Using marginal mean models with data from a longitudinal survey having a complex design: some advances in methods

Authors:
Georgia Roberts, Statistics Canada (robertg@statcan.ca)
Qunshu Ren, Carleton University and research student at Statistics Canada (qren@math.carleton.ca)
J.N.K. Rao, Carleton University (jrao34@rogers.com)

1. INTRODUCTION

In recent years, longitudinal surveys, where sample subjects are observed over two or more time points, are being undertaken by government agencies in order to provide longitudinal data for analytic studies and to aid in the development of public policy. At Statistics Canada, for example, the National Population Health Survey (NPHS), the National Longitudinal Survey of Children and Youth (NLSCY) and the Survey of Labour and Income Dynamics (SLID) were all launched in the mid 1990’s for this purpose and now have available several cycles of data on the same samples of individuals. Data from such surveys can be used for a variety of purposes including (a) gross flows estimation, (b) event history modeling, (c) conditional modeling of the response at a given time point as a function of past responses and present and past co-variables, and (d) modeling of marginal means of responses as functions of co-variables. Binder (1998) gives an excellent account of the possible uses of longitudinal survey data.

Because of the repeated interviewing of the same individuals, longitudinal surveys typically lead to dependent observations over time. Furthermore, longitudinal surveys often have complex sampling designs that involve clustering, which results in cross-sectional dependencies among subjects. Because of this additional complexity, new methods must be developed to replace the methods of analysis used for longitudinal data where subjects are independent. In this paper, we focus on some aspects of marginal mean modeling with survey data, in particular binary responses and marginal logistic regression models. The first problem that we discuss here is the estimation of the model parameters. The case of a simple random sample, where individuals are considered to be independent and to have equal chances of being selected, has been studied extensively in the literature, especially in applications in biomedical and health science. Liang and Zeger (1986) used generalized estimating equations (GEE) to estimate the model parameters and their variances, assuming a “working” correlation structure for the repeated measurements. The GEE method can be readily adapted to the case of a complex survey design, with the model parameters being estimated as the solution to survey-weighted estimating equations.

An odds ratio is one measure of association between a pair of binary responses. Lipsitz, Laird and Harrington (1991) used odds ratios to model the working covariance structure in GEE, instead of working correlation coefficients used by Liang and Zeger (1986) and others. It makes more sense to use odds ratios rather than correlations with binary responses and is easier to interpret. Moreover, the odds ratio approach is less seriously constrained than the correlation coefficient approach (Liang, Qaqish and Zeger, 1992). We demonstrate in Section 2 how this approach may be extended to the case of complex survey data.
2. Survey-weighted Estimating Equations (SEE) and Odds Ratio Approach

Suppose a sample, \( s \), of size \( n \), is selected by a complex survey design from a population \( U \) of size \( N \), and that the same sampled units are observed for \( T \) occasions. Let the data have the form \( \{(y_{it}, x_{it}), i \in s, t = 1, \ldots, T\} \), where \( y_{it} \) is the response of the \( ith \) individual on occasion \( t \) and \( x_{it} \) is a \( p \times 1 \) vector of fixed covariates associated with \( y_{it} \). In the case of a binary response variable (i.e. \( y_{it} = 0 \) or \( 1 \)), the marginal logistic regression model is a natural choice for describing the relationship between \( y_{it} \) and \( x_{it} \). In the marginal logistic regression model, the marginal density of response \( y_{it} \) given covariate \( x_{it} \) is the Bernoulli density with \( E(y_{it}) = p_{it} \), where

\[
\text{logit} (p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) = \beta_0 + x_{it}'\beta_1. \tag{1}
\]

Assuming independence between sample individuals (or simple random sampling with negligible sampling fraction \( n / N \)), an estimator of the vector of model parameters \( \beta = (\beta_0, \beta_1)' \) is obtained as the solution of the Generalized Estimating Equations (GEE):

\[
\hat{u}(\beta) = \sum_{i \in s} D_i^t V_i^{-1} (y_i - p_i(\beta)) = 0, \tag{2}
\]

where \( y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})' \), \( p_i(\beta) = (p_{i1}, p_{i2}, \ldots, p_{iT})' \), \( D_i = \partial p_i(\beta)/\partial \beta \), and \( V_i \) is the “working” covariance matrix of \( y_i \) (Liang and Zeger, 1986). Note that \( V_i \) is the identity matrix under a working independence assumption for the observations from the \( i \)’th individual, or is a positive definite matrix under a working correlation assumption. It should be kept in mind that, while...
V_i may differ from the true covariance matrix of y_i, we assume that the mean of y_i is correctly specified, i.e. E(y_i) = p_i(\beta).

