STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1305–1324
Published online 20 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3011

Testing logistic regression coefficients with clustered data and few positive outcomes‡

Sally Hunsberger[1,*,†], Barry I. Graubard[2] and Edward L. Korn[1]

[1]*Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.*
[2]*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, U.S.A.*

SUMMARY

Applications frequently involve logistic regression analysis with clustered data where there are few positive outcomes in some of the independent variable categories. For example, an application is given here that analyzes the association of asthma with various demographic variables and risk factors using data from the third National Health and Nutrition Examination Survey, a weighted multi stage cluster sample. Although there are 742 asthma cases in all (out of 18 395 individuals), for one of the categories of one of the independent variables there are only 25 asthma cases (out of 695 individuals). Generalized Wald and score hypothesis tests, which use appropriate cluster-level variance estimators, and a bootstrap hypothesis test have been proposed for testing logistic regression coefficients with cluster samples. When there are few positive outcomes, simulations presented in this paper show that these tests can sometimes have either inflated or very conservative levels. A simulation-based method is proposed for testing logistic regression coefficients with cluster samples when there are few positive outcomes. This testing methodology is shown to compare favorably with the generalized Wald and score tests and the bootstrap hypothesis test in terms of maintaining nominal levels. The proposed method is also useful when testing goodness-of-fit of logistic regression models using deciles-of-risk tables. Published in 2007 by John Wiley & Sons, Ltd.

KEY WORDS: bootstrap; clustering; generalized Wald statistics; generalized score statistics; goodness of fit; survey methods

1. INTRODUCTION

During 1980–1994, self-reported asthma increased in the U.S. by 75 per cent, resulting in about 13.5 million asthmatics [1]. Understanding the risk factors for asthma may be useful in controlling future increases. Arif *et al.* [2] used logistic regression to examine the risk factors associated

*Correspondence to: Sally Hunsberger, Biometric Research Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.
†E-mail: sallyh@ctep.nci.nih.gov

Received 14 November 2006
Accepted 19 June 2007

with asthma and wheezing for 18 825 U.S. adults aged $\geqslant$20 years sampled in the third National Health and Nutrition Examination Survey (NHANES III). NHANES III is a complex multistage survey involving the sampling of clusters at various stages and having sample weights that reflect the differential rates of sampling of subjects [3]. Of the 18 825 adults sampled, there were only 742 (3 per cent) positive responses for the outcome of asthma. For some of the risk factors for asthma, there were categories with small numbers of asthma cases, e.g. there were only 25 asthma cases (out of 695 individuals, 3.6 per cent) in one of the defined race categories. In this setting of testing logistic regression coefficients with few positive outcomes, hypothesis tests whose reference distributions are based on asymptotic theory may not attain their nominal levels, and this problem can be exacerbated by the clustering. In addition, for analyzing surveys like the NHANES III, sample weighting needs to be taken into account in the testing methods.

With clustered data, the method of analysis depends on the inference of interest. In this paper, we focus on marginal models where the population-average effect of the covariate on the outcome is of interest and the cluster effects are considered to be nuisance terms that are not explicitly modeled. Marginal models account for dependence due to clustering in estimators of variances of regression coefficients. In particular, the classical Wald and score tests for independent observations that are used for testing regression coefficients can be modified (generalized Wald and score tests) to accommodate clustered data and also to accommodate sample weights that are used in surveys.

In this paper, we consider the general problem of testing logistic regression coefficients. Graubard *et al*. [4] proposed using a simulation-based goodness-of-fit test that only partially accounts for the correlation in the data due to clustering. We propose a simulation-based method of testing that more fully accounts for correlation in the data. In the proposed simulation-based method, the approximate null-hypothesis distribution of the generalized-score test statistic is generated to lessen the reliance on asymptotic approximations. To account for the correlation, we estimate an intraclass correlation parameter from the original data set and use this estimate of correlation in the generation of the distribution of the test statistic. The proposed method can account for sample weights in the survey data.

The idea of simulating the distribution of a test statistic under a null hypothesis to avoid an asymptotic approximation is well known and not new [5]. What is different in the proposed method is the fact that the simulated data sets do not necessarily accurately mimic the structure of the original data set; the dependence of the data due to the clustering could be quite complex and not amenable to simple modeling. However, since the test statistics utilized do accommodate arbitrary correlation due to clustering, we show *via* simulation that the statistical properties of the test statistics are similar, regardless of whether the dependence is modeled completely accurately or not. In particular, problems with asymptotic approximations and small numbers of positive events are (partially) ameliorated. For large sample sizes and numbers of positive events, the test statistics have approximate chi-squared distributions for both the observed data and the simulated data sets [6], ensuring the asymptotic validity (type I error) of the proposed method.

In the next section, we give some details concerning the generalized test statistics, the proposed method, as well as a bootstrap approach. In Section 3, we present some simulation results to show that our approach works better than the standard approaches when the expected number of positive outcomes is small. Section 4 presents the analysis of the asthma data from NHANES III, for which we consider the effect of various demographic variables and risk factors on the binary outcome, asthma. We utilize our methodology for testing the regression coefficients of these covariates as well as for testing the goodness of fit of the model. We end with a discussion in Section 5

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

of alternative approaches to the problem of logistic regression with a small number of positive outcomes.

2. APPROACHES TO CALCULATING TEST STATISTICS AND p-VALUES

We begin by considering a general parametric setting. Consider a p-dimensional parameter vector (θ, λ) where the q-dimensional θ is the parameter of interest to be tested *versus* θ_0, and λ is a vector of nuisance parameters. Using (unclustered) likelihood theory, let $L(\theta, \lambda)$ be the likelihood, $(\hat{\theta}, \hat{\lambda})$ be the unconstrained maximum likelihood estimator, and $(\theta_0, \hat{\lambda}_0)$ be the constrained maximum likelihood estimator. The classical Wald, score, and likelihood ratio tests for independent observations are described in Kotz *et al*. [7], Tarone [8], and Strawderman [9], respectively. To accommodate clustered data, the generalized Wald and score tests substitute cluster-level covariance estimators (sometimes referred to as robust estimators or sandwich estimators) for the model-based unclustered estimators used in the classical tests; for details, see Binder [10], Rotnitzky and Jewell [11], Boos [12], and Rao *et al*. [6], and for the specific case of logistic regression, Horton *et al.* [13]. Typically, these test statistics are compared with chi-squared quantiles to obtain p-values. In this paper, we compare the generalized Wald and score test statistics to $(dq)/(d - q + 1)$ times the quantiles of an F distribution with d and $d - q + 1$ degrees of freedom, where d equals the number of clusters minus one and q is the number of parameters being tested [14, 15].

