whether the coefficients for the $g-1$ indicator variables are simultaneously equal to zero, rather than calculating a score test, an F-adjusted Wald test was performed where linearization was used to obtain variance estimators for the parameter estimates. Although Tsiatis proposed forming the g groups based on a partition of the covariate space, the formation of the g groups for this proposed test statistic was based on deciles of risk. This is similar to the proposed generalization of the Hosmer–Lemeshow test for evaluating goodness-of-fit of repeated binary outcome data using generalized estimating equations (Horton et al., 1999).

3.2. Simulation study

A simulated population exhibiting intra-cluster correlation was created by generating $m = 375$ clusters, each containing a random number of observations. This resulted in a population consisting of 150,852 observations. A large population was constructed because our primary goal was to understand how well goodness-of-fit tests could be applied when modeling data from a national health survey, such as the National Health Interview Survey (NHIS). The number of observations within each cluster was generated as an integer from a uniform distribution where $N_m \sim \text{Unif}(300, 500)$. Then, within each of the m clusters, cluster-specific parameter values for three different variables were generated. First, a uniform random variable c_{1m} on the interval [0.3, 0.7] was randomly generated to represent the intra-cluster probability of some dichotomous risk factor (e.g., gender, smoking status). Second, within each cluster a uniform random variable c_{2m} on the interval [0, 1] was randomly generated to represent the intra-cluster probability of some second dichotomous risk factor. Third, within each cluster a normally distributed random variable $c_{3m} \sim N(45, 81)$, was generated to represent some intra-cluster mean of a continuous covariate (e.g., age). That is, within the mth cluster, the c's were randomly generated once and the result was taken to be the cluster-specific parameter values, which were subsequently used in the generation of individual observations $i = 1, \ldots, N_m$, as a means of inducing homogeneity of observations within clusters. Once the cluster-specific parameters (c_{1m}, c_{2m}, c_{3m}) were obtained for the mth cluster, the independent covariates $x_1, x_2, x_3, x_4, x_5, x_6$, and x_7 for individual observations $i = 1, \ldots, N_m$ within the mth cluster

were generated as

$$x_{1m_i} \sim \text{Bernoulli}\left(\Pr(x = 1) = c_{1m}\right),$$

$$x_{2m_i} \sim \text{Bernoulli}\left(\Pr(x = 1) = c_{2m}\right).$$

$$x_{3m_i} \sim N\left(c_{3m}, 4\right),$$

$$x_{4m_i} \sim N\left(c_{3m}, 16\right),$$

$$x_{5m_i} \sim N\left(c_{3m}, 64\right),$$

$$x_{6m_i} \sim N\left(c_{3m}, 144\right),$$

$$x_{7m_i} \sim U(-5, 5).$$

Although for each observation the covariates were independently generated, $x_{1m_i}, x_{2m_i}, x_{3m_i}, x_{4m_i}, x_{5m_i}$ and x_{6m_i} exhibit intra-cluster correlation because observations within each cluster were randomly generated using the same within cluster-specific parameter values, c_{1m}, c_{2m}, and c_{3m}. The design effect, estimated as the ratio of the estimate's variance induced by the complex sampling design to the estimate's variance induced by a simple random sample design (Skinner et al., 1989), was consistently around 1.5, leading us to conclude that our simulated population consisted of observations that were homogeneous within clusters.

Using these covariate values, dichotomous outcomes were randomly generated from a Bernoulli distribution, where for each observation the outcome was

$$y_{\text{Scenario}} \sim \text{Bernoulli}\left(\Pr(y = 1) = \frac{\exp\left\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7\right\}}{1 + \exp\left\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7\right\}}\right). \quad (13)$$

Twenty-four different outcomes were generated, each using different covariates and varying parameter values for the β_i's as listed in Table 1. The third column of Table 1 identifies the covariates and their parameter values that were used in generating the outcome using Eq. (13) while the second column identifies the independent covariates that were included in the fitted regression model. For example, for scenario A1, the outcome was generated as $y_{A_1} \sim \text{Bernoulli}\left(\Pr(y = 1) = e^{0.3x_1 + 0.009x_4} / \left(1 + e^{0.3x_1 + 0.009x_4}\right)\right)$ and the logistic regression model was fit to this outcome using x_1 and x_4 as the independent predictors. The various randomly generated outcomes based on known probability models permitted a comparison of the properties of the design-based and iid-based logistic regression goodness-of-fit tests under a variety of scenarios. Scenarios A1–F3 represent situations in which the correct covariates were included in the fitted model, whereas scenarios G1–H3 represent situations in which important covariates were either omitted or their functional form was misspecified in the fitted model.

