

QUASI-SCORE TESTS WITH SURVEY DATA

J.N.K. Rao and A.J. Scott ¹

ABSTRACT

Although most survey texts are concerned primarily with problems of estimating finite population parameters, survey data are often used to develop and fit stochastic models describing the underlying structure of the population. In this paper we develop an analogue of the score test for use with survey data where the use of multi-stage sampling and variable selection probabilities cause special problems.

KEY WORDS: Estimating equations; Score test; Superpopulation parameters; Survey data.

RÉSUMÉ

Alors que la plupart des articles concernant les enquêtes traitent surtout de l'estimation des paramètres de population finies, les données d'enquête sont souvent utilisées pour élaborer et ajuster des modèles stochastiques qui décrivent la structure sous-jacente de la population. Le présent article a pour but de présenter le développement d'un analogue du test des scores applicable aux données d'enquête dans les cas où l'utilisation de l'échantillonnage à plusieurs degrés et des probabilités variables de sélection pose des problèmes particuliers.

MOTS CLÉS: Équations d'estimation; test des scores; paramètres des superpopulations; données d'enquête.

1. INTRODUCTION

In most conventional statistics courses, a clear distinction is made between sample survey methods on the one hand and the rest of applied statistics on the other. Traditional survey methods are concerned with estimating population means, totals and proportions, along with related quantities like ratios, while the rest of applied statistics concentrates on model-building for explanation, prediction and so on. In reality, many surveys (especially in the health and social sciences) are aimed at exploring relationships and building predictive models, just as in the rest of statistics. Surveys are conducted to find out what effect education has on unemployment or income, what factors affect crib deaths in infants or strokes in older people and so on. For example, the data shown in Table 1 comes from a stratified case control sample drawn from records of people under the age of 35 in northern Malawi given in Clayton and Hills (1993). Here "cases" are new cases of leprosy and Scar is a binary variable taking the value one if a person has a BCG vaccination scar and zero otherwise. The aim in this

study is to gain some insight into whether or not a BCG vaccination affects the chance of contracting leprosy rather than estimating population totals and proportions.

What makes the analysis of survey data different? One obvious problem is that, by their very nature, analytical surveys are observational studies and we are always faced with the difficulty of making causal inferences in situations where we have no control over the assignment of experimental treatments. There are two other major features that distinguish the analysis of survey data. The first is the correlation induced by the hierarchical structure of multi-stage sampling. Because of the cost, most large scale surveys are carried out in two or more stages. The lack of independence within primary sampling units (census blocks, doctors' practices, schools, households) means that standard errors, confidence intervals and *P*-values produced by standard computer packages are invalid. This is by no means unique to surveys, however. Many experimenters have to cope with correlation between repeated measurements on the same subject, siblings from the same litter, and so on. As an aside, it is

¹ J.N.K. Rao, Dept. of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6; A.J. Scott, Department of Statistics, University of Auckland, Auckland, New Zealand.

interesting to note that a number of techniques developed to handle survey data are starting to find uses in other areas that have to deal with correlated data (see Rao and Scott, 1991, for a simple example). Perhaps more important is the use of variable selection probabilities. If some part of the population are sampled more intensively than others, then the resulting sample can look very different from the population from which it is drawn and about which we want to make inferences. The data in Table 1, from a survey in which cases of leprosy are heavily oversampled, gives an illustration of this phenomenon, although it is difficult to see the impact with a binary response variable. A more graphic illustration is shown in Figure 1 of Scott and Wild (1986) which plots blood alcohol readings against readings from a blood test for a sample of respondents. The sample was a stratified one, with strata defined by values of the response variable and with very different sampling fractions between strata. When the data were weighted to allow for these varying selection probabilities, the fitted straight line gave a perfectly adequate fit.

In this paper, we attempt to show how standard survey methods for estimating totals can be adapted indirectly to overcome both these problems and produce valid methods for fitting models to survey data and testing hypotheses about model parameters.

2. THEORY

Suppose that, attached to all units of a finite population of size N , we have measurements (\mathbf{x}_i, y_i) made on a vector of explanatory variables, \mathbf{x} , and a response variable, Y . We assume that for a given value of \mathbf{x} , Y is generated by some random process with mean

$$E(Y_i) = \mu_i = \mu(\mathbf{x}_i; \boldsymbol{\beta}) \quad (1)$$

and we suppose that we have in mind some working model for the variance, say

$$\text{var}(Y_i) = V_{0i} = V_0(\mu_i) \quad (2)$$

for $i = 1, \dots, N$. The model for the mean is assumed to be valid, but the working variance may only be a rough approximation at best.

We do not observe values for all the population units but only for those in a sample drawn from the finite population according to some well-defined sampling scheme. We are interested in estimating $\boldsymbol{\beta}$ and, more particularly, in testing the hypothesis that

$\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ using the sample data where $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ with $\boldsymbol{\beta}_2$ a $q \times 1$ vector.