Another approach that has been suggested is a bootstrap test of $\theta = \theta_0$ [16]. In our implementation, the generalized score statistic for testing $\theta = \theta_0$, as well as the maximum likelihood estimator $(\hat{\theta}, \hat{\lambda})$, is first calculated using the observed data. Then the clustered data are resampled multiple times by resampling the clusters. For each of these bootstrap data sets, the generalized score statistic for testing $\theta = \hat{\theta}$ is calculated. The bootstrap p-value is then given by the proportion of the bootstrap data sets for which the score statistic is larger than or equal to the score statistic calculated using the original data.

We now describe our proposed simulation-based method for computing p-values. The procedure is as follows:

(1) Using the observed cluster data, calculate a test statistic that accommodates clustering (e.g. the generalized score statistic).
(2) Fit the null hypothesis model to the observed data to obtain the regression coefficients.
(3) Estimate the Oman–Zucker intraclass correlation parameter [17] under the null hypothesis model using the observed data. We use a modification of the estimation procedure proposed by Oman and Zucker (see the Appendix for details).
(4) Create a large number of data sets (e.g. 999), with each data set containing the same number of observations, covariate values, and cluster designations as the original data. Generate the binary outcomes according to the Oman–Zucker method (see the Appendix for details) so that these binary outcomes have marginal expectations that satisfy the null model with coefficients equal to the estimated coefficients from (2) and intraclass correlation equal to the estimated intraclass correlation parameter in (3).
(5) Calculate the test statistic for each of the simulated data sets.
(6) Estimate the p-value as the proportion of simulated test statistics that are greater than or equal to the test statistic using the observed data.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

Throughout this paper we will base the proposed simulation-based method on the generalized score test. Note that, for a univariate θ, one can also obtain a test-based confidence interval for θ using this approach. Also note that we utilize the Oman–Zucker model for intraclass correlation rather than a simpler nonlinear mixed-effects model because the latter would not be consistent with the assumed marginal (cluster-free) logistic regression model; see also the Discussion.

We can modify the proposed simulation-based method to analyze sample survey data with sample-weighted observations. For applications involving sample surveys, each observation will typically have a sample weight associated with it. The sample weight estimates the number of people in the population that the sampled person represents. Weighted logistic regression is used for estimating the regression coefficients so that the results will be representative of the population. Standard errors for coefficients and tests of coefficients using generalized Wald and score statistics are constructed by using appropriate covariance estimators that are based on generalizations of the methods described above to weighted data [6, 10, 18].

For the proposed method, weights are incorporated as follows:

(1) Calculate a *weighted* generalized score statistic using the observed data [6].
(2) Fit the null hypothesis model to the *unweighted* observed data to obtain the estimated regression coefficients.
(3) Estimate the Oman–Zucker intraclass correlation parameter under the null hypothesis using the *unweighted* observed data (see the Appendix for details).
(4) Create a large number of data sets (e.g. 999), with each data set containing the same number of observations, covariate values, and cluster designations as the original data. Generate the binary outcomes according to the Oman–Zucker method (see the Appendix for details) so that these binary outcomes have marginal expectations that satisfy the null model with coefficients equal to the estimated coefficients from (2) and intraclass correlation equal to the estimated intraclass correlation parameter in (3).
(5) Calculate the *weighted* test statistic for each of the simulated data sets.
(6) Estimate the p-value as the proportion of simulated test statistics that are greater than or equal to the *weighted* generalized score statistic using the observed data.

The reason why we simulate data sets using unweighted regression coefficients rather than using weighted regression coefficients is that we want the simulated data sets to appear similar to the observed data, and not the population from which the observed data were sampled. In particular, if there are a small number of positive observations in a given category of a variable in the observed data, we desire there to be a small number of positive observations in that category in the simulated data sets even if a weighted analysis suggests that the positivity rate is much higher in the population for that category.

Since the outcomes are being generated in the simulated data sets without using the sample weights, the sample weights in the simulated data sets are noninformative. That is, the sample weights are unrelated to the outcome conditional on the covariates in the simulated data sets. This will typically not be the case in the observed data. However, since the test statistics for the observed data are calculated using the sample weights, these test statistics accommodate informative sample weights. In particular, for large sample sizes they will have approximate chi-squared distributions for the observed data (and simulated data sets) [6]. This ensures the asymptotic validity of the approach. This is the same argument that was used in the Introduction concerning the correlation structure of the simulated data sets potentially not being the same as in the observed data.

3. SIMULATIONS

3.1. Simulation methods

We evaluate the various methods discussed in Section 2 in the context of logistic regression *via* some limited simulations. All simulation results are based on 50 000 repetitions. For the proposed simulation-based method, 999 data sets were simulated for each repetition. For the bootstrap method, 999 bootstrap samples were used for each repetition. As a baseline model, we consider a marginal model with a single continuous covariate (Z) and a categorical variable (X) with C categories:

$$\log \frac{P(Y_i=1)}{1-P(Y_i=1)} = \alpha + \eta Z_i + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_{C-1} I_{i,C-1} \tag{1}$$

Here, the I_{ij} are indicator variables indicating whether X for observation i is in the jth category; when $I_{i1} = I_{i2} = \cdots = I_{i,C-1} = 0$, X for observation i is in the Cth category. We will usually consider the null hypothesis of interest to be that the probability of response is independent of category controlling for Z, that is, $\beta_1 = \beta_2 = \cdots = \beta_{C-1} = 0$, but sometimes we will consider testing $\eta = 0$.

The data are generated in k clusters with m observations per cluster. Three intraclass correlation structures are evaluated when testing the β coefficients: in one, there is no intraclass correlation in Y, X, or Z; in a second, there is intraclass correlation in Y but not in X or Z; and in a third, there is intraclass correlation in Y and X but not in Z. Two intraclass correlation structures are evaluated when testing the η coefficient: no intraclass correlation in Y, X, or Z and intraclass correlation in Y and Z but not in X.