After the covariates and responses were generated for this simulated population, a simulation study was conducted to evaluate the properties of the proposed goodness-of-fit tests. To mimic an actual survey sample, we repeatedly conducted two-stage cluster samples from this simulated population. For each two-stage cluster sample, the logistic regression model was fit (according to Table 1, column 2 for each outcome listed in column 1) and goodness-of-fit was tested. Each two-stage cluster sample was drawn as follows. At the first stage, k clusters were randomly selected. At the second stage, a random sample within each of the k clusters selected at the first stage was drawn, where the proportion of observations sampled within each sampled cluster depended upon the cluster ID. That is, for cluster IDs ranging from 1 to 125 that were included at the first stage of sampling, 5% of their observations were randomly sampled in the second stage. For cluster IDs ranging from 126 to 250 that were included at the first stage of sampling, 10% of their observations were randomly sampled in the second stage. Likewise, for cluster IDs ranging from 251 to 375 that were included at the first stage of sampling, 25% of their observations were randomly sampled in the second stage. We examined the results for different numbers of clusters sampled in the first stage, letting $k = 25$, 75, or 125. For each simulation study, the two-stage cluster sampling procedure was replicated 2000 times. For each replication, the logistic regression model was fit using both maximum likelihood and pseudo-maximum likelihood to enable a comparison of the properties of traditional iid-based goodness-of-fit methods to the proposed design-based tests. That is, aside

Table 1
Description of the covariates and their corresponding parameter values used in generating the Bernoulli outcomes for 25 different scenarios with the corresponding list of covariates included in the fitted logistic regression model after each two-stage cluster sample was drawn

Scenario	Covariate(s) included in the fitted logistic regression model	Covariates with their corresponding parameter values that were used in randomly generating the Bernoulli outcomes
A1	x_1 and x_4	$\beta_1 = 0.3$ and $\beta_4 = 0.009$
A2	x_1 and x_4	$\beta_1 = 0.5$ and $\beta_4 = 0.009$
A3	x_1 and x_4	$\beta_1 = 0.9$ and $\beta_4 = 0.009$
B1	x_2 and x_4	$\beta_2 = 0.3$ and $\beta_4 = 0.009$
B2	x_2 and x_4	$\beta_2 = 0.5$ and $\beta_4 = 0.009$
B3	x_2 and x_4	$\beta_2 = 0.9$ and $\beta_4 = 0.009$
C1	x_3	$\beta_5 = 0.010$
C2	x_3	$\beta_5 = 0.025$
C3	x_3	$\beta_5 = 0.045$
D1	x_4	$\beta_3 = 0.010$
D2	x_4	$\beta_3 = 0.025$
D3	x_4	$\beta_3 = 0.045$
E1	x_5	$\beta_4 = 0.010$
E2	x_5	$\beta_4 = 0.025$
E3	x_5	$\beta_4 = 0.045$
F1	x_6	$\beta_6 = 0.010$
F2	x_6	$\beta_6 = 0.025$
F3	x_6	$\beta_6 = 0.045$
G1	x_1 and x_4	β_1=0.4 and β_4=0.001, where x_4^2 is the proper functional form of the covariate
G2	x_1 and x_4	$\beta_1 = 0.4$ and $\beta_4 = 0.0015$, where x_4^2 is the proper functional form of the covariate
G3	x_1 and x_4	$\beta_1 = 0.4$ and $\beta_4 = 0.0025$, where x_4^2 is the proper functional form of the covariate
H1	x_1 and x_4	$\beta_0 = -0.5$, $\beta_1 = 0.2$, $\beta_4 = 0.009$, $\beta_{14} = 0.01$, where x_{14} is the interaction term for x_1 and x_4
H2	x_1 and x_4	$\beta_0 = -0.5$, $\beta_1 = 0.2$, $\beta_4 = 0.009$, $\beta_{14} = 0.05$, where x_{14} is the interaction term for x_1 and x_4
H3	x_1 and x_4	$\beta_0 = -0.5$, $\beta_1 = 0.02$, $\beta_4 = 0.009$, $\beta_{14} = 0.10$, where x_{14} is the interaction term for x_1 and x_4

from estimating the proposed test statistics, for each replication the traditional Hosmer–Lemeshow goodness-of-fit test was also applied using the design-based logistic regression estimated probabilities and also after fitting the iid-based logistic regression model which ignored the sampling design altogether. The cluster sampling as well as all additional calculations and analyses were performed using Stata (StataCorp, 2005).

3.3. Application: 2004 NHIS

The 2004 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. This release includes 31,326 observations from a multi-stage sampling design (National Center for Health Statistics, 1999). For the illustrative example in this paper, we estimated 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 (AGE_P), "ever smoked 100 cigarettes" (SMKEV), "frequency of vigorous activity" (VIGNO), and "ever had 12+ drinks in any one year" (ALC1YR).