In the case of a complex survey design, let the final survey weights be \{w_i, i \in s\}. The weights \{w_i, i \in s\} may represent calibration weights that satisfy post-stratification constraints. Rao (1998) proposed the following survey-weighted estimating equations (SEEI) for estimating \beta:

\[ \hat{u}_i(\beta) = \sum_{i=s} w_i D_i^{-1}(y_i - p_i(\beta)) = 0. \] (3)

Denote the solution of (3) as \hat{\beta}_w. Note that \hat{\beta}_w is a survey-weighted estimator of the census parameter, \beta_N, which is the solution of the census estimating equations

\[ u_N(\beta) = \sum_{i=1} D_i^{-1}(y_i - p_i(\beta)) = 0. \]  The census parameter \beta_N would be a consistent estimator of \beta if the population, U, of individuals is regarded as a self-weighting sample from a superpopulation obeying the marginal model. The survey-weighted estimator, \hat{\beta}_w, is consistent for \beta_N (and hence for \beta) if \hat{u}_i(w) is design-unbiased or consistent for \beta_N. We assume that \sqrt{n}(\hat{\beta}_w - \beta) is small relative to \sqrt{n}(\hat{\beta}_w - \beta_N) so that \sqrt{n}(\hat{\beta}_w - \beta) = \sqrt{n}(\hat{\beta}_w - \beta_N), and thus it is not necessary to distinguish \beta from \beta_N.

In the case of a marginal model with binary responses, Lipsitz et. al (1991) used the odds ratio as a measure of association between pairs of binary responses. The major reason for this choice is that the odds ratio is not constrained by the means of the two binary variables, like the correlation. As well, we can use a working model for the odds ratios to define V_i. If we let y_{ist} = y_{ist} for all s = 1,..., T - 1, t = s + 1,...T, and p_{ist} = E(y_{ist}) = Pr(y_{ist} = 1, y_{ist} = 1), then for given s≠t, the odds ratio \gamma_{ist} is defined as:

\[ \gamma_{ist} = \frac{p_{ist}(1-p_{ist})}{p_{ist}(1-p_{ist})}. \] (4)

Suppose that the odds ratio \gamma_{ist} is modeled as a function of covariates (e.g., log odds ratio is a linear function of some covariates), and that \alpha is the vector of parameters in that model, i.e. \gamma_{ist} = \gamma_{ist}(\alpha). Then the elements of the working covariance matrix V_i can be written:

\[ Var(y_{ist}) = V_{ist} = p_{ist}(1-p_{ist}) \]
\[ Cov(y_{ist}, y_{ist}) = V_{ist}(\beta, \alpha) = p_{ist}(\beta, \alpha) - p_{ist}(\beta)p_{ist}(\beta) \] (5)

where, from the quadratic equation (4), p_{ist} can be expressed as p_{ist} = g(\gamma_{ist}, p_{ist}, p_{ist}), which is a function of both \alpha and \beta.
Since $\alpha$ and $\beta$ are both unknown, we need to use a second set of survey-weighted estimating equations (SEEII). Let $u_i = (y_{i12}, \ldots, y_{i(T-1)T})'$ and

$$\theta_i(\beta, \alpha) = (p_{i12}(\beta, \alpha), p_{i13}(\beta, \alpha), \ldots, p_{i(T-3)T}(\beta, \alpha))'.$$

Then SEEII are given by:

$$\hat{u}_{2w}(\beta, \alpha) = \sum_{i \in s} w_i C_i F_i^{-1} [u_i - \theta_i(\beta, \alpha)] = 0,$$

(6)

where $C_i = \frac{\partial \theta'}{\partial \alpha}$ and $F_i = \text{diag}\{p_{iu}(1 - p_{iu})\}$. A Newton Raphson type iterative method may be used to solve (3) and (6) simultaneously using initial values $\hat{\alpha}_0, \hat{\beta}_0$. Iterations are given by

$$\hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} - \left(\sum_{i \in s} w_i D'_i(\hat{\beta}_{(m)}) F_i^{-1} D_i(\hat{\beta}_{(m)})\right)^{-1} \left[\sum_{i \in s} w_i D'_i(\hat{\beta}_{(m)}) V_i^{-1} (y_i - p_i(\hat{\beta}_{(m)}))\right]$$

(7)

and

$$\hat{\alpha}_{(m+1)} = \hat{\alpha}_{(m)} - \left(\sum_{i \in s} w_i C'_i(\hat{\alpha}_{(m)}) F_i^{-1} C_i(\hat{\alpha}_{(m)})\right)^{-1} \left[\sum_{i \in s} w_i C'_i(\hat{\alpha}_{(m)}) F_i^{-1} (u_i - \theta_i(\hat{\alpha}_{(m)}), \hat{\beta}_{(m+1)})\right],$$