The intraclass correlation in X was generated as follows. For each cluster, $C-1$ uniform random numbers were generated and ordered, $0 \equiv t_{(0)} < t_{(1)} < \cdots < t_{(C-1)} < t_{(C)} \equiv 1$. Then, for each observation within that cluster, X was generated to take the value i with probability $t_{(i)} - t_{(i-1)}$. This implies that $P(X_j = i \mid X_k = i) = [2C/(C+1)](1/C)$ if the X_j and X_k are in the same cluster (rather than $1/C$ corresponding to an intraclass correlation of 0). The intraclass correlation in Y was generated using the Oman–Zucker model with $\gamma = 0.7$. To get an idea of how large this correlation is, we note that if the $P(Y_i = 1)$ were constant, this would be equivalent to a constant intraclass correlation of $0.49 = (0.7)^2$. In particular, if $P(Y_i = 1) \equiv 0.01$ then $P(Y_i = 1 \mid Y_j = 1) = 0.495$ if the Y_i and Y_j are in the same cluster.

The continuous covariate Z was distributed as uniform $(0, 1)$, where for observation j in cluster i, $Z_{ij} = \Phi(R_{ij})$, $\Phi(x)$ denotes the cumulative normal distribution, and R_{ij} is a random variable generated from a standard normal with an intracluster correlation ρ. In the simulations, $\rho = 0$ or 0.49, which corresponds to an intracluster correlation for Z of 0 and approximately 0.49, respectively. In all the simulations, $\eta = 0$ (η is the coefficient for Z).

3.2. Computational issues

A Newton–Raphson algorithm is used for calculating the maximum likelihood estimates of the parameters. Calculating the maximum likelihood estimates of a parameter will be a problem when it is the parameter associated with a category (of a categorical variable) that has no positive events. For the generalized score test, this is a problem only if the parameter is considered to be a nuisance parameter because if it is the parameter of interest, the parameter is set to the null hypothesis and will not need to be estimated. For the generalized Wald test, this is a problem both when the

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

parameter is of interest and when it is considered a nuisance parameter, since it is estimated in both cases.

If this problem occurs when calculating the generalized score test, individuals with that given category are removed from the data set and that category is dropped from the model before calculating the score statistic. Computationally, the same results can be obtained by restricting the estimated logistic regression coefficient to be larger than some small negative number, e.g. -50.

If this problem occurs when calculating the generalized Wald test the following procedure is used. A cluster is randomly selected and a subject with a positive event is added to each level of the covariate. The added observations are given a weight of 0.5 and a Z covariate value of 0.5.

Finally, in the simulations, because we are considering the case of a small event rate, some replications of the 50 000 data sets have either no positive outcomes or so few positive outcomes that the Newton–Raphson algorithm convergence criteria are not met. When this occurs these replications are removed from the simulation (i.e. the total number of simulated data sets will be less than 50 000). Data sets that are generated to simulate the null distribution of the test statistic (for both the proposed method and the bootstrap method) can also have no positive outcomes or no convergence. In these cases, the test statistic is given a value smaller than the observed test statistic.

At the end of the results section, data are presented on the frequency of data sets with no events and data sets that do not meet convergence criteria (both for the 50 000 simulated data sets and for the data sets to calculate the null distribution).

3.3. Simulation results

Table I presents simulated levels for the various tests with $C = 4$ categories, $k = 100$ clusters, $m = 20$ observations per cluster, and the marginal response probability being 20, 2, and 1 per cent. For a marginal response probability of 20 per cent, there are an expected 400 events; hence, this is not a small-event situation. With regard to the classical tests, simulation #1 demonstrates that intraclass correlation in Y alone is not sufficient for these tests to fail, but intraclass correlation in X is also required. This is similar to the results for linear regression [19]. Simulations #2 and 3 consider small-event situations. The generalized Wald test has a highly inflated level, and the bootstrap test has a highly conservative level. The generalized score test sometimes has an inflated level (simulations #2a and 3a) and sometimes has a conservative level (simulations #2c, 3b, and 3c). The proposed test sometimes has a conservative level (simulations #2c, 3b, and 3c). (The previous version of our simulation-based test [4], which generated simulated data sets with zero intraclass correlation, leads to even more conservative results when there are few events, e.g. the simulated levels for simulations #2c and 3c were 0.019 and 0.005, respectively.) Note that when the classical score test is appropriate (simulations #1a, 2a, and 3a), the levels of this test are less affected by the small number of positive events than the generalized score test, verifying the statement made in the Introduction. On the basis of the simulation results in Table I, we consider only the generalized score test and our proposed test in the additional simulations.

Table II presents simulated levels under some different situations (the results for simulations #3a and 3c from Table I are repeated). Simulation #3d is identical to simulation #3c, except that the intraclass correlation of X is 1. That is, for each cluster, all the observations are in only one of the four categories. A situation in which this may arise is cluster randomization, in which everyone in the same cluster receives the same treatment [20]. The simulated levels for this situation for both tests are very conservative.

Table I. Simulated levels (nominal level 0.05) using various tests of the null hypothesis in which the response probabilities [$P(Y=1)$] are constant across $C=4$ categories in model (1) with $k=100$ clusters and $m=20$ observations per cluster.

Simulation number	Intraclass correlation		Marginal $P(Y=1)$ (per cent)	Simulated levels*					
				Classical tests			Generalized tests		
	Y†	X‡		Likelihood ratio	Wald	Score	Wald	Score	Bootstrap	Proposed
1a	No	No	20	0.053	0.052	0.053	0.050	0.052	0.051	0.052
1b	Yes	No	20	0.051	0.050	0.051	0.062	0.049	0.046	0.050
1c	Yes	Yes	20	0.431	0.427	0.430	0.073	0.051	0.044	0.051
2a	No	No	2	0.055	0.033	0.049	0.033	0.063	0.015	0.050
2b	Yes	No	2	0.051	0.0287	0.049	0.236	0.052	0.007	0.050
2c	Yes	Yes	2	0.391	0.323	0.380	0.364	0.033	0.000	0.040
3a	No	No	1	0.042	0.0153	0.047	0.025	0.080	0.000	0.051
3b	Yes	No	1	0.039	0.0145	0.044	0.309	0.035	0.000	0.028
3c	Yes	Yes	1	0.357	0.225	0.311	0.386	0.023	0.000	0.019

*The standard error for a simulation with a true 0.05 level is 0.0010.
†Intraclass correlation in the Y variable is zero ('no') or generated as described in the text with $\gamma=0.7$ ('yes').
‡Intraclass correlation in the X variable is zero ('no') or generated as described in the text ('yes').

Published in 2007 by John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table II. Simulated levels (nominal level 0.05) using the generalized score test and the proposed test of the null hypothesis in which the response probabilities $[P(Y=1)]$ are constant across C categories in model (1) with k clusters and m observations per cluster.