The following rules were followed in recoding the variables in the NHIS Adult Core dataset (Table 2). 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

Table 2
Recodings applied to the 2004 NHIS Sample Adult Core dataset. The lowest coded category served as the referent in the fitted logistic regression model

Covariate	Original coding	Recoding
SEX	1 = Male 2 = Female	1 = Male 2 = Female
RACRECI2 "Race Recode"	1 = White 2 = Black 3 = All other race groups	−1 = All other race groups 0 = White 1 = Black
SMKEV "Ever smoked 100 cigarettes"	1 = Yes 2 = No 7 = Refused 8 = Not ascertained 9=Don't know	0 = No 1 = Yes Missing = refused, not ascertained, don't know
ALC1YR "Ever had 12+ drinks in any one year"	1 = Yes 2 = No 7 = Refused 8 = Not ascertained 9=Don't know	0 = No 1 = Yes Missing = refused, not ascertained, don't know
VIGNO "Frequency of vigorous activity"	0 = Never 1–500 units 996 = Unable to do this activity 997 = Refused 998 = Not ascertained 999=Don't know	−1 = Unable to do this activity 0 = Never 1 = 1–500 units ("Some") Missing = refused, not ascertained, don't know
HYPEV "Ever been told you have hypertension"	1 = Yes 2 = No 7 = Refused 8 = Not ascertained 9=Don't know	0 = No 1 = Yes Missing = refused, not ascertained, don't know

were used to determine the most appropriate form of the continuous covariate (AGE_P). The best fitting second order model $\left(\text{age}^2 + \text{age}^3\right)$ was significantly better than any other simpler model, however, it was not substantially different than the more readily interpretable full polynomial $\left(\text{i.e., } \text{age} + \text{age}^2 + \text{age}^3\right)$. Therefore, the full polynomial was carried forward in all subsequent models. Prior to estimating 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. After estimating the logistic regression model, the F-adjusted mean residual goodness-of-fit test was applied to assess the fit of the final model.

4. Results

4.1. Simulation study: type I error

The results from the simulations for the correctly specified models when 25, 75, and 125 clusters were sampled appear in Tables 3–5 , respectively. The tabled value is the percent of times the p-value from the goodness-of-fit test was less than 0.05. With 2000 replications, this percent should range from 4% to 6%. Therefore, when the percent ranged from 4% to 6%, the goodness-of-fit test was interpreted to have Type I error rate close to the nominal level; when the percent was above 6%, the test was interpreted to have an inflated Type I error rate; when the percent was below 4% the test was interpreted to have a Type I error rate below the nominal level. The results are also presented graphically as boxplots in Figs. 1–3. Generally, the Tsiatis-like F-adjusted Wald test, total residual test, mean residual test, mean residual test assuming the covariances are zero, and the design-based Hosmer–Lemeshow test gave an inflated Type I error rate. The F-adjusted total residual test and the total residual test assuming the covariance was zero gave Type

Table 3
Percent of times the p-value for the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 25$

Scenario	Tsiatis-like F-adjusted Wald	Total residual test	Mean residual test	F-adjusted total residual test	F-adjusted mean residual test	Total residual test, assuming cov = 0	Mean residual test assuming cov = 0	Design-based Hosmer–Lemeshow	iid-Hosmer–Lemeshow[a]
A1	9.40	28.45	34.40	3.65	6.45	4.50	10.95	7.30	4.70
A2	13.45	29.95	36.40	4.00	7.45	4.85	12.70	7.95	3.95
A3	16.45	27.85	35.15	3.80	7.30	4.45	12.55	6.45	3.35
B1	10.30	28.80	37.90	4.15	8.50	5.60	14.55	7.30	4.05
B2	13.15	28.40	39.40	3.05	8.50	4.45	14.90	6.65	3.95
B3	14.75	28.55	37.75	3.40	9.05	3.85	13.60	6.40	3.15
C1	12.05	32.20	51.60	3.75	17.25	4.15	29.75	7.95	5.30
C2	12.25	34.50	52.65	3.30	17.20	4.80	30.95	8.40	4.90
C3	10.25	33.55	53.55	4.60	19.45	6.30	34.95	7.30	4.15
D1	7.70	31.95	40.50	4.70	10.50	4.95	16.40	7.45	4.95
D2	7.00	31.25	40.90	3.90	9.85	5.65	18.55	7.85	5.10
D3	6.60	36.95	46.55	6.75	16.85	10.20	26.50	7.45	4.85
E1	4.95	31.70	36.20	3.75	6.15	5.95	10.55	8.15	5.45
E2	4.40	31.75	36.95	4.75	6.45	5.70	11.10	7.15	4.80
E3	3.95	40.25	46.45	8.40	12.70	13.40	22.60	8.20	5.25
F1	4.55	32.30	34.90	4.65	6.40	5.50	9.20	7.55	5.10
F2	3.90	32.70	35.55	4.90	6.60	7.45	10.90	7.40	4.75
F3	3.45	41.25	45.90	10.05	14.35	17.20	24.80	8.75	5.50

[a]Goodness-of-fit tests applied to iid-based models.