(8)

where the subscripts $(m)$ and $(m + 1)$ indicate that quantities are evaluated at $\beta = \hat{\beta}_{(m)}$ and $\alpha = \hat{\alpha}_{(m)}$ in (7) and at $\beta = \hat{\beta}_{(m+1)}$ and $\alpha = \hat{\alpha}_{(m)}$ in (8). At convergence of the iterations, we obtain $\hat{\beta}$ and $\hat{\alpha}$, where $\hat{\beta}$ is a consistent estimator of $\beta$ even under misspecification of the means of the $u_i$. We assume that $\hat{\alpha}$ converges in probability to some $\alpha^*$ which agrees with $\alpha$ only when the working model $\gamma_{iu} = \gamma_{iu}(\alpha)$ is correctly specified.

3. Variance Estimation: One Step EF-Bootstrap

In order to make inferences from an estimated marginal model, an estimator of the covariance matrix of the estimated model parameters is required. Assuming independence between sample individuals, Liang and Zeger (1986) used Taylor linearization to derive consistent sandwich-type estimators which are widely used. Sandwich-type estimators of the covariance matrix have also been developed, through linearization, for many analytical problems applied to survey data. See, for example, Binder (1983) for an application of this approach to generalized linear models and Rao, Scott and Skinner (1998) for this approach in developing Wald and quasi-score tests. However, as the forms of parameter estimates become more complex and as nonresponse and calibration adjustments to survey weights become more involved, such as in the case of some longitudinal analyses, it becomes more difficult to carry out a full linearization. Because of this difficulty, attention has turned to studying replication methods for design-based variance estimation. As examples, Rao, Yung and Hidiroglou (2002) and Rao and Tausi (2004) have proposed jackknife and bootstrap re-sampling approaches for variance estimation. For many of its analytical surveys, Statistics Canada is now releasing design information for variance estimation only in the form of survey bootstrap weights.
The direct bootstrap method for variance estimation (see, for example, Rust and Rao, 1996) involves obtaining point estimates of the parameters of interest with the full-sample survey weights, \( w_i \), and then, in an identical fashion, with each set of survey bootstrap weights. This method, consisting of many repetitive operations, can be computationally intensive and time consuming. Furthermore, Binder, Kovacevic and Roberts (2004) found that, when using this approach for logistic regression, it was possible to have many sets of bootstrap weights for which the parameter estimation algorithm would not converge due to ill-conditioned matrices that were not invertible. To overcome these problems, Binder, Kovacevic and Roberts (2004) and Rao and Tausi (2004) proposed estimating function (EF) bootstrap approaches, motivated by the work of Hu and Kalbfleish (2000) for the non-survey case. Here, we extend the one-step EF bootstrap approach of Rao and Tausi (2004) to the marginal logistic regression model.

Let \{\( w_i^{(b)} \), \( i \in s \)\}, for \( b = 1, ..., B \) be B sets of bootstrap weights for the sample \( s \). Let

\[
\hat{u}^{(b)}_{iw}(\hat{\beta}) = \sum_{i \in s} w_i^{(b)} D_i V_i^{-1} [y_i - p_i(\hat{\beta})]
\]

and

\[
\hat{u}^{(b)}_{2w}(\hat{\alpha}, \hat{\beta}) = \sum_{i \in s} w_i^{(b)} C_i F_i^{-1} [y_i - \theta_i(\hat{\beta}, \hat{\alpha})],
\]

be the corresponding bootstrap estimating functions evaluated at \( \beta = \hat{\beta} \) and \( \alpha = \hat{\alpha} \) where \( \hat{\beta} \) and \( \hat{\alpha} \) are obtained from (7) and (8). Now compute one-step Newton-Raphson solutions to the following EF equations using \( \hat{\beta} \) and \( \hat{\alpha} \) as starting values:

\[
\hat{u}^{(b)}_{iw}(\beta) = \hat{u}^{(b)}_{iw}(\hat{\beta})
\]

\[
\hat{u}^{(b)}_{2w}(\alpha, \beta) = \hat{u}^{(b)}_{2w}(\hat{\alpha}, \hat{\beta}).
\]

This is equivalent to Taylor linearization of the left hand sides of (11) and (12). The one-step bootstrap estimators for the \( b \)-th bootstrap sample are given by:

\[
\tilde{\beta}^{(b)} = \hat{\beta} - (\sum_{i \in s} w_i \hat{D}_i V_i^{-1} \hat{D}_i)^{-1} \hat{u}^{(b)}_{iw}(\hat{\beta})
\]

\[
\tilde{\alpha}^{(b)} = \hat{\alpha} - (\sum_{i \in s} w_i \hat{C}_i F_i^{-1} \hat{C}_i)^{-1} \hat{u}^{(b)}_{2w}(\hat{\alpha}, \hat{\beta}),
\]

where the matrices \( \hat{D}_i, \hat{V}_i, \hat{C}_i \), and \( \hat{F}_i \) in (13) and (14) are obtained by evaluating \( D_i, V_i, C_i \), and \( F_i \) at \( \hat{\beta} \) and \( \hat{\alpha} \).

