

# Marginal Logistic Regression Models for Longitudinal Complex Survey Data

Qunshu Ren and Georgia Roberts<sup>1</sup>

## ABSTRACT

In this paper we study some problems associated with extending marginal logistic modeling to data from a longitudinal survey with a complex design. Design-weighted Generalized Estimating Equations (GEE) are used to estimate the model parameters. Odds ratios are used as measures of association between pairs of binary responses in the working covariance matrix. A one-step estimating function (EF) bootstrap method is used for variance estimation. The methods are illustrated through their application to data from Statistics Canada's National Population Health Survey.

KEY WORDS: Bootstrap, Estimating equations, GEE, Logistic regression, Longitudinal surveys, Odds ratio

## RÉSUMÉ

Dans cet article, nous étudions quelques problèmes associés à la généralisation de la modélisation logistique marginale à des données d'enquêtes longitudinales avec un plan de sondage complexe. Les Équations Estimantes Généralisées Pondérées selon le plan de sondage (EEG) sont utilisées pour estimer les paramètres du modèle. Le rapport des cotes est utilisé comme une mesure d'association entre les paires de réponses binaires dans la matrice de covariance de travail. La méthode de la fonction d'estimation bootstrap en une seule étape (FE) est utilisée pour l'estimation de la variance. Les méthodes sont illustrées par application à des données de l'enquête nationale de la santé et de la population (ENSP) de Statistique Canada.

MOTS CLÉS : bootstrap, équations estimantes, EEG, modèles logistiques, enquêtes longitudinales, rapport des cotes

## 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.

---

<sup>1</sup>Carleton University and Statistics Canada, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada Email: [qren@math.carleton.ca](mailto:qren@math.carleton.ca)

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.

Standard errors of estimated model parameters are used in inference. In the non-survey case, Liang and Zeger (1986) proposed a sandwich type variance estimator which they used in Wald tests, while Rotnitzky and Jewell (1990) developed quasi-score tests. For the case of complex survey data, where the non-independence among individuals must be accounted for, the variance estimation techniques must be modified. Rao (1998) explained how to obtain an appropriate sandwich-type estimator for the case of complex survey data, but, in practice, this can be difficult to carry out when linearization to account for post-stratification and nonresponse adjustments is required. As an alternative, a design-based bootstrapping method becomes a good choice as it is simple and robust. As well, Statistics Canada is releasing design information for variance estimation only in the form of survey bootstrap weights for many of its analytical surveys. In Section 3 we will follow the work of Rao and Tausi (2004) to extend the EF-bootstrap method developed by Hu and Kalbfleisch (2000) to the situation of the marginal logistic regression model.

To illustrate the methods discussed, longitudinal data from Statistics Canada's National Population Health Survey are used in an example in Section 4.

## 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  $\{(y_{it}, x_{it}), i \in s, t = 1, \dots, T\}$ , where  $y_{it}$  is the response of the  $i$ 'th individual on occasion  $t$  and  $x_{it}$  is a  $p \times 1$  vector of associated covariates. 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,

$$f(y_{it} | x_{it}) = p_{it}^{y_{it}} (1 - p_{it})^{(1-y_{it})}, \quad (1)$$

where  $E(y_{it} | x_{it}) = p_{it}$ ,  $\text{logit}(p_{it}) = X'_{it}\beta$ , and  $X_{it} = (1, x'_{it})'$ . Let  $X_i = \{X_{i1}, X_{i2}, \dots, X_{iT}\}$ ,  $y_i = \{y_{i1}, y_{i2}, \dots, y_{iT}\}'$ ,  $p_i(\beta) = \{p_{i1}, p_{i2}, \dots, p_{iT}\}'$ , and  $V_i$  be the "working" covariance matrix of  $y_i$ . Assuming independence between sample individuals (or simple random sampling with negligible sampling fraction  $n/N$ ), an estimator of the vector of model parameters  $\beta$  is obtained as the solution of the Generalized Estimating Equations (GEE):

$$\hat{u}(\beta) = \sum_{i \in s} D_i V_i^{-1} (Y_i - p_i(\beta)) = 0, \quad (2)$$

where  $D_i = \frac{\partial p_i(\beta)}{\partial \beta}$  (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 survey weights be  $\{w_i, i \in s\}$ . Rao (1998) proposed the following survey-weighted estimating equations (SEEI) for estimating  $\beta$ :

$$\hat{u}_{1w}(\beta) = \sum_{i \in S} w_i D_i' V_i^{-1} (Y_i - p_i(\beta)) = 0. \quad (3)$$