Table 4
Percent of times the p-value for the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 75$

Scenario	Tsiatis-like F-adjusted Wald	Total residual test	Mean residual test	F-adjusted total residual test	F-adjusted mean residual test	Total residual test, assuming cov = 0	Mean residual test assuming cov = 0	Design-based Hosmer–Lemeshow	iid-Hosmer–Lemeshow[a]
A1	7.50	7.80	9.30	2.90	4.05	2.90	5.05	6.65	4.50
A2	9.40	6.15	7.10	2.45	3.40	2.35	3.95	6.15	3.15
A3	14.00	6.00	8.00	2.35	4.00	2.70	4.45	8.00	3.55
B1	7.80	7.30	9.60	3.05	4.35	2.90	5.50	7.65	4.40
B2	8.35	6.25	8.70	2.20	3.75	2.45	4.90	6.60	3.00
B3	10.40	6.35	8.65	2.45	4.10	2.75	5.35	7.15	3.40
C1	7.80	8.15	13.40	3.50	6.70	3.85	9.50	6.50	4.45
C2	8.45	8.30	13.05	3.70	7.10	3.75	9.45	9.15	6.10
C3	7.35	10.40	16.85	4.25	8.30	4.85	12.30	7.50	4.65
D1	4.85	7.30	10.60	2.40	4.25	2.90	5.80	8.40	5.30
D2	6.35	9.35	12.25	3.95	5.55	3.80	7.80	9.10	6.10
D3	5.95	10.40	13.65	4.70	7.10	5.80	9.40	7.55	4.30
E1	4.85	8.95	10.15	3.25	4.65	4.10	5.30	7.35	4.20
E2	5.05	8.95	10.10	4.05	4.45	4.70	6.40	8.05	5.00
E3	5.75	12.65	14.80	6.50	7.60	8.35	10.65	7.90	5.00
F1	4.50	8.40	9.35	3.30	4.05	3.75	4.70	7.60	4.90
F2	4.10	8.70	9.60	3.30	4.00	3.70	4.85	7.15	4.30
F3	5.45	13.25	15.10	7.30	8.45	7.70	8.95	9.20	5.85

[a]Goodness-of-fit tests applied to iid-based models.

Table 5
Percent of times the p-value from the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 125$

Scenario	Tsiatis-like F-adjusted Wald	Total residual test	Mean residual test	F-adjusted total residual test	F-adjusted mean residual test	Total residual test, assuming cov = 0	Mean residual test assuming cov = 0	Design-based Hosmer–Lemeshow	iid-Hosmer–Lemeshow[a]
A1	7.35	6.15	6.95	3.45	3.80	3.90	4.85	7.40	4.45
A2	9.40	3.75	4.75	2.05	2.60	2.35	3.15	7.15	4.35
A3	12.35	4.10	4.75	2.35	3.20	2.40	3.20	5.15	2.35
B1	5.95	4.45	5.75	2.40	3.40	2.60	4.35	7.20	4.05
B2	9.05	4.50	5.95	2.40	3.55	2.90	4.15	6.30	2.75
B3	9.70	4.05	5.95	2.20	3.20	1.75	3.40	7.45	3.35
C1	7.35	5.55	7.90	3.55	5.20	4.00	6.70	6.90	4.60
C2	6.65	5.80	7.95	3.30	4.90	3.55	6.35	10.35	6.95
C3	7.15	7.85	10.85	4.50	7.00	4.70	8.15	8.65	5.65
D1	5.85	5.85	7.35	3.45	4.30	3.05	4.65	7.75	4.85
D2	6.95	6.95	8.50	4.50	5.50	5.20	6.75	9.75	6.30
D3	6.15	7.40	9.10	4.85	6.25	4.55	6.60	6.55	4.00
E1	6.25	6.50	7.30	3.70	4.45	4.05	4.95	7.65	4.50
E2	5.90	6.90	7.75	3.90	4.55	3.80	4.95	9.30	6.15
E3	5.35	8.20	8.90	4.95	5.65	5.15	6.35	7.00	4.05
F1	4.15	5.55	6.15	3.20	3.30	3.45	3.85	7.25	4.70
F2	4.80	6.50	7.20	3.50	4.10	3.80	4.85	6.45	3.85
F3	5.70	7.95	8.95	5.25	5.70	5.95	6.65	7.70	4.80

[a]Goodness-of-fit tests applied to iid-based models.