Note that for all \( b = 1, ..., B \), the inverse matrices in (13) and (14) remain the same, so that further inversion for each bootstrap sample is not needed. Moreover, there are no convergence problems since we iterate only once. The Rao-Tausi EF-bootstrap estimator of \( \text{cov}(\hat{\beta}) \) is given by
Diagonal elements of (15) give the variance estimators while the covariance estimators are obtained from the off-diagonal elements. The estimator (15) is algebraically identical to the bootstrap estimator proposed by Binder, Kovacevic and Roberts (2004), which is obtained by equating \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \), to \(-\hat{u}_{1w}(\hat{\beta})\) and \(-\hat{u}_{2w}(\hat{\alpha}, \hat{\beta})\) respectively and then using one-step Newton-Raphson iteration.

\[
\sum_{b=1}^{B} \frac{1}{B} (\widehat{\beta}(b) - \hat{\beta})(\widehat{\beta}(b) - \hat{\beta})'.
\]

4. Goodness-of-fit tests

Inferences on regression parameter \( \beta \), outlined in Sections 2 and 3, assume that the mean of \( y_i \) is correctly specified as \( E(y_i | x_i) = p_{i\alpha}(\beta) \), with \( \logit p_{i\alpha}(\beta) = \beta_0 + x_i' \beta_1 \). It is therefore important to check the adequacy of the mean specification using goodness-of-fit tests, before making inferences on \( \beta \). Assuming independence among sample individuals, several goodness-of-fit tests have been proposed in the literature for the logistic regression model. In particular, for the case of no repeated measurements on a sample individual, the well-known Hosmer and Lemeshow (1980) test uses a Pearson chi-squared statistic after partitioning the subjects into groups based on the values of estimated response probabilities under the assumed model for the mean. Horton et al. (1999) proposed a score test for the case of longitudinal data after partitioning all the estimated probabilities \( \hat{p}_{i\alpha} = p_{i\alpha}(\hat{\beta}) \) into groups under the assumed marginal model and working independence.

The above tests are generally not applicable to complex survey data because of cross-sectional dependencies among subjects. For the case of no repeated measurements, Graubard et al. (2004) proposed a Wald test that takes account of survey weights \( w_i \) and the cross-sectional dependencies. This test uses the survey weights \( w_j \) and the associated estimates of response probabilities to partition the subjects into groups. Graubard et al. (2004) also proposed a simulated Wald test, based on a parametric bootstrap method, that performed better than the Wald test in controlling the Type I error rate under the absence of sample selection bias (i.e., the superpopulation model holds for the sample).

In this section, we extend the Horton et al. (1999) score test to the case of longitudinal survey data and show that it is equivalent to an extension of the Graubard et al. (2004) Wald test to the longitudinal case.

4.1 Construction of groups

In the case of no repeated measurements with data \((y_i, x_i; i \in s)\), Graubard et al. (2004) obtained weighted decile groups \( G_1, G_2, \ldots, G_{10} \): \( n_1 \) subjects with the smallest estimated probabilities \( \hat{p}_i = p_i(\hat{\beta}) \) under the null hypothesis \( H_0 : p_i = p_i(\beta) \), are in the first group \( G_1 \), the next \( n_2 \) in
the second group $G_2$, and so forth until the group $G_{10}$ with $n_{10}$ observations is formed. The $n_i$ observations in $G_i$ are chosen such that $\sum_{i \in G_i} w_i/\sum_{i \in G_i} w_i = 1/10$; $i = 1, 2, \ldots, 10$. In the special case of equal weights $w_i = w$, this grouping method reduces to the Hosmer and Lemeshow (1980) method of grouping the subjects.