Suppose that, if we had values for the whole finite population, we could obtain a consistent estimator of $\boldsymbol{\beta}$ by solving the estimating equation

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_1^N \mathbf{u}_i(\boldsymbol{\beta}) = 0, \quad (3)$$

where $\mathbf{u}_i(\boldsymbol{\beta})$ has k -th component

$$u_{ik} = \frac{\partial \mu_i}{\partial \beta_k} \frac{(y_i - \mu_i)}{V_{0i}}. \quad (4)$$

Thus we are working in the general estimating equation framework considered by Godambe and Thompson (1986) and Godambe (1991), although there is no requirement that the estimating equation be optimal or that the units be sampled independently from the superpopulation. Note that the resulting estimator is the quasi-likelihood estimator if the finite population is regarded as a random sample from the superpopulation, but the estimator is consistent under much more general conditions. Essentially, all we need to assume is that the finite population can be regarded as a self-weighting sample from the superpopulation.

In reality, we do not know the values for the whole finite population but only for those units in a sample drawn from the population. We suppose only that the sample design provides consistent, asymptotically normal estimators of population totals, and associated standard errors. Then, since $\mathbf{S}_N(\boldsymbol{\beta})$ is a vector of population totals for fixed $\boldsymbol{\beta}$, we can produce an estimator of $\mathbf{S}_N(\boldsymbol{\beta})$ say

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_{i \in s} w_{is} \mathbf{u}_i(\boldsymbol{\beta}), \quad (5)$$

where the survey weights, w_{is} , may depend on the sample s (e.g., post-stratified weights). Our sample estimator, $\hat{\boldsymbol{\beta}}$, is obtained by solving $\hat{\mathbf{S}}(\hat{\boldsymbol{\beta}}) = 0$. This approach was suggested by Fuller (1975) for linear regression with two-stage sampling and by Binder (1983), for generalized linear models and any survey design.

Under suitable conditions (see Binder, 1983, for details), $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$, and we can estimate $\text{cov}(\hat{\boldsymbol{\beta}})$ consistently by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = [\mathbf{J}(\hat{\boldsymbol{\beta}})]^{-1} \hat{\mathbf{V}}_s(\hat{\boldsymbol{\beta}}) [\mathbf{J}(\hat{\boldsymbol{\beta}})]^{-1}, \quad (6)$$

where

$$\mathbf{J}(\boldsymbol{\beta}) = -\frac{\partial \hat{\mathbf{S}}}{\partial \boldsymbol{\beta}^T} = \sum_{i \in S} w_{is} \frac{\partial \mathbf{u}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \quad (7)$$

and $\hat{\mathbf{V}}_s(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of $\hat{\mathbf{S}}(\boldsymbol{\beta})$ under the specified survey design evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Note that $\hat{\mathbf{V}}_s(\hat{\boldsymbol{\beta}})$ is obtained from the standard survey variance estimator for a total.

Now consider the problem of testing the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$. One approach is to base the test on the corresponding Wald statistic

$$X_W^2 = (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20})^T \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_2)^{-1} (\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{20}). \quad (8)$$

This has the usual problems associated with the Wald test. For example, it is not invariant under reparameterization, and often has poor small sample behaviour. In addition, with survey data the effective degrees of freedom for estimating $\text{cov}(\hat{\boldsymbol{\beta}})$ are often rather small, resulting in instability of $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_2)^{-1}$ when the dimension of $\boldsymbol{\beta}_2$ is large (see Thomas and Rao, 1987). Ideally, we would prefer to use a likelihood ratio test, which is invariant and usually has better small sample properties, but we have no likelihood from which to construct such a test here. However, the score test shares many of the desirable properties of the likelihood ratio test, and it is relatively straightforward to construct a simple analogue of the score test in our framework. Our development of the test and its properties parallels the development of Boos (1992) for the case of random sampling from an infinite population.

Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_{20}^T)^T$ be the solution of $\hat{\mathbf{S}}_1(\tilde{\boldsymbol{\beta}}) = 0$ where $\hat{\mathbf{S}} = (\hat{\mathbf{S}}_1^T, \hat{\mathbf{S}}_2^T)^T$ is partitioned in the same way as $\boldsymbol{\beta}$. The analogue of the score test, which we shall call the quasi-score test, is based on the statistic

$$X_S^2 = \tilde{\mathbf{S}}_2^T \tilde{\mathbf{V}}_{2S}^{-1} \tilde{\mathbf{S}}_2, \quad (9)$$

where $\tilde{\mathbf{S}}_2 = \hat{\mathbf{S}}_2(\tilde{\boldsymbol{\beta}})$ and $\tilde{\mathbf{V}}_{2S}$ is a consistent estimator of $\text{Cov}(\tilde{\mathbf{S}}_2)$. The asymptotic distribution of $\tilde{\mathbf{S}}_2$ under H_0 can be obtained as in Boos (1992), who treated the case of random sampling from an infinite population, by expanding $\hat{\mathbf{S}}_1(\boldsymbol{\beta})$ and $\hat{\mathbf{S}}_2(\boldsymbol{\beta})$ as a function of $\tilde{\boldsymbol{\beta}}$ about the true value, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \boldsymbol{\beta}_{20}^{*T})^T$. We give a brief sketch of the development here.