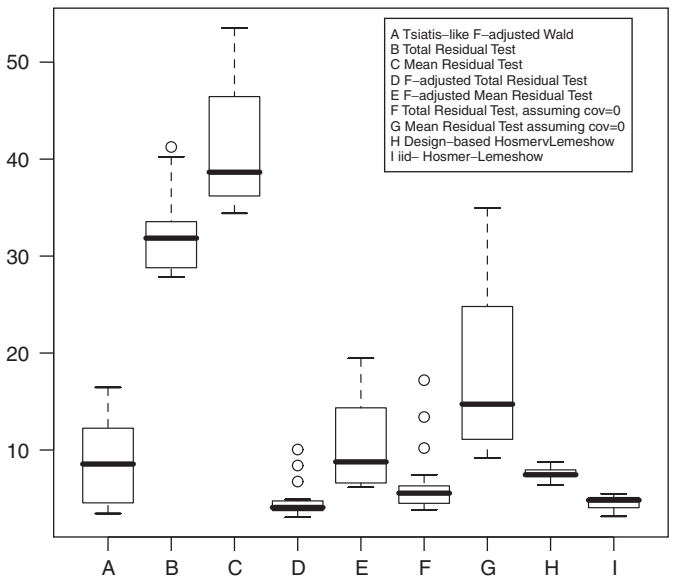

Fig. 1. Boxplot of percent of times the p-value for the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 25$ by test statistic.

I error rates below the nominal level as the number of sampled clusters increased. The F-adjusted mean residual test and MLE Hosmer–Lemeshow tests gave Type I error rates close to the nominal level, particularly as the number of sampled clusters increased.

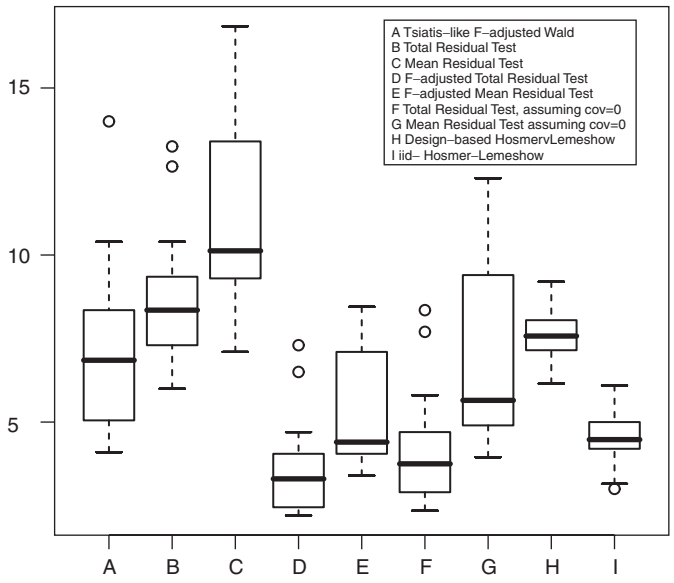

Fig. 2. Boxplot of percent of times the p-value for the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 75$ by test statistic.

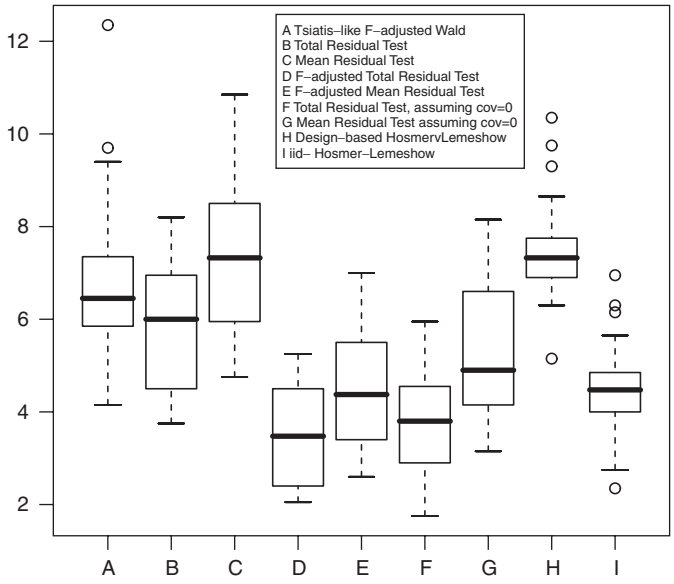

Fig. 3. Boxplot of percent of times the p-value from the goodness-of-fit test was less than 0.05 for each scenario when the number of sampled clusters was $k = 125$ by test statistic.

4.2. Simulation study: power

The results from the simulations for the misspecified models when 25, 75, or 125 clusters were sampled appear in Table 6 and in Fig. 4, where again the tabled value represents the percentage of times the goodness-of-fit test was less than 0.05. For the two test statistics with Type I error rates close to the nomimal level (the F-adjusted mean residual test and the traditional Hosmer–Lemeshow test) the Hosmer–Lemeshow goodness-of-fit test applied to the data as if were from a simple random sample both gave low power when the proper functional form of a continuous covariate

Table 6
Power for each goodness-of-fit test estimated as the percent of times the p-value < 0.05 for each scenario when a mis-specified model was fit