We now extend the above weighted decile grouping method to the case of longitudinal survey data, following Horton et al. (1999). We make use of all the estimated probabilities $\dot{\pi}_i = p_i(\beta)$, under the null hypothesis $H_0: p_i = p_i(\beta)$ with $\text{logit } p_i(\beta) = \beta_0 + x_i' \beta$. Let $I_{it} = 1$ if $\dot{\pi}_i$ is in group $G_i$ and $I_{it} = 0$ otherwise, $i = 1, \ldots, 9$. The alternative model is then given by

$$\text{logit } p_i(\beta) = \beta_0 + x_i' \beta + \gamma_1 I_{i1} + \cdots + \gamma_9 I_{i9}$$ (16)

We treat the indicator variables as fixed covariates even though they are based on the random $\pi_i$. In the case of independence among sample individuals, Moore and Spruill (1975) provided asymptotic justification for treating the partition $\{G_1, G_2, \ldots, G_{10}\}$ as though based on the true $p_i$.

Under the set-up (16), our $H_0$ is equivalent to testing $H_0^* : \gamma_1 = \gamma_2 = \ldots = \gamma_9 = 0$.

Following Rao et al. (1998), we now develop a quasi-score test of $H_0^*$, taking account of the weights and the design. We assume working independence and obtain survey-weighted estimating equations under (16) as

$$\hat{u}_w(\beta, \gamma) = \begin{bmatrix} \hat{u}_{1w}(\beta, \gamma) \\ \hat{u}_{2w}(\beta, \gamma) \end{bmatrix} = \begin{bmatrix} \sum_{i \in s} \sum_{t=1}^T w_i x_i \left( y_{it} - p_i(\beta; \gamma) \right) \\ \sum_{i \in s} \sum_{t=1}^T w_i I_i \left( y_{it} - p_i(\beta; \gamma) \right) \end{bmatrix}$$ (17)

where $I_{it} = (I_{i1}, \ldots, I_{i9})'$ and $\gamma = (\gamma_1, \ldots, \gamma_9)'$. Under $H_0^*$, we solve $\hat{u}_{1w}(\beta, 0) = 0$ to get the estimator $\hat{\beta}_w$ of $\beta$. Note that $\hat{\beta}_w$ is identical to $\hat{\beta}_w$ obtained from (3) under working independence. We substitute $\hat{\beta}_w$ into the second component $\hat{u}_{2w}(\beta, \gamma)$ of (17) to get $\hat{u}_{2w}(\hat{\beta}_w, 0)$ under $H_0^*$. We have $\hat{u}_{2w}(\beta, 0) = o_w - e_w$, where $o_w = \sum_{i \in s} \sum_{t=1}^T w_i I_i y_{it}$ and $e_w = \sum_{i \in s} \sum_{t=1}^T w_i I_i \hat{\pi}_i$, where
\( \hat{p}_n = p_n(\hat{\beta}_w,0) = p_n(\hat{\beta}_w) \) and \( p_n(\beta) \) is the null model. Note that \( o_w \) and \( e_w \) are the weighted observed and expected counts respectively, under \( H^*_0 \).

A quasi-score statistic for testing \( H^*_0 \) is now given by

\[
X^2_{QS} = \hat{u}_{2w}(\hat{\beta}_w,0)\left[\text{var}(\hat{u}_{2w}(\hat{\beta}_w,0))\right]^{-1}\hat{u}_{2w}(\hat{\beta}_w,0),
\]

where \( \text{var}(\hat{u}_{2w}(\hat{\beta}_w,0)) \) is a consistent estimator of variance of \( \hat{u}_{2w}(\hat{\beta}_w,0) \) under \( H^*_0 \). For the NPHS data (Section 5) we have used a bootstrap variance estimator, using the Rao and Wu (1988) bootstrap weights to calculate the bootstrap estimates \( \hat{u}^{(b)}_{2w} = \hat{u}_{2w}(\hat{\beta}^{(b)}_w,0), b = 1,...,B \), where \( B \) is the number of bootstrap replicates and \( \hat{\beta}_w^{(b)} \) is obtained from

\[
\hat{u}^{(b)}_{2w}(\beta,0) = \sum_{s=1}^{s} \sum_{i=1}^{I} w_i^{(b)}(y_{it} - p_u(\beta)) = 0,
\]

where \( w_i^{(b)}, b = 1,...,B \) are the bootstrap weights. To simplify the computations we propose a one-step Newton-Raphson iteration with \( \hat{\beta}_w \) as starting value to obtain \( \hat{\beta}_w^{(b)} \). The bootstrap variance estimator of \( \hat{u}_{2w}(\hat{\beta}_w,0) \) is then given by

\[
v^{\text{BOOT}}(\hat{u}_{2w}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{u}_{2w}^{(b)} - \hat{u}_{2w})(\hat{u}_{2w}^{(b)} - \hat{u}_{2w})^\prime.
\]