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  $u_N(\beta) = \sum_{i \in U} D_i' V_i^{-1} (Y_i - p_i(\beta)) = 0$ .  $\beta_N$  would be a consistent estimator of  $\beta$  if the population, U, of individuals is 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  $\hat{u}_{1w}(\beta)$  is design-unbiased or consistent for  $u_N(\beta)$ . We assume that  $n/N$  is negligible so that  $\sqrt{n}(\hat{\beta}_w - \beta) \approx \sqrt{n}(\hat{\beta}_w - \beta_N)$ , 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 is that the odds ratio is not constrained by the means of the two binary variables, which is a problem with the correlation. As well, we can use a working model for the odds ratios to define  $V_i$ . If we let  $Y_{ist} = Y_{is} Y_{it}$  for all  $s=1, \dots, T-1$ ,  $t=s+1, \dots, T$ , and  $p_{ist} = E(Y_{ist}) = \Pr(Y_{is} = 1, Y_{it} = 1)$ , then for given  $s \neq t$ , the odds ratio  $\gamma_{ist}$  is defined as:

$$\gamma_{ist} = \frac{P(Y_{is} = 1, Y_{it} = 1)P(Y_{is} = 0, Y_{it} = 0)}{P(Y_{is} = 1, Y_{it} = 0)P(Y_{is} = 0, Y_{it} = 1)} = \frac{p_{ist}(1 - p_{is} - p_{it} + p_{ist})}{(p_{is} - p_{ist})(p_{it} - p_{ist})}. \quad (4)$$

Suppose that the odds ratio  $\gamma_{ist}$  is modeled as a function of covariates (e.g., the log odds ratio is the 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:

$$\begin{aligned} \text{Var}(Y_{it}) &= V_{itt} = p_{it}(\beta)(1 - p_{it}(\beta)) \\ \text{Cov}(Y_{is}, Y_{it}) &= V_{ist}(\beta, \alpha) = p_{ist}(\beta, \alpha) - p_{is}(\beta)p_{it}(\beta) \end{aligned} \quad (5)$$

where, from the quadratic equation (4),  $p_{ist}$  can be expressed as  $p_{ist} = g(\gamma_{ist}, p_{is}, p_{it})$ , 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 (SEII). Let  $U_i = (Y_{i12}, \dots, Y_{i(T-1)T})'$  and  $\theta_i(\beta, \alpha) = (p_{i12}(\beta, \alpha), p_{i13}(\beta, \alpha), \dots, p_{i(T-1)T}(\beta, \alpha))'$ . Then SEII are given by:

$$\hat{u}_{2w}(\beta, \alpha) = \sum_{i \in S} w_i C_i' F_i^{-1} [U_i - \theta_i(\beta, \alpha)] = 0, \quad (6)$$

where  $C_i = \frac{\partial \theta_i'}{\partial \alpha}$  and  $F_i = \text{diag}\{p_{ist}(1 - p_{ist})\}$ . The Newton Raphson iterative method may be used to solve (3) and (6) simultaneously using initial values  $\hat{\alpha}_0, \hat{\beta}_0$ , and then letting

$$\hat{\beta}_{(m+1)} = \hat{\beta}_{(m)} - \left( \sum_{i \in S} w_i D_{i(m)}' V_{i(m)}^{-1} D_{i(m)} \right)^{-1} \left\{ \sum_{i \in S} w_i D_{i(m)}' V_{i(m)}^{-1} (Y_i - p_i(\hat{\beta}_{(m)})) \right\} \quad (7)$$

and

$$\hat{\alpha}_{(m+1)} = \hat{\alpha}_{(m)} - \left( \sum_{i \in S} w_i C'_{i(m)} F_{i(m)}^{-1} C_{i(m)} \right)^{-1} \left\{ \sum_{i \in S} w_i C'_{i(m)} F_{i(m)}^{-1} (U_i - \theta_i(\hat{\alpha}_{(m)}, \hat{\beta}_{(m+1)})) \right\}. \quad (8)$$

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  $\gamma_{ist} = \gamma_{ist}(\alpha)$  is correctly specified.