Scenario	Tsiatis-like F-adjusted Wald	Total residual test	Mean tesidual test	F-adjusted total residual test	F-adjusted mean residual test	Total residual test, assuming cov = 0	Mean residual test assuming cov = 0	Design-based Hosmer–Lemeshow	iid-Hosmer–Lemeshow[a]
$K = 25$									
G1	5.80	52.10	63.70	15.30	36.45	25.90	49.55	9.70	5.85
G2	7.55	86.50	92.80	55.85	86.45	74.25	88.90	9.20	5.15
G3	10.75	99.85	100.00	99.85	100.00	99.55	100.00	6.75	2.80
H1	13.00	32.85	39.80	4.75	9.65	5.30	15.70	11.10	6.15
H2	40.50	71.80	80.30	27.95	47.75	38.35	62.00	52.50	42.95
H3	34.95	94.95	96.80	92.80	96.10	86.60	95.90	29.30	18.70
$K = 75$									
G1	13.60	27.85	32.35	18.30	24.10	20.70	28.05	15.95	11.50
G2	12.35	65.55	70.85	58.05	64.45	61.75	68.25	14.20	8.55
G3	8.70	100.00	100.00	100.00	100.00	100.00	100.00	3.55	1.00
H1	16.75	14.05	17.30	7.10	9.40	7.85	11.85	19.00	12.15
H2	85.50	86.90	89.75	78.65	83.20	82.25	87.15	97.10	95.50
H3	72.80	95.40	96.85	92.40	94.10	93.50	95.15	83.45	78.35
$K = 125$									
G1	18.55	28.85	31.40	21.90	25.55	23.50	28.20	23.95	18.95
G2	21.45	61.50	64.90	56.30	60.45	59.05	63.30	21.20	15.05
G3	10.70	100.00	100.00	100.00	100.00	100.00	100.00	3.75	1.25
H1	22.00	16.25	19.80	11.35	13.50	12.45	15.60	29.20	20.70
H2	98.10	97.95	98.35	96.55	97.25	97.35	98.15	99.90	99.75
H3	94.85	98.10	98.35	97.35	97.55	97.30	98.05	98.45	98.05

[a]Goodness-of-fit tests applied to iid-based model.

was misspecified but good power when an important interaction term was omitted. Generally, the F-adjusted mean residual test demonstrated better power. The power of both tests increased as the parameter value increased.

4.3. Results from 2004 NHIS

Age, smoking, alcohol use, gender, frequency of vigorous activity, and race were all important predictors for hypertension (Table 7). The odds of having hypertension among those having ever smoked at least 100 cigarettes was greater compared to those who had not ever smoked 100 cigarettes (Odds ratio (OR) 1.16; 95% CI 1.08–1.24). The odds of having hypertension was lower among females compared to males (OR 0.92; 95% CI 0.86–0.99); among those who drank at least 12 drinks in any one year compared to those who did not (OR 0.89; 95% CI 0.82–0.96); and among those who never or had some vigorous activity compared to those who were unable to participate in vigorous activities, with OR 0.57 and 0.45, respectively. The odds of having hypertension among whites was significantly greater compared to those in all other races (OR 1.80; 95% CI 1.64–1.97), however, blacks were not significantly different from all other races with respect to having ever been told they had hypertension. The F-adjusted mean residual goodness-of-fit test did not indicate any overall model departure from the observed data ($p = 0.31$). We recommend that this design-based goodness-of-fit test be examined after a design-based logistic regression model has been estimated, particularly if the purpose of the model is to estimate probabilities of event occurrences.

5. Discussion

The simulations for the simulated population were run under a variety of conditions. Conditions that were varied in the simulation study included (i) different number of sampled clusters ($m = 25$, 75, 125), and (ii) examination of the Type I error using the probabilities from both the design-based model as well as the probabilities from the