Simulation number	Intraclass correlation		Marginal $P(Y=1)$ (per cent)	Number of categories (C)	Number of clusters (k)	No. of observations per cluster (m)	Simulated levels*	
	$Y^{\dagger}$	$X^{\ddagger}$					Generalized score test	Proposed test
3a	No	No	1	4	100	20	0.080	0.051
3c	Yes	Yes	1	4	100	20	0.023	0.019
3d	Yes	1[§]	1	4	100	20	0.012	0.010
4a	No	No	1	4	1000	2	0.079	0.050
4b	Yes	Yes	1	4	1000	2	0.073	0.053
5a	No	No	1	10	100	20	0.388	0.010
5b	Yes	Yes	1	10	100	20	0.000	0.005
6a	No	No	2.5	10	100	20	0.139	0.054
6b	Yes	Yes	2.5	10	100	20	0.017	0.038
7a	No	No	1	$Z(4)^{\P,\|}$	100	20	0.048	0.050
7b	Yes	No	1	$Z(4)^{\P,**}$	100	20	0.015	0.030

*The standard error for a simulation with a true 0.05 level is 0.0010.

†Intraclass correlation in the Y variable is zero ('no') or generated as described in the text with $\gamma = 0.7$ ('yes').

‡Intraclass correlation generated in the X variable as described in the text.

§For each cluster, all of the observations are in only one of the four categories, i.e. the intraclass correlation for X is 1.

¶The continuous variable Z in model (1) is tested rather than the categorical variable (which has four categories).

‖Intraclass correlation generated in the Z variable is 0.

**Intraclass correlation generated in the Z variable is generated as described in the text with intraclass correlation of 0.49.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324 DOI: 10.1002/sim

Simulations #4a and 4b consider the scenario with 1000 clusters with two observations per cluster. The generalized score test has an inflated level for this situation. Simulations #5 and 6 consider the situation where $C = 10$ categories are being tested. For simulations #5 the marginal response probability remains 1 per cent, whereas for simulations #6 it is 2.5 per cent hence; the expected number of events per category (five) remains the same as the $C = 4$ category case. For these simulations, the level of the generalized score test is highly inflated when there is intraclass correlation and can be highly conservative when there is intraclass correlation. Our proposed test has very conservative levels for simulation #5 and a conservative level for simulation #6b. The last pair of simulations in Table II (no. 7) addresses the situation in which the regression coefficient of the continuous covariate Z is being tested, controlling for the categorical variable. Both the generalized score test and the proposed test are conservative when there is intraclass correlation in Z and Y.

In Tables I and II, all cluster sizes within a simulation are the same. Table III examines the performance of the test when the cluster sizes are not the same. In each simulation there are two sizes of clusters. In simulations 8a–e, the cluster sizes vary from being similar, 20 and 15 observations per cluster, to being very different, 20 and 2 observations per cluster. In simulations 8a–e and 10a, the proportion of large and small cluster sizes varies dramatically (10 and 90 per cent), and in simulations 9 and 10b the proportions are approximately 50–50. In the table, columns 8 and 9 show the results for a 1 per cent event rate and columns 10 and 11 show the results for a 2 per cent event rate. Simulations with a 2 per cent event rate follow the patterns we have seen previously, that is, the simulated method approximately maintains the nominal level or is conservative while the generalized score can have an inflated level. For the 1 per cent event rate, the proposed method breaks down when there is wide variation in the cluster sizes (8c, 20 and 5, or 8d, 20 and 2 observations per cluster). The inflated level seems to be related to the poor estimation of γ (results not shown).

The results given in Tables I–III suggest that the proposed test is superior to the other tests based on the maintenance of level. However, one could argue that these simulations are not a good test of the proposed test since the correlation structure of the simulated data is in the set of correlation structures assumed by our model. Therefore, to test the robustness of our method to a correlation misspecification, we repeated simulations #3c, 3d, 4b, 5b, 6b, and 7b (Table II) but with the intraclass correlation of Y being zero for half of the clusters and as before ($\gamma = 0.7$) for the other half. These are simulations #3e, 3f, 4c, 5c, 6c, and 7c, respectively, in Table IV. For simulations #3e, 3f, and 4c, the results look similar to the situation in which the intraclass correlation was zero (#3a for #3e and 3f, and #4a for #4c). For simulation #5c, the results of the generalized score test and the proposed test were similar and less conservative than when the correlation was always positive (simulation #5b). Both tests are anticonservative for simulation #6c, with the generalized score test being further from the nominal level. Having seen the simulation #5c results being conservative, and the #6c results being anticonservative, simulation #11 was performed to ensure that with the increasing number of events the simulated levels do approach the nominal level. For simulation #7c, the results look similar to the situation when the intraclass correlation was zero (simulation #7a). In summary, the results of Tables I–IV suggest that the proposed test is the best of the tests considered for preserving level, although there is no guarantee that the nominal level will be preserved.

To examine the power of the proposed test, we compare it with the power of the generalized score statistic when that test maintains the nominal level. In particular, we consider cases in Table V with $C = 4$, $m = 20$, intraclass correlation in Y, and $P(Y = 1)$ equal to 2 per cent (simulation no. 12) or

Table III. Simulated levels (nominal level 0.05) using the generalized score test and the proposed test of the null hypothesis in which the response probabilities $[P(Y=1)]$ are constant across C categories in model (1) with two types of cluster sizes.

	No. of clusters and no. of observations per cluster in the first type of cluster			No. of clusters and no. of observations per cluster in the second type of cluster			Simulated levels*			
							1 per cent event rate		2 per cent event rate	
Simulation number	Target per cent of total clusters	No. of clusters	No. of observations	Target per cent of total clusters	No. of clusters	No. of observations	Generalized score test	Proposed test	Generalized score test	Proposed test
3d and 4d	100	100	20	—	—	—	0.021	0.010	0.033	0.040
8a	10	13	20	90	116	15	0.011	0.014	0.016	0.026
8b	10	18	20	90	164	10	0.011	0.029	0.026	0.037
8c	10	31	20	90	276	5	0.025	0.078	0.060	0.057
8d	10	52	20	90	480	2	0.059	0.097	0.041	0.046
8e	—	—	—	100	1000	2	0.109	0.060	0.076	0.051
9a	50	58	20	50	56	15	0.012	0.015	0.017	0.024
9b	50	67	20	50	66	10	0.011	0.026	0.017	0.038
9c	50	80	20	50	80	5	0.012	0.038	0.018	0.037
9d	50	91	20	50	90	2	0.017	0.042	0.021	0.034
10a	10	91	4	90	818	2	0.110	0.064	0.078	0.050
10b	50	333	4	50	334	2	0.103	0.057	0.088	0.051

The total number of clusters is varied so that the number of observations is always maintained at 2000. In each simulation $\gamma = 0.7$, and the intraclass correlation in X is 1.