### 3. Variance Estimation: One Step EF-Bootstrap

In order to make inferences from an estimated marginal model, (co)variance estimates are required for the estimated model parameters. Assuming independence between sample individuals, Liang and Zeger (1986) used linearization to derive consistent sandwich-type variance estimates which are widely used. Sandwich-type variance estimates 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 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  $\{ \{ w_i^{(b)}, i = 1, \dots, n \}, b = 1, \dots, B \}$  be  $B$  sets of bootstrap weights for the sample  $s$ . Let

$$\hat{u}_{1w}^{(b)}(\hat{\beta}) = \sum_{i \in S} w_i^{(b)} D_i V_i^{-1} \{ Y_i - p_i(\hat{\beta}) \} \quad (9)$$

and

$$\hat{u}_{2w}^{(b)}(\hat{\alpha}, \hat{\beta}) = \sum_{i \in S} w_i^{(b)} C_i' F_i^{-1} \{ U_i - \theta_i(\hat{\beta}, \hat{\alpha}) \}, \quad (10)$$

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:

$$\hat{u}_{1w}(\beta) = \hat{u}_{1w}^{(b)}(\hat{\beta}) \quad (11)$$

$$\hat{u}_{2w}(\alpha, \beta) = \hat{u}_{2w}^{(b)}(\hat{\alpha}, \hat{\beta}). \quad (12)$$

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:

$$\tilde{\beta}^{(b)} = \hat{\beta} - (\sum_{i \in S} w_i \hat{D}_i \hat{V}_i^{-1} \hat{D}_i)^{-1} \hat{u}_{1w}^{(b)}(\hat{\beta}) \quad (13)$$

$$\tilde{\alpha}^{(b)} = \hat{\alpha} - (\sum_{i \in S} w_i \hat{C}_i' \hat{F}_i^{-1} \hat{C}_i)^{-1} \hat{u}_{2w}^{(b)}(\hat{\alpha}, \hat{\beta}), \quad (14)$$

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  $b=1, \dots, B$ , the inverse matrices in (13) and (14) remain the same, so that no further inversion is needed for each bootstrap sample. Since we only iterate once, there are no convergence problems. The EF-bootstrap variance estimator of  $\hat{\beta}$  is given by

$$v_{BOOT}^{EF}(\hat{\beta}) = \frac{1}{B} \sum_{b=1}^B (\tilde{\beta}^{(b)} - \hat{\beta})(\tilde{\beta}^{(b)} - \hat{\beta})'. \quad (15)$$

#### 4. Application to 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 psychosocial 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  $x_{i1}$ ) and occasion  $t=2$  is 1998/99 (so that data from both 1996/97 and 1998/99 are used to generate  $x_{i2}$ ). A survey weight variable appropriate for respondents to the first three cycles of NPHS was chosen, along with a set of  $B=500$  bootstrap weights.

We used the following approaches to model our data:

1. Separate: Logistic models were fit separately to the data for each occasion – thus different  $\beta$ 's for each;
2. SEE-Ind: SEE with a working independence assumption;
3. SEEII-OR<sub>constant</sub>: SEE with a constant odds ratio model for the SEE working covariance structure
4. SEEII-OR<sub>f(age)</sub>: SEE with a working odds ratio modeled as a function of an individual's age group by

$$\log(\gamma_i) = \alpha_0 + \alpha_1 * a_i + \alpha_2 * a_i^2,$$

where  $a_i = 1$  for age 25-34,  $a_i = 2$  for age 35-44, ...,  $a_i = 5$  for age 65-74, and  $a_i = 6$  for age >75.

In approaches 3 and 4, a second set of estimating equations,  $\hat{u}_{2w}(\beta, \alpha) = 0$ , 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  $\hat{u}_{2w}(\beta, \alpha)$  is not needed. The following two approaches use this option:

5. SEE- OR<sub>constant</sub>-E: empirical constant odds ratio; and
6. SEE- OR<sub>f(age)</sub>-E: empirical constant odds ratio within each age group.

For approaches 2 to 6 we fitted the marginal logistic regression model assuming (i) separate  $\beta$ 's for each occasion (thus 82 coefficients to be estimated) and also assuming (ii) a common  $\beta$ . Values of  $\beta$  for (i) are denoted as Time1 and Time2 in Table 1. The one-step EF approach was used to estimate variances for approaches 2 to 6 while the direct survey bootstrap was used for approach 1.

Table 1 illustrates the coefficient estimates and associated standard errors under the 6 different approaches for two of the binary explanatory variables, namely “functionally restricted (yes/no)” and “heavy smoker (yes/no)” used in the logistic regression models.