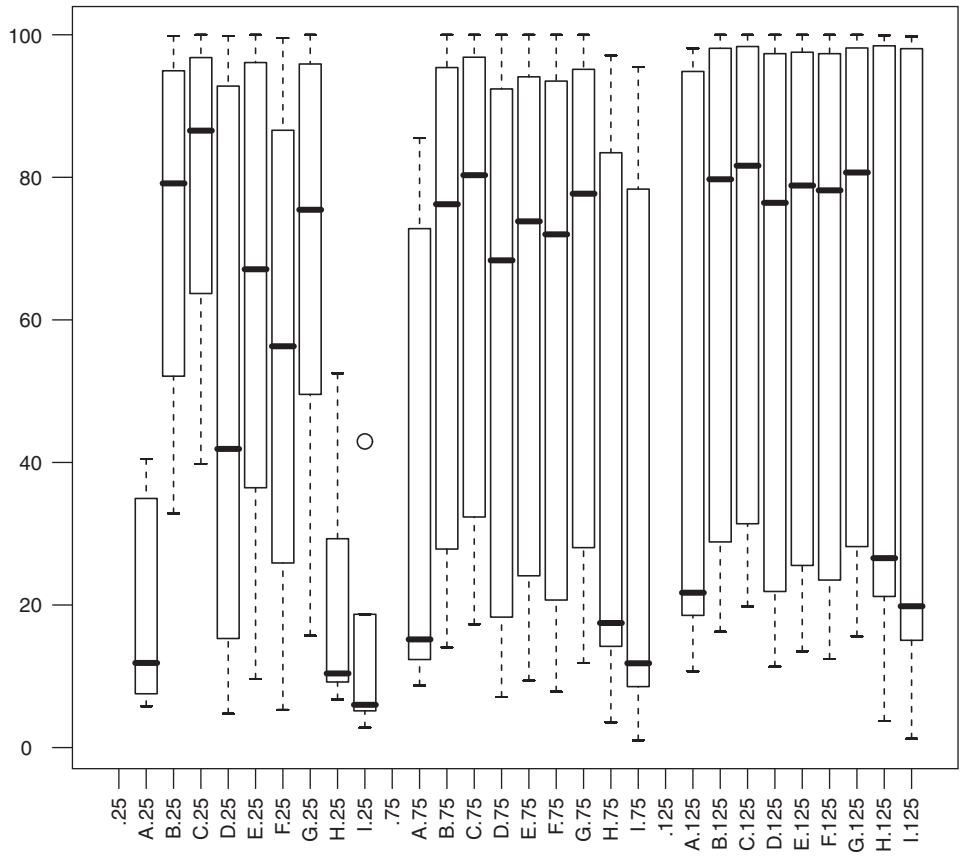

Fig. 4. Boxplot of power for each goodness-of-fit test estimated as the percent of times the p-value < 0.05 for each scenario when a mis-specified model was fit by test statistic. X-axis labels indicate test statistic (A="Tsiatis-like F-adjusted Wald", B="total residual test", C="mean residual test", D="F-adjusted total residual test", E="F-adjusted mean residual test", F="total residual test, assuming cov = 0", G="mean residual test assuming cov = 0", H="design-based Hosmer–Lemeshow", I="iid-Hosmer–Lemeshow") and number of sampled clusters $k = 25, 75, 125$.

Table 7
Estimated odds ratios and 95% confidence intervals for variables predicting hypertension using the 2004 NHIS Adult Sample Core data

	Odds ratio	95% Estimated confidence interval
SMKEV="Yes"	1.156	1.080, 1.237
ALC1YR="Yes'	0.887	0.822, 0.958
SEX="Female"	0.924	0.863, 0.991
VIGNO="Never"	0.569	0.477, 0.678
VIGNO="Some"	0.451	0.375, 0.542
RACERECI2="White"	1.798	1.643, 1.968
RACERECI2="Black"	0.899	0.756, 1.069

population parameter values for several proposed goodness-of-fit tests. The Type I error rates estimated when the Hosmer–Lemeshow goodness-of-fit test was applied using the design-based logistic regression estimates were compared to the Type I error rates when the sampling design was ignored altogether. Misspecified models were fit to assess the statistical power of the goodness-of-fit tests. Results from these simulations were verified by examining the calculations by hand, and by replicating the simulation studies using different initial random seeds. The different sets of simulations yielded similar results.

Although the iid-based Hosmer–Lemeshow test gave a Type I error rate close to the nominal level for most scenarios studied, it gave low power to detect model misspecification when the proper functional form of a continuous variable was misspecified, regardless of the number of clusters sampled. This is consistent with previous findings (Hosmer et al., 1997). The goodness-of-fit tests for complex sampling data were more powerful in detecting whether a continuous covariate was misspecified or an interaction term was omitted from the model than were the iid-based tests previously reported in the literature.

It is important to note that the Tsiatis-like F-adjusted Wald, and total and mean residual goodness-of-fit tests in this simulation study were based on deciles of risk, similar to the frequently used Hosmer–Lemeshow goodness-of-fit test. One problem with the Tsiatis-like F-adjusted Wald test was that sometimes the indicator variables representing decile of risk were highly collinear in the logistic regression model. This resulted in certain indicator(s), representing a given decile of risk, being dropped from the model. In so doing, there could be problems, such as loss of power due to decreased degrees of freedom. Since one cannot control problems of multicollinearity with the decile of risk indicator variables, the Tsiatis-like F-adjusted Wald test is not recommended. Nevertheless, all of these tests suffer from the same liability as discussed by other investigators (le Cessie and van Houwelingen, 1995). Namely, they are unable to detect departures that occur within a given decile and that the resulting p-value is dependent upon the partitioning into deciles of risk. In particular, different software packages have implemented the Hosmer–Lemeshow test in different ways and the strategy they use may result in different test statistics.