Since the NPHS data file provides the weights \( \{w_i, i \in s\} \) and the bootstrap weights \( \{w_i^{(b)}, i \in s, b = 1,...,B\} \) along with the response variables, it is straightforward to implement \( v^{\text{BOOT}}(\hat{u}_{2w}) \) from the NPHS data file. The EF method of Section 4 was designed for variance estimation of regression parameter estimates \( \hat{\beta}_w \), and further theoretical work is needed to study its applicability to quasi-score tests. For the NPHS data (Section 5), we did not encounter ill-conditioned matrices in calculating \( \hat{\beta}_w^{(b)} \); therefore, we used the Rao-Wu bootstrap method to obtain \( \hat{u}_{2w}^{(b)} \) needed in \( v^{\text{BOOT}}(\hat{u}_{2w}) \) given by (20). An alternative to (20) is obtained by substituting \( \hat{u}_{2w}^{(b)} = \frac{1}{B} \sum_{b=1}^{B} \hat{u}_{2w}^{(b)} \) for \( \hat{u}_{2w} \) in (20), but the result should be close to (20).

We can write \( X^2_{QS} \) as a Wald statistic based on \( o_w - e_w \):

\[
X^2_{QS} = (o_w - e_w)\left[\text{v}(o_w - e_w)\right]^{-1}(o_w - e_w),
\]

where \( \text{v}(o_w - e_w) \) is a consistent variance estimator of \( o_w - e_w \). Graubard et al. (2004) proposed a Wald statistic similar to (21) for the case of no repeated measurements. Under \( H^*_0 \), the statistic \( X^2_{QS} \) is asymptotically distributed as a \( \chi^2 \) variable with 9 degrees of freedom. Hence, the p-value may be obtained as \( \Pr(\chi^2_9 \geq X^2_{QS}(\text{obs})) \), where \( X^2_{QS}(\text{obs}) \) is the observed value of \( X^2_{QS} \). The p-value could provide evidence against \( H^*_0 \).

### 4.3 Adjusted Hosmer-Lemeshow test
We now consider a survey-weighted version of the Hosmer-Lemeshow (HL) chi-squared statistic for testing the null hypothesis $H_0$ in the context of longitudinal data, and then adjust the statistic to account for the survey design effect following the method of Rao and Scott (1981). For $l = 1, \ldots, 10$, let

\[
\hat{p}_{nl} = \left( \sum_{i \in s_l} \sum_{t=1}^{T} w_{it} I_{itl} \right) \left( \sum_{i \in s_l} \sum_{t=1}^{T} w_{it} \right)^{-1},
\]

\[
\hat{x}_{nl} = \left( \sum_{i \in s_l} \sum_{t=1}^{T} w_{it} I_{itl} \hat{x}_{nl} \right) \left( \sum_{i \in s_l} \sum_{t=1}^{T} w_{it} \right)^{-1}, \text{ and}
\]

\[
\hat{W}_l = \left( \sum_{i \in s_l} \sum_{t=1}^{T} w_{it} I_{itl} \right) \left( T \sum_{i \in s_l} w_i \right)^{-1}.
\]

Then the survey-weighted, longitudinal data version of the HL statistic is given by

\[
X_{HL}^2 = nT \sum_{l=1}^{10} \hat{W}_l \left( \hat{p}_{nl} - \hat{x}_{nl} \right)^2 \hat{W}_{nl} (1 - \hat{x}_{nl}).
\]  

(22)

If the weights $w_i$ are equal and $T = 1$, then $X_{HL}^2$ reduces to the HL statistic for the case of one time point and simple random sampling. In the latter case, Hosmer and Lemeshow (1980) approximated the null distribution by a $\chi^2$ with $10 - 2 = 8$ degrees of freedom, $\chi^2_8$. This result is not applicable in the survey context and we adjust (22) using the Rao-Scott (1981) method.

Rao and Scott (1981) proposed a first-order correction and a more accurate second-order correction to a chi-squared statistic for categorical data. Roberts, Rao and Kumar (1987?) applied the method to logistic regression of estimated cell proportions, and their results are now applied to our grouped proportions $(\hat{p}_{nl}, \hat{x}_{nl}), l = 1, \ldots, 10$. Let $\hat{p}_w = (\hat{p}_{w1}, \ldots, \hat{p}_{w10})$, $\hat{x}_w = (\hat{x}_{w1}, \ldots, \hat{x}_{w10})$ and $\hat{V}_w$ be the estimated covariance matrix of $\hat{p}_w - \hat{x}_w = \hat{r}_w$. We used the Rao-Wu bootstrap method to get an estimator $\hat{V}_w$ using the one-step Newton-Raphson iteration. Let $\hat{p}_w^{(b)}$ be the $b$th bootstrap version of $\hat{p}_w$ and $\hat{x}_w^{(b)}$ be the corresponding bootstrap version of $\hat{x}_w$ using $\hat{p}_w^{(b)}$ obtained from (22). Then,

\[
v_{BOOT}(\hat{r}_w) = \frac{1}{B} \sum_{b=1}^{B} \left( \hat{p}_w^{(b)} - \hat{r}_w \right) \left( \hat{p}_w^{(b)} - \hat{r}_w \right)^\prime
\]

(23)

is a bootstrap estimator of $\text{cov}(\hat{r}_w)$.