*The standard error for a simulation with a true 0.05 level is 0.0010.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324 DOI: 10.1002/sim

Table IV. Simulated levels (nominal level 0.05) using the generalized score test and the proposed test of the null hypothesis in which the response probabilities $[P(Y=1)]$ are constant across C categories in model (1) with k clusters and m observations per cluster.

Simulation number	Intraclass correlation		Marginal	Number of	Number of	No. of observations	Simulated levels*	
	Y†	X‡	$P(Y=1)$ (per cent)	categories (C)	clusters (k)	per cluster (m)	Generalized score test	Proposed test
3e	Yes/no	Yes	1	4	100	20	0.083	0.057
3f	Yes/no	1§	1	4	100	20	0.071	0.050
4c	Yes/no	Yes	1	4	1000	2	0.080	0.052
5c	Yes/no	Yes	1	10	100	20	0.026	0.029
6c	Yes/no	Yes	2.5	10	100	20	0.169	0.061
11	Yes/no	No	5	10	100	20	0.059	0.051
7c	Yes/no	No	1	$Z(4)$¶,‖	100	20	0.038	0.045

*The standard error for a simulation with a true 0.05 level is 0.0010.
†Intraclass correlation generated in the Y variable is zero for half the clusters and as described in the text with $\gamma = 0.7$ for the other half of the clusters.
‡Intraclass correlation generated in the X variable as described in the text ('yes') or is 1.
§For each cluster, all of the observations are in only one of the four categories, i.e. the intraclass correlation for X is 1.
¶The continuous variable Z in model (1) is tested rather than the categorical variable (which has four categories).
‖Intraclass correlation generated in the Z variable is generated as described in the text with the intraclass correlation of 0.49.

Published in 2007 by John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table V. Simulated powers (for nominal level 0.05 tests) using the generalized score test and the proposed test of the null hypothesis in which the response probabilities $[P(Y=1)]$ are constant across $C=4$ categories in model (1) with $k=300$ clusters and $m=20$ observations per cluster.

Simulation number	Intraclass correlation		Marginal $P(Y=1)$ (per cent)	$P(Y=1)$ for the four categories	Simulated powers	
	Y*	X†			Generalized score test	Proposed test
12a	Yes	No	2	3.73 per cent, 3@1.42 per cent	0.809	0.831
12b	Yes	Yes	2	3.73 per cent, 3@1.42 per cent	0.583	0.653
12c	Yes	Yes	2	4.27 per cent, 3@1.24 per cent	0.806	0.848
13a	Yes	No	1	2.40 per cent, 3@0.53 per cent	0.791	0.792
13b	Yes	Yes	1	2.40 per cent, 3@0.53 per cent	0.568	0.646
13c	Yes	Yes	1	2.80 per cent, 3@0.40 per cent	0.792	0.836

*Intraclass correlation in the Y variable is zero ('no') or generated as described in the text with $\gamma=0.7$ ('yes').
†Intraclass correlation in the X variable is zero ('no') or generated as described in the text ('yes').

1 per cent (simulation no. 13). For the alternative hypothesis under which the data are generated, the marginal probability of one of the cells is set larger than the other three as designated in Table V. To achieve reasonable powers, the number of clusters k was set to 300 rather than 100 as used in Table I. Going from simulation #12a to 12b (and from 13a to 13b) shows the loss of power incurred with the addition of intraclass correlation in X when there is intraclass correlation in Y. On the basis of the limited results in Table V, the power of the proposed test appears to be at least as good as the power of the generalized score statistic. Simulated levels for the tests under the associated null hypotheses in Table V are $\leqslant 5$ per cent for both tests in all cases, with the generalized score test being more conservative (results not shown).)

In order to examine the properties of the test when sample weights were used, a limited number of simulations with noninformative weights were performed. We considered the extreme case of a 1 per cent marginal event rate. In the simulations, the data were generated as previously described. Weights were generated such that $\frac{1}{6}$ of the observations were randomly assigned weights of 5 and the other $\frac{5}{6}$ were assigned weights of 1. The noninformative weights made the generalized score test more conservative as compared with the simulation with no weights. For example, in the simulation with no weights and no correlation in Y and X (simulation 3a), the level was 0.080. In an analogous simulation with noninformative weights, the level was 0.066. In a simulation with no weights and correlation in Y and X (simulation 3c), the level was 0.023. When noninformative weights were included, the level was 0.006. The proposed simulation method moved closer to the nominal level, simulation 3a went from 0.051 to 0.050, and simulation 3c went from 0.019 to 0.029.

Computational problems as described in Section 3.2 did not appear to be an issue. In the 'original' simulated data sets ('original' is used to differentiate data sets that were generated under specified parameters for the simulation from data sets generated for the null distribution), the proposed simulated method had very few data sets with no events (maximum of 0.89 per cent, in simulation 3d). In the original data sets, the maximum number of times there were convergence problems was 8 out of 50 000 replications.

The data sets generated for the null distribution had more data sets with no events and data sets with convergence problems, especially when there was intraclass correlation in the outcomes. In

the extreme case of a 1 per cent event rate and intraclass correlation in the Y variable, 66 per cent of the simulations had at least 1 of the 999 null data sets with either no events or a convergence problem. Of the null distributions where at least one test statistic could not be calculated, the median number of times a test statistic could not be calculated was 4. For a 2 per cent event rate, the percent of null distributions that had at least one test statistic that could not be calculated decreased to 6 per cent. The bootstrap simulations were similar with regard to data sets with no events and data sets with convergence problems.

4. AN APPLICATION USING NHANES III

The previously introduced NHANES III survey has a complex multistage sample design. In particular, at the first stage of sampling, there was a stratified selection of 81 clusters, known as primary sampling units (PSUs), consisting of counties or sometimes two or more adjacent smaller counties. Additional stages of sampling involve a second stage of sampling segments consisting of city or suburban blocks or other contiguous geographic areas from sampled PSUs, a third stage of sampling households from the sampled segments, and a fourth stage of sampling individuals from sampled households. Sampled individuals have an associated sample weight, which represents the number of people in the population that he or she represents. As part of their analyses, Arif *et al*. [2] used a logistic regression model for the binary outcome asthma, defined by a positive answer to both the following questions in NHANES III: 'Has a doctor ever told you that you had asthma?' and 'Do you still have asthma?' All people who indicated that they had emphysema were excluded. There were 742 positive binary outcomes in the data set analyzed. As is typically carried out with the survey data, the PSUs (the 'ultimate cluster') were used as clusters for the variance estimation. The independent variables are given in Table VI, along with the odds ratios estimated from the weighted logistic regression. The unweighted Oman–Zucker parameter is estimated to be $\hat{\gamma} = 0.19$. For each variable, the p-value calculated using the generalized score statistic as well as the proposed method is given. In Table VI, the p-values from our proposed method are at least as large as from the generalized score statistics, except for smoking status.