Expanding $\hat{\mathbf{S}}_1(\tilde{\boldsymbol{\beta}})$ and $\hat{\mathbf{S}}_2(\tilde{\boldsymbol{\beta}})$ gives

$$\hat{\mathbf{S}}_1(\tilde{\boldsymbol{\beta}}) \approx \hat{\mathbf{S}}_1(\boldsymbol{\beta}^*) - \mathbf{J}_{11}^* (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) = 0 \quad (10)$$

and

$$\hat{\mathbf{S}}_2(\tilde{\boldsymbol{\beta}}) \approx \hat{\mathbf{S}}_2(\boldsymbol{\beta}^*) - \mathbf{J}_{21}^* (\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*), \quad (11)$$

where

$$\mathbf{J}^* = \mathbf{J}(\boldsymbol{\beta}^*) = \begin{bmatrix} \mathbf{J}_{11}^* & \mathbf{J}_{12}^* \\ \mathbf{J}_{21}^* & \mathbf{J}_{22}^* \end{bmatrix}.$$

Then, replacing \mathbf{J}^* by its expected value, \mathbf{I}^* say, and substituting for $(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ from (10) into (11) yields

$$\tilde{\mathbf{S}}_2 = \hat{\mathbf{S}}_2(\tilde{\boldsymbol{\beta}}) \approx \hat{\mathbf{S}}_2(\boldsymbol{\beta}^*) - \mathbf{I}_{21}^* \mathbf{I}_{11}^{*-1} \hat{\mathbf{S}}_1(\boldsymbol{\beta}^*) = \sum_{i \in S} w_{is} \mathbf{z}_i \quad (12)$$

where

$$\mathbf{z}_i = \mathbf{u}_{2i}(\boldsymbol{\beta}^*) - \mathbf{A} \mathbf{u}_{1i}(\boldsymbol{\beta}^*) \quad (13)$$

with $\mathbf{A} = \mathbf{I}_{21}^* \mathbf{I}_{11}^{*-1}$ and $\mathbf{u}_i = (\mathbf{u}_{1i}^T, \mathbf{u}_{2i}^T)^T$. It then follows from our assumptions about the survey estimator of a total that $\tilde{\mathbf{S}}_2$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\text{cov}(\tilde{\mathbf{S}}_2)$ under H_0 . Thus, X_S^2 is asymptotically a χ_q^2 variable under H_0 .

The quasi-score test based on X_S^2 shares most of the advantages of its infinite population counterpart. With an appropriate choice for the variance estimator $\tilde{\mathbf{V}}_{2S}$ (we discuss this point further in Section 3), the test is invariant under reparameterization. Moreover, we need only ever fit the simple null model, which is a considerable advantage if the full model contains a large number of terms as will be the case, for example, with a factorial structure of explanatory variables containing a large number of interactions. The few studies that have been carried out so far indicate that the small sample behaviour tends to be better than that of the corresponding Wald test but much more work needs to be done here yet.

3. ESTIMATION OF $\text{cov}(\tilde{\mathbf{S}}_2)$

Calculation of X_S^2 requires an estimator of $\text{cov}(\tilde{\mathbf{S}}_2)$. A resampling method, such as the jackknife or balanced repeated replication (BRR), in the case of stratified multi-stage sampling is particularly attractive because post-stratification and unit nonresponse adjustment are automatically taken into account. For example, a jackknife estimator of $\text{cov}(\tilde{\mathbf{S}}_2)$ under stratified multi-stage sampling with n_h sampled clusters from h -th stratum is given by

$$\hat{\mathbf{V}}_J(\tilde{\mathbf{S}}_2) = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\tilde{\mathbf{S}}_{2(hi)} - \tilde{\mathbf{S}}_2) (\tilde{\mathbf{S}}_{2(hi)} - \tilde{\mathbf{S}}_2)^T. \quad (14)$$

Here $\tilde{S}_{2(hi)}$ is obtained in the same manner as \tilde{S}_2 when the data from the (hi) -th sample cluster is deleted, but using jackknife weights and recalculating $\tilde{\beta}$, say $\tilde{\beta}_{(hi)}$. Computation of $\tilde{\beta}_{(hi)} = (\tilde{\beta}_{1(hi)}^T, \tilde{\beta}_{20}^T)^T$ can be simplified by performing only a single Newton-Raphson iteration for the solution of $\hat{S}_{1(hi)}(\beta_1, \beta_{20}) = 0$, using $\tilde{\beta}$ as the starting value, where $\hat{S}_{1(hi)}(\beta_1, \beta_{20})$ uses the jackknife weights instead of the original weights. We refer the reader to Rao (1996) for details on the jackknife method.