The Rao-Scott (RS) first-order correction uses the factor

\[
\hat{\delta}_1 = \frac{1}{9} \left[ (nT) \sum_{l=1}^{10} \hat{V}_{r, l} (\hat{V}_r (1 - \hat{x}_r))^{-1} \right],
\]

(24)

which represents a generalized design effect, where $\hat{V}_{r, l}$ is the $l$th diagonal element of $\hat{V}_r$. The first-order adjusted HL statistic is then given by

\[
X_{RS}^2(1) = X_{HL}^2 / \hat{\delta}_1,
\]

(25)

which is treated as $\chi^2_8$ under $H_0$. Since the choice of degrees of freedom is not clear-cut, we may follow what is done with the Horton et al. score statistic and treat (25) as $\chi^2_9$ instead of $\chi^2_8$.

We need new theory on the asymptotic null distribution of (25).
$X_{R_2}^2$ (1) takes account of the survey design, but not as accurately as the second-order adjusted statistic which requires the knowledge of the off-diagonal elements $\hat{V}_{r,m}$ of $\hat{V}_r$, $l \neq m$. We need the factor

$$\hat{a}^2 = \hat{\delta}^2 \sum_{i=1}^{10} \left( \frac{1}{9} \sum_{j=1}^{9} \left( \hat{\delta}_j \right)^2 \right)$$

where

$$\sum_{i=1}^{10} \hat{\delta}_i^2 = \sum_{i=1}^{10} \sum_{m=1}^{10} \hat{V}_{r,m} \left( nTW_i \right) \left( nTW_m \right) \pi_i \pi_m (1 - \pi_i)(1 - \pi_m).$$

The second-order adjusted HL statistic is given by

$$X_{R_2}^2 (2) = X_{R_2}^2 (1)/\left( 1 + \hat{a}^2 \right),$$

which is treated as $\chi^2$ with degrees of freedom $8/(1 + \hat{a}^2)$ or $9/(1 + \hat{a}^2)$ under $H_0$. Note that $X_{R_2}^2 (2) \approx X_{R_2}^2 (1)$ if $\hat{a}^2 \approx 0$.

5. Example from NPHS data

We applied the marginal logistic regression model and the EF bootstrap to data from Statistics Canada’s National Population Health Survey (NPHS). The NPHS began in 1994/95, and collects information every two years from the same sample of individuals. A stratified multistage design was used to select households within clusters, and then one household member 12 years or older was chosen to be the longitudinal respondent. The longitudinal sample consists of 17,276 individuals. Currently, 5 cycles of data are available.

Motivated by the research of Shields and Shooshtari (2001) who used NPHS data and logistic regression in order to study the relationship between a self-perceived health measure and various socio-economic, lifestyle, physical and psycho-social health variables, we formulated a marginal logistic regression model for $T=2$ occasions. We took the same sample of 5380 females who were 25+ years of age at the time of sample selection, were respondents in all of the first three cycles of the survey and did not have proxy responses to the health component of the questionnaire. For occasion $t$, our binary response variable $y_{it}$ is 1 if self-perceived health of the $i$'th individual at time $t$ is excellent or very good and is 0 if self-perceived health at time $t$ is good, fair or poor. The associated vector of covariates $x_{it}$ consists of 41 dichotomous variables similar to those used by Shields and Shooshtari. Some of the covariates describe the status of the individual at the previous survey cycle, while other covariates describe changes in status between the previous and current survey cycles. For our example, occasion $t=1$ is 1996/97 (so that data from both 1994/95 and 1996/97 are used to generate $x_{it}$) and occasion $t=2$ is 1998/99 (so that data from both 1996/97 and 1998/99 are used to generate $x_{i2}$). Survey weights $\{w_i, i \in s\}$ appropriate for respondents to the first three cycles of NPHS were chosen, along with $B=500$ sets of bootstrap weights $\{w_i^{(b)}, i \in s\}, b = 1, \ldots, B$.

Parameter estimates and standard errors

We used the following approaches for estimating the model parameters:
1. SEE-Ind: SEE with a working independence assumption;
2. SEEII-ORconstant: SEE with a constant odds ratio model for the SEE working covariance structure
3. SEEII-ORf(age): SEE with a working odds ratio modeled as a function of an individual’s age group by \( \log (\gamma_t) = \alpha_0 + \alpha_1 * a_t + \alpha_2 * a_t^2 \), where \( a_t = 1 \) for age 25-34, \( a_t = 2 \) for age 35-44, \( ..... \), \( a_t = 5 \) for age 65-74, and \( a_t = 6 \) for age >75.