Along with testing individual covariates in the logistic regression model, it is also of interest to test the goodness-of-fit test of the logistic regression model. A test, based on deciles of risk, can be formulated as testing logistic regression coefficients [21]. This is an important application for the proposed method since in this context, by design, the lower decile categories will have the fewest positive events. When there are few positive events overall the lower deciles will have even fewer events, thus potentially causing problems when testing the logistic regression coefficients using asymptotic methods.

The test is formulated as follows: the predicted probabilities of an event are divided into 10 categories based on their deciles. Nine dummy variables representing the 10 risk categories are added to the model in question. The goodness of fit is tested by simultaneously testing that the corresponding nine regression coefficients are zero.

Table VII presents the goodness-of-fit table for the model given in Table VI. The p-value associated with the goodness of fit of this model is $p = 0.10$ using a generalized score statistic. The last column is the standard error of the difference in the expected and observed probabilities based on the jackknife estimate of the standard error for complex sample surveys [22]. To apply the proposed test, we first use the model fit under the null hypothesis (that the nine dummy variables are zero) to estimate the intraclass correlation parameter γ. Second, binary outcome data are

Table VI. Estimated odds ratios and 95 per cent confidence intervals from multiple logistic regression analysis of asthma using NHANES III (following Arif *et al.* [2].

Variable (number of asthma cases/sample size for category)	Adjusted odds ratio (95 per cent confidence interval)*	*p*-value	
		Generalized score test	Proposed test
Intercept = −4.50			
Age category		0.59	0.62
20–29 (134/3738)	1.61 (1.09, 2.40)		
30–39 (154/3555)	1.29 (0.91, 1.83)		
40–49 (134/2746)	1.56 (1.03, 2.37)		
50–59 (97/1974)	1.29 (0.85, 1.95)		
>59 (223/6094)	1.00†		
Race/ethnicity		0.007	0.018
Non-Hispanic White (341/7708)	1.00†		
Non-Hispanic Black (254/4950)	1.06 (0.85, 1.33)		
Mexican-American (122/4754)	0.60 (0.44, 0.83)		
Other (25/695)	0.82 (0.47, 1.40)		
Sex		0.017	0.022
Male (267/8436)	1.00†		
Female (475/9671)	1.44 (1.06, 1.97)		
Poverty income ratio‡		0.039	0.052
Below poverty (183/3738)	1.41 (1.04, 1.92)		
At or above or missing Poverty (559/14369)	1.00†		
Education		0.002	0.005
<12 years (439/10754)	1.64 (1.25, 2.17)		
High school (303/7353)	1.00†		
Year house built		0.127	0.154
Before 1946 (194/3969)	1.32 (1.02, 1.70)		
After 1946 (467/12221)	1.00†		
Missing (81/1917)	0.97 (0.67, 1.41)		
Smoking status		0.298	0.296
Nonsmokers (349/9042)	1.00†		
Past smokers (188/4460)	1.19 (0.82, 1.70)		
Current smokers (205/4605)	1.26 (0.94, 1.71)		
Pet ownership		0.0120	0.0124
No (289/5825)	1.00†		
Yes (453/12282)	1.35 (1.08, 1.68)		
Hay fever		<0.001	0.001
No (256/1626)	1.00†		
Yes (486/16481)	5.52 (4.26, 7.15)		
Body Mass Index		0.012	0.008
<18.5 (18/400)	1.00 (0.44, 2.27)		
18.5–24.9 (255/6913)	1.00†		
25.0–29.9 (209/6325)	0.90 (0.64, 1.27)		
⩾30 (260/4469)	1.64 (1.19, 2.26)		

*Odds ratios are adjusted for all the other variables in the table.

†Reference category.

‡The poverty income ratio is the ratio of the family income to the poverty threshold. The threshold was based on the age of the family reference person and the calendar year of the family interview. Poverty threshold values (in dollars) are produced annually by the U.S. Census Bureau.

Table VII. Deciles-of-risk goodness-of-fit table for logistic regression model given in Table VI; goodness-of-fit p-value = 0.18 using the proposed method.

Decile	Sample size	Number of events	Estimated proportion of population in category (per cent)	Weighted observed proportion of Asthma (per cent)	Weighted mean of predicted probabilities (per cent)	Difference (per cent)	Standard error of difference (per cent)*
1	2504	34	10.0	1.07	1.36	−0.29	0.27
2	2162	56	10.1	2.21	1.81	0.41	0.39
3	1627	31	9.9	2.65	2.13	0.52	0.89
4	1723	46	10.0	2.05	2.46	−0.41	0.47
5	1671	39	10.0	2.65	2.79	−0.14	0.55
6	1729	44	10.0	2.27	3.22	−0.95	0.52
7	1806	66	10.0	4.82	3.78	1.04	0.73
8	1774	89	10.1	5.24	4.70	0.54	0.76
9	1755	105	9.9	5.48	7.03	−1.55	0.78
10	1356	232	10.0	17.01	16.20	0.81	0.32

*Based on jackknifed standard error estimation.

Table VIII. p-value for 'hay fever' variable and goodness-of-fit test from analysis in Table VI performed on ten subsets of the data set obtained by a random partition of the original data set.