The jackknife quasi-score test resulting from $\hat{V}_J(\tilde{S}_2)$ is invariant under reparameterization, unlike the Wald test X_w^2 . Westat Inc. is currently planning to incorporate our quasi-score test into their WESVAR software based on the jackknife and BRR.

Alternatively, we can use a Taylor linearization variance estimator which amounts to applying the survey variance estimator for a total to the representation in (12) replacing z_i by

$$\tilde{z}_i = \mathbf{u}_{2i}(\tilde{\beta}) - \tilde{\mathbf{A}} \mathbf{u}_{1i}(\tilde{\beta}), \quad (15)$$

where $\tilde{\mathbf{A}}$ is an estimator of $\mathbf{A}^* = \mathbf{I}_{21}^* \mathbf{I}_{11}^{*-1}$. Note that the survey variance estimator used should account for post-stratification and unit nonresponse.

There are several possible choices for $\tilde{\mathbf{A}}$. It might seem natural to use $\mathbf{J}(\tilde{\beta})$ in place of \mathbf{I}^* , where $\mathbf{J}(\beta)$ is defined in (7) and, in fact, this form of the score statistic (9) (with $q=1$) is used by Binder and Patak (1994) to construct confidence intervals for β_2 , although their derivation is quite different from that given here. However, this choice does not have the desired invariance property in general. We can get an invariant test by taking the expectation of $\mathbf{J}(\beta)$ under the mean specification defined by (1), giving

$$\mathbf{I}(\tilde{\beta}) = \sum_{i \in S} w_{is} \mathbf{D}_i(\tilde{\beta}) \mathbf{D}_i(\tilde{\beta})^T / V_{0i}(\tilde{\mu}_i), \quad (16)$$

where

$$\mathbf{D}_i(\beta) = \partial \mu_i(\beta) / \partial \beta^T \text{ and } \tilde{\mu}_i = \mu(\mathbf{x}_i, \tilde{\beta}).$$

We suspect that $\mathbf{I}(\tilde{\beta})$ is also more stable than $\mathbf{J}(\tilde{\beta})$ and it is the one we recommend as the choice for $\tilde{\mathbf{A}}$, although again, much more work is needed here. Note that \mathbf{I} and \mathbf{J} are identical for models with canonical link functions (e.g., logistic regression). We could replace $\tilde{\beta}$ by β in either of the above choices, but this would require fitting the full model and thus negate one of the principal attractions of the score test.

4. SPECIAL CASE: SIMPLE LOGISTIC REGRESSION

We illustrate the preceding theory in the simple special case in which the response variable, Y is binary and we are fitting a simple linear logistic regression model. Thus

$$E(Y_i) = \mu_i = \exp(\beta_1 + \beta_2 x_i) / [1 + \exp(\beta_1 + \beta_2 x_i)] \quad (17)$$

here. As working model for the variance, we take the standard binomial form with $V_{0i} = \mu_i(1 - \mu_i)$ so that $\mathbf{u}_i = \mathbf{x}_i(y_i - \mu_i(\beta))$ where $\mathbf{x}_i = (1, x_i)^T$ and $\beta = (\beta_1, \beta_2)^T$. Suppose that we want to test the null hypothesis that $\beta_2 = 0$. Then,

$$\hat{S}_1(\beta) = \sum_{i \in S} w_{is} (y_i - \mu_i(\beta)) \quad (18)$$

and

$$\hat{S}_2(\beta) = \sum_{i \in S} w_{is} x_i (y_i - \mu_i(\beta)). \quad (19)$$

Setting $\hat{S}_1(\tilde{\beta}) = 0$ gives $\tilde{\beta}_1 = \log[\tilde{p}/(1 - \tilde{p})]$ with $\tilde{p} = \sum_s w_{is} y_i / \sum_s w_{is}$. Thus, the score test is particularly simple in this case since we can write down \tilde{S}_2 explicitly, viz., $\tilde{S}_2 = \sum_s w_{is} x_i (y_i - \tilde{p})$. Note that since this is a canonical model, \mathbf{I} and \mathbf{J} are identical:

$$\mathbf{I}(\beta) = \mathbf{J}(\beta) = \sum_{i \in S} w_{is} \mu_i(1 - \mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}. \quad (20)$$

Substituting \tilde{p} for $\tilde{\mu}_i = \mu_i(\tilde{\beta})$ gives

$$\tilde{\mathbf{A}} = \tilde{\mathbf{