Based on these results, the suggestion that one could use design-based methods for parameter estimation but then carry out goodness-of-fit tests using iid-based methods (Hosmer and Lemeshow, 2000) seems appropriate if the correctly specified model was fit. However, for some types of misspecified models or when observations are sampled with respect to the outcome variable, traditional goodness-of-fit tests may not reject a model whose probabilities are not consistent with the observed outcomes. Since one rarely knows the correct model, as such knowledge would make model estimation a moot point, we recommend the F-adjusted mean residual test for testing goodness-of-fit for data collected using a complex sampling design, particularly when the number of sampled clusters is large. This test gave good power for the misspecified models, with increased power as the strength of the relationship between the covariate and outcome variable increased. This test also gave a Type I error rate close to the nominal level for many of the correctly specified scenarios. A Stata ado function `svylogitgof` for estimating the F-adjusted mean residual test after `svylogit` fit has been recently described (Archer and Lemeshow, 2006) and is available at http://www.people.vcu.edu/∼kjarcher/Research/Data.htm.

Since the hoped-for outcome from a goodness-of-fit test is to fail to reject the null hypothesis, one risks committing a Type II error whenever one concludes that there is no evidence of lack of fit. Therefore, statistical power is important in assessing goodness-of-fit. Additional simulation studies investigating the power of these proposed goodness-of-fit tests on difference populations using different misspecified models would be worthwhile. Additionally, increasing the number of replications for each simulation to more accurately quantify the statistical power would be beneficial. Finally, investigations into the use of a weighted version of the Hosmer–Lemeshow test is currently being explored.

References

Archer, K.J., Lemeshow, S., 2006. Goodness-of-fit test for a logistic regression model estimated using survey sample data. The Stata J. 6 (1), 97–105.

Azzalini, A., Bowman, A.W., Härdle, W., 1989. On the use of nonparametric regression for model checking. Biometrika 76 (1), 1–11.

Brown, C.C., 1982. On a goodness-of-fit test for the logistic model based on score statistics. Comm. Statist. Theory Meth. 11 (10), 1087–1105.

Cox, D.R., 1958. Two further applications of a model for binary responses. Biometrika 45, 562–565.

Graubard, B.I., Korn, E.L., Midthune, D., 1997. Testing goodness-of-fit for logistic regression with survey data. In: Proceedings of the Section on Survey Research Methods.American Statistical Association,

Horton, N.J., et al., 1999. Goodness-of-fit for GEE: an example with mental health service utilization. Statist. Medicine 18, 213–222.

Hosmer, D.W., Lemeshow, S., 1980. Goodness-of-fit tests for the multiple logistic regression model. Comm. Statist. Theory Meth. A A9 (10), 1043–1069.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. Wiley, New York.

Hosmer, D.W., et al., 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statist. Medicine 16, 965–980.

Korn, E.L., Graubard, B.I., 1991. Surveys: accounting for the sampling design. Amer. J. Public Health 81 (9), 1166–1173.

Korn, E.L., Graubard, B.I., 1995. Analysis of large health surveys: accounting for the sampling design. J. Roy. Statist. Soc. A 158, 263–295.

Korn, E.L., Graubard, B.I., 1999. Analysis of Health Surveys. Wiley, New York.

le Cessie, S., van Houwelingen, J.C., 1991. A goodness-of-fit test for binary regression models, based on smoothing methods. Biometrics 47, 1267–1282.

le Cessie, S., van Houwelingen, H.C., 1995. Testing the fit of a regression model via score tests in random effects models. Biometrics 51, 600–614.
Lemeshow, S., et al., 1998. Illustration of analysis taking into account complex survey considerations: the association between wine consumption and dementia in the PAQUID study: personnes ages quid. Amer. J. Epidemiology 148, 298–306.
Levy, P.S., Lemeshow, S., 1999. Variance estimation in complex sample surveys. In: Sampling of Populations: Methods and Applications. Wiley, New York, pp. 365–390.
National Center for Health Statistics, 1999. National Health Interview Survey: Research for the 1995–2004 redesign. vol. Series 2. Vital and Health Statistics, Hyattsville, MD.
Osius, G., Rojek, D., 1992. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. J. Amer. Statist. Assoc. 87 (420), 1145–1152.
Pigeon, J.G., Heyse, J.F., 1999a. An improved goodness-of-fit statistic for probability prediction models. Biometrical J. 41 (1), 71–82.
Pigeon, J.G., Heyse, J.F., 1999b. A cautionary note about assessing the fit of logistic regression models. J. Appl. Statist. 26 (7), 847–853.
1989. In: Skinner, C.J., Holt, D., Smith, T.M.F. (Eds.), Analysis of Complex Surveys. Wiley, New York.
StataCorp, 2005. Stata Statistical Software: Release 9. College Station, TX.
Su, J.Q., Wei, L.J., 1991. A lack-of-fit test for the mean function in a generalized linear model. J. Amer. Statist. Assoc. 86 (414), 420–426.
Thomas, D.R., Rao, J.N.K., 1987. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. J. Amer. Statist. Assoc. 82 (398), 630–636.
Tsiatis, A.A., 1980. A note on a goodness-of-fit test for the logistic regression model. Biometrika 67 (1), 250–251.