Random subset number	p-value for hay fever variable		p-value for goodness-of-fit	
	Generalized score test	Proposed test	Generalized score test	Proposed test
1	<0.001	0.006	0.273	0.367
2	<0.001	0.022	0.194	0.407
3	0.003	0.016	0.326	0.476
4	0.019	0.045	0.040	0.148
5	<0.001	0.009	0.328	0.500
6	0.006	0.039	0.165	0.381
7	<0.001	0.004	0.011	0.038
8	0.040	0.084	0.087	0.192
9	0.001	0.018	0.614	0.812
10	0.001	0.006	0.999	1.000
Original data set	0.004	0.009	0.101	0.183

simulated using the fixed set of covariate values in Table VI (excluding the nine dummy variables) as well as sample weights and PSU designations. Intraclass correlation in the binary variable is generated using the estimated γ. The generalized score statistic for testing the nine dummy variables, calculated using the simulated data sets, are then used as the reference distribution for the generalized score statistic calculated using the original data set. Note that the dummy variables are not recalculated for each simulated data set (which we calculated using the predicted values for that simulated data set), although that is a possibility. The p-value calculated for goodness of fit using the proposed method is $p = 0.18$. As with other goodness-of-fit tests, (1) a rejection of the null hypothesis should be followed with a careful examination of the pattern of observed and expected events (like Table VII) to see if there is a substantive misfit and (2) nonrejection of the null hypothesis may be due to the lack of power of such omnibus tests.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

To examine the relationship of the score and proposed p-values further, we split the original data set randomly into 10 subsets and fit the same logistic regression model to each subset. Table VIII presents the p-values for subsets for hay fever variable (which was the most significant variable in Table VI) and the goodness-of-fit p-values. Again, the score statistic gives smaller p-values than the proposed method. The smaller p-values for the proposed method in the 10 subsets of data are consistent with the simulation results of Section 3, which show that the generalized score statistic can be anticonservative.

5. DISCUSSION

We have assumed that the parameters in the logistic regression model (1) are the parameters of interest. With respect to the clustering, this is considered a marginal approach since the clustering does not appear in the model (for the mean response). For some applications, the clusters are an intrinsic part of the model, and it is desired to know whether the outcome is associated with the independent variables controlling for cluster. The intercept in model (1) would then be indexed by the cluster. Conditional logistic regression can be used for analyzing data where the effect of the cluster is modeled. In the conditional logistic regression model, one conditions on the number of positive outcomes in each cluster. One can perform asymptotic tests on the parameters of interest or use an 'exact' conditional logistic regression approach, which makes no asymptotic assumptions [23]. We suspect that the small-sample behavior of test statistics for conditional logistic regression, conditioning on the clusters, may be better than for unconditional logistic regression because the conditioning would remove the effects of clustering and speed up the asymptotics to be closer to those of simple random samples.

Another possibility for analyzing data where the cluster is of interest, models the intercepts (clusters) with a random effect. The effect of few positive outcomes on this type of modeling would need investigation. It is important to note that the marginal and cluster-specific models are not interchangeable; they may yield different conclusions and the choice of model should be made on substantive grounds [24–26].

With the marginal model that is the focus of this paper, we have found that although the generalized score statistic performs better than the generalized Wald statistic, its actual significance level can be higher than its nominal level when there are few positive outcomes. Without clustering, it is sometimes possible to perform an 'exact' unconditional logistic regression analysis [27]. However, based on results for estimating a simple proportion [28, 29] such exact methods will not be useful in marginal logistic regression applications with clustering.

The proposed simulation approach to testing suggested in this paper is to simulate the observed data under a reasonable model that involves generating correlated data according to the Oman–Zucker method and use a test statistic that would work well with large samples regardless of the model. We chose to use the Oman–Zucker model because it is one of the few models that permit correlated binary data within clusters whose probabilities can vary by individual while maintaining the population expectation to follow a specified marginal model, e.g. a logistic regression model. There are other available models for generating correlated data with these properties [30–33], but they can be more computationally intensive.

We present a modification of the simulation-based approach that is applicable to data with sample weights such as sample survey data. Even though the simulation approach with sample weights generates outcomes in the simulated data sets without using the sample weights (treating

the weights as noninformative), it produces test statistics that have correct large sample properties for informative sample weighting. However, our approach of generating data with noninformative weights for the null distribution may result in small sample bias when the weights are informative in the original data. Our results show that the simulation approach performed well for unweighted data and noninformative weighted data for the logistic regression model, but further research would be useful for extending our simulations to informative weighted data.

APPENDIX

We use a modification of the approach of Oman and Zucker [17] to estimate their intraclass correlation model parameter γ. First, for all distinct pairs of observations within clusters, the pairwise-squared difference $W_{ijk} = (Y_{ij} - Y_{ik})^2$ is calculated for all clusters i, and individuals $j \neq k$ in cluster i. Let π_{is} be the probability that $Y_{is} = 1$, and let $\hat{\pi}_{is}$ be the estimator of this probability from the fit of the null hypothesis model. (Oman and Zucker [17] are only considering a single model so that there is no null hypothesis model in their formulation.) Using only those pairs of observations such that the predicted $\pi_{ij} \leqslant \pi_{ik}$, Oman and Zucker [17] note that

$$E(W_{ijk}) = f_{ijk} + h_{ijk}\gamma^2$$

where

$$f_{ijk} = \pi_{ij}(1 - \pi_{ik}) + \pi_{ik}(1 - \pi_{ij})$$

and

$$h_{ijk} = 2\pi_{ij}(\pi_{ik} - 1)$$

Thus, a linear regression can be used for estimating γ^2 if we substitute $\hat{\pi}_{is}$ for π_{is} in the expressions for f_{ijk} and h_{ijk}, and γ can be estimated by the square root of the estimated γ^2. Oman and Zucker [17] suggest using a weighted linear regression with the weight for observation W_{ijk} being the inverse of $f_{ijk}(1 - f_{ijk})$. (Note that this weighting is not related to the sample weights of survey data.) In this paper, we use an unweighted linear regression.

Table AI shows that in our context of small event rates the unweighted linear regression leads to a more accurate estimator of γ than the weighted regression. This is likely due to the variability in the predicted probabilities. Table AI includes simulations with event response rates of 1, 2, and 20 per cent. Data are generated according to model (1) with η the coefficient for the continuous covariate set at 1, 2.59, and 3.5 and the true value of $\gamma = 0.7$. For event rates of 1 and 2 per cent the median of the unweighted estimator is much closer to 0.7 and the 10th and 90th percentiles show increased variability with the weighted estimator. With a 20 per cent event rate there is little difference in the difference between the performance of the weighted estimates and the unweighted estimator.

We applied the method described in Oman and Zucker [17] to generate binary outcomes that are correlated within clusters of observations. For each cluster i, generate e_i^0 and for each observation j in cluster i generate e_{ij}, where e_i^0 and e_{ij} are independent random variates from a standard normal distribution. Generate U_{ij} according to a Bernoulli distribution with parameter γ and let $\theta_{ij} = \Phi^{-1}(\hat{\pi}_{ij})$ for each observation j in cluster i, and $\hat{\pi}_{ij}$ is the estimated probability of a positive outcome based on the set of covariates and estimated regression coefficients. Generate the binary

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

Table AI. Simulated percentiles of the estimators of γ using weighted (Oman–Zucker) and unweighted linear regressions.