In approaches 2 and 3, a second set of estimating equations,\( \hat{\beta}_2 = \alpha \), was used to estimate the unknown parameters \( \alpha \) associated with the working odds ratios. Another option is to use empirical odds ratios to estimate \( \alpha \) directly from the data, so that \( \hat{\beta}_2 = \alpha \) is not needed. The following two approaches use this option:
4. SEE- ORconstant-E: empirical constant odds ratio;
5. SEE- ORf(age)-E: empirical constant odds ratio within each age group.

For all approaches we fitted the marginal logistic regression model (1) using the 41 covariates \( x_{it} \). The one-step EF approach was used to obtain variance estimates.

Table 1 illustrates the coefficient estimates and associated standard errors under the five different approaches for two of the binary covariates, “functionally restricted (yes/no)” and “heavy smoker (yes/no)”, used in the logistic regression models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Functionally restricted</th>
<th>Heavy smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>S.E.</td>
</tr>
<tr>
<td>SEE-Ind</td>
<td>-1.13</td>
<td>0.14</td>
</tr>
<tr>
<td>SEEII-ORconstant</td>
<td>-0.99</td>
<td>0.13</td>
</tr>
<tr>
<td>SEEII-ORf(age)</td>
<td>-0.99</td>
<td>0.13</td>
</tr>
<tr>
<td>SEE- ORconstant-E</td>
<td>-0.94</td>
<td>0.12</td>
</tr>
<tr>
<td>SEE- ORf(age)-E</td>
<td>-0.94</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 1 shows that, for our example, the standard errors are quite similar under the four methods of modelling odds ratios. Also, the gain in efficiency (in terms of S.E.) over SEE-Ind is quite small in this example. A possible explanation for this small efficiency gain is that the two covariates are nearly cluster-specific in the sense of not changing from t=1 to t=2. In the case of a cluster-specific covariate, Fitzmaurice (1995) has shown that the efficiency gain over the working independence method is small. On the other hand, working independence can lead to a considerable loss of efficiency in estimating parameters associated with time varying (or within cluster) covariates.

**Goodness-of-fit tests**
We tested the goodness-of-fit of our logistic regression model with the 41 dichotomous covariates used by Shields and Shooshtari (2001). Using the quasi-score test (18) based on the bootstrap variance estimator (20), we obtained \( \chi^2_{Qs} = 6.57 \) and p-value = \( P(\chi^2 \geq \chi^2_{Qs}) = 0.682 \), which suggests that there is no evidence against the assumed logistic regression model. Note that the p-value is calculated under the framework of the alternative model (16) based on weighted decile groups.

We now turn to the HL chi-squared statistic and its Rao-Scott adjustments for design effect. For the survey weighted HL statistic (22), we obtained \( \chi^2_{HL} = 10.11 \), but its null distribution is not asymptotically \( \chi^2 \) or \( \chi^2 \). Hence, we use the Rao-Scott adjustments. For the first-order version (25), we obtained \( \delta = 1.49 \), \( \chi^2_{RS} = 6.78 \) and corresponding p-values \( P(\chi^2 \geq 6.78) = 0.56 \) and \( P(\chi^2 \geq 6.78) = 0.66 \), suggesting that there is no evidence against the assumed logistic regression model. Note that \( \chi^2_{HL} \) is substantially larger than \( \chi^2_{RS} \) because it ignores the cross-sectional dependencies. Turning to the more accurate second-order version (28), we obtained \( \delta = 0.16 \), \( \chi^2_{RS} = 6.78/1.16 = 5.86 \) and degrees of freedom \( 8/(1+\delta^2) = 6.9 \) or \( 9/(1+\delta^2) = 7.8 \), with corresponding p-values of 0.54 and 0.64 respectively. The above p-values again suggest no evidence against the assumed logistic regression model.

5. SUMMARY

This paper shows how the marginal logistic regression model for binary responses – widely used in biostatistical research – can be extended to the case of design-based analysis of complex survey data. Estimation of the model through survey-weighted estimating equations using odds ratios for describing the working covariance matrix is illustrated. Also, the one-step EF bootstrap approach to variance estimation is extended to this model. Methods for assessing the goodness of fit of the marginal model, taking account of the survey design, are also given. Finally, longitudinal data from Statistics Canada’s National population Health Survey are used to illustrate the methods proposed in this paper.

REFERENCES