Table 1. Coefficient estimates and their standard errors (in brackets)

Method	Functionally restricted			Heavy Smoker		
	Time1	Time2	Common $\beta$	Time1	Time2	Common $\beta$
Separate	-0.94 (0.18)	-1.37 (0.17)	-	-0.41 (0.15)	-0.28 (0.18)	-
SEE-Ind	-0.94 (0.18)	-1.37 (0.17)	-1.13 (0.14)	-0.41 (0.15)	-0.28 (0.18)	-0.32 (0.12)
SEEII-OR <sub>constant</sub>	-0.91 (0.16)	-1.23 (0.15)	-0.99 (0.13)	-0.43 (0.14)	-0.25 (0.16)	-0.34 (0.11)
SEEII-OR <sub>f(age)</sub>	-0.90 (0.16)	-1.23 (0.15)	-0.99 (0.13)	-0.43 (0.14)	-0.25 (0.16)	-0.34 (0.11)
SEE- OR <sub>constant</sub> -E	-0.90 (0.15)	-1.19 (0.15)	-0.94 (0.12)	-0.43 (0.14)	-0.24 (0.16)	-0.35 (0.11)
SEE- OR <sub>f(age)</sub> -E	-0.88 (0.15)	-1.20 (0.15)	-0.94 (0.12)	-0.42 (0.14)	-0.24 (0.16)	-0.35 (0.11)

Table 1 shows that, for our example, the standard errors are quite similar under the four different methods of modelling odds ratios. Standard errors under the working independence model (SEE-Ind) are slightly larger than the corresponding values under the odds ratio models when the logistic regression model uses separate  $\beta$ 's for each occasion. For example, the standard error for the variable “functionally restricted” under working independence (SEE-Ind) is 0.18 compared to 0.15 under empirical constant odds ratios within age groups (SEE- OR<sub>f(age)</sub>-E). Finally, for a given approach, the standard errors under the common  $\beta$  model are smaller than the corresponding standard errors under the separate  $\beta$ 's model. For example, for SEE- OR<sub>f(age)</sub>-E and the variable “heavy smoker”, the standard error under the common  $\beta$  model is 0.11 compared to 0.16 for Time2 under the separate  $\beta$ 's model. This reduction in standard error is achieved because the common  $\beta$  model borrows more strength from the two time points than the separate  $\beta$ 's model; the latter model uses both time points only through the covariance matrix for the two time points. Although we have not shown it here, we could test whether the common  $\beta$  model explains the data as well as the separate  $\beta$ 's model.

## 5. CONCLUSIONS

This paper shows how the marginal logistic regression model – 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 model, taking account of the survey design, are currently being researched.

## Acknowledgements

This work is supported by the internship program co-sponsored by NPCDS and Statistics Canada. Many thanks to:

Professor J. N. K. Rao & Professor M. Mojirsheibani, who are Qunshu's supervisors at Carleton University; and also to Dr. M. Kovacevic, M. Shields, and others in Statistics Canada who provided technical support.

## REFERENCES

- Binder, D.A. (1993). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51, 279-292.
- Binder, D.A. (1998). Longitudinal surveys. Why are these surveys different from all other surveys? *Survey Methodology*, 24, 101-108.
- Binder, D.A., Kovacevic, M. and Roberts, G. (2004). Design-based methods for survey data: alternative uses of estimating functions. *American Statistical Association 2004 Proceedings of the Survey Research Methods Section*.
- Hu, F. and Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canadian Journal of Statistics*, 28, 449-499.
- Liang, K.-Y. and Zeger, S. L.(1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K.-Y., Zeger, S., and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society*, B, 54, 3-40.
- Lipsitz, S. R., Laird, N. M. and Harrington, D. P.(1991). Generalized estimating equations for correlated binary data: using odds ratios as a measure of association, *Biometrika*, 78, 153-160.
- Rao, J.N.K. (1998) Marginal models for repeated observation: Inference with survey data. *American Statistical Association 1998 Proceedings of the Survey Research Methods Section*, 76-82.
- Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Rao, J. N. K., Scott, A. J. and Skinner, C. J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.
- Rao, J.N.K. and Tausi, M. (2004). Estimating function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics* 33, 9, 2087-2095.
- Rao, J. N. K., Yung, W. and Hidiroglou, M. A. (2002). Estimating equations for the analysis of survey data using post-stratification information. *Sankhya*, A, 64, 364-378.
- Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semi-parametric generalized linear models for cluster correlated data. *Biometrika*, 77, 485-497.
- Rust, K.F. and Rao, J.N.K.(1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* 5, 283-310.
- Shields, M and Shooshtari, S. (2001) Determinants of Self-perceived Health. *Health Reports*, 13, No. 1, 35-52.
- Sutraduhar, B. and Kovacevic, M. (2000). Analysing ordinal longitudinal survey data: generalized estimating equations approach. *Biometrika*, 87, 837-848.
- Ziegler, A., Kastner, C. and Chang-Claude, J. (2003). Analysis of pregnancy and other factors on detection of human papilloma virus (HPV) infection using weighted estimating equations for follow-up data. *Statistics in Medicine*, 22, 2217-2233.