		True value of α	True value of η	Percentiles of estimates of γ (true $\gamma = 0.7$)				
				10 per cent	25 per cent	50 per cent	75 per cent	90 per cent
1 per cent event rate	Weighted	−5.11	1	0.000	0.229	0.649	0.760	0.823
		−6.11	2.59	0.000	0.389	0.648	0.766	0.849
		−6.81	3.50	0.000	0.436	0.679	0.802	0.903
	Unweighted	−5.11	1	0.000	0.308	0.675	0.768	0.825
		−6.11	2.59	0.219	0.524	0.690	0.773	0.831
		−6.81	3.50	0.307	0.553	0.695	0.783	0.847
2 per cent event rate	Weighted	−4.43	1	0.226	0.551	0.678	0.745	0.789
		−5.44	2.59	0.289	0.528	0.669	0.751	0.807
		−6.10	3.50	0.233	0.531	0.679	0.764	0.832
	Unweighted	−4.43	1	0.344	0.588	0.687	0.749	0.792
		−5.44	2.59	0.499	0.618	0.696	0.756	0.801
		−6.10	3.50	0.524	0.625	0.700	0.760	0.806
20 per cent event rate	Weighted	−1.91	1	0.655	0.676	0.698	0.717	0.734
		−2.84	2.59	0.646	0.672	0.697	0.720	0.739
		−3.43	3.50	0.637	0.667	0.695	0.721	0.743
	Unweighted	−1.91	1	0.657	0.677	0.698	0.717	0.733
		−2.84	2.59	0.657	0.678	0.699	0.719	0.736
		−3.43	3.50	0.655	0.676	0.698	0.719	0.738

Data are generated according to model (1) with the β's $= 0$.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2008; **27**:1305–1324
DOI: 10.1002/sim

outcomes as $Y_{ij}=1$ if $(U_{ij}e_i^0+(1-U_{ij})e_{ij})\leqslant\theta_{ij}$ and $Y_{ij}=0$ otherwise. This method for generating the binary outcomes Y_{ij} preserves its marginal distribution, i.e. $E(Y_{ij})=\pi_{ij}$.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers for their thorough review and insightful comments.

All simulations were performed on the high-performance Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

REFERENCES

1. Mannino DM, Homa DM, Pertowski CA, Ashizawa A, Nixon LL, Ball LB, Jack E, Kang DS. Surveillance for asthma—United States, 1960–1995. CDC Surveillance Summaries, 24 April 1998. *Morbidity and Mortality Weekly Report* 1998; **47**(SS-1):1–27.
2. Arif AA, Delclos GL, Lee ES, Tortolero SR, Whitehead LW. Prevalence and risk factors of asthma and wheezing among US adults: an analysis of the NHANES III data. *European Respiratory Journal* 2003; **21**(5):827–833.
3. National Center for Health Statistics. Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–94. *Vital and Health Statistics* 1994; **1**(32):20–22.
4. Graubard BI, Korn EL, Midthune D. Testing goodness-of-fit for logistic regression with survey data. *American Statistical Association Proceedings of the Section on Survey Research Methods*, Alexandria, VA, 1997; 170–174.
5. Barnard GA. Discussion of 'The spectral analysis of point processes' by Bartlett, M.S. *Journal of the Royal Statistical Society*, *Series B* 1963; **25**:264–296.
6. Rao JNK, Scott AJ, Skinner CJ. Quasi-score tests with survey data. *Statistical Sinica* 1998; **8**:1050–1070.
7. Kotz S, Johnson NL, Read CB. Wald's W-statistics. *Encyclopedia of Statistical Sciences*, vol. 9. Wiley: New York, 1988; 525–526.
8. Tarone RE. Score statistics. *Encyclopedia of Statistical Sciences*, vol. 8. Wiley: New York, 1988; 304–308.
9. Strawderman WE. Likelihood ratio tests. *Encyclopedia of Statistical Sciences*, vol. 4. Wiley: New York, 1988; 647–650.
10. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 1983; **51**:279–292.
11. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990; **77**:485–497.
12. Boos DD. On generalized score tests. *The American Statistician* 1992; **46**:327–333.
13. Horton NJ, Bebchuk JD, Jones CL, Lipsitz SR, Catalano PJ, Zahner GEP, Fitzmaurice GM. Goodness-of-fit for GEE: an example with mental health service utilization. *Statistics in Medicine* 1999; **18**:213–222.
14. Thomas DR, Rao JNK. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association* 1987; **82**:630–636.
15. Korn EL, Graubard BI. Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t statistics. *American Statistician* 1990; **44**:270–276.
16. Sherman M, Le Cessie S. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics—Simulation and Computation* 1997; **26**:901–925.
17. Oman SD, Zucker DM. Modelling and generating correlated binary variables. *Biometrika* 2001; **88**(1):287–290.
18. Rao JNK. Marginal models for repeated observations: inference with survey data. *ASA Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, Alexandria, VA, 1998; 76–82.
19. Scott AJ, Holt D. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 1982; **77**:848–854.
20. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978; **108**:100–102.
21. Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 1980; **67**:250–251.
22. Korn EL, Graubard BI. *Analysis of Health Surveys*. Wiley: New York, 1999; 30.
23. Mehta CR, Patel NR, Senchaudhuri P. Efficient Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association* 2000; **95**:99–108.
24. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**:1049–1060.

25. Neuhaus JM. Estimation efficiency and tests of covariate effects with clustered binary data. *Biometrics* 1993; **49**:989–996.
26. Graubard BI, Korn EL. Regression-analysis with clustered data. *Statistics in Medicine* 1994; **13**(5–7):509–522.
27. Mehta CR, Patel NR, Gray R. Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. (Corr: vol. 81, 1132). *Journal of the American Statistical Association* 1985; **80**:969–973.
28. Gross ST, Frankel MR. Confidence limits for small proportions in complex samples. *Communications in Statistics*: *Theory and Methods* 1991; **20**:951–975.
29. Korn EL, Graubard BI. Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology* 1998; **24**:193–201.
30. Emrich LJ, Peidmonte MR. A method for generating high-dimensional multivariate binary variates. *American Statistician* 1991; **45**:302–304.
31. Lee AJ. Generating random binary deviates having fixed marginal distributions and specified degrees of association. *American Statistician* 1993; **49**:209–215.
32. Gange SJ. Generating multivariate categorical variates using the iterative proportional fitting algorithm. *American Statistician* 1995; **49**:134–138.
33. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 2003; **90**(2):455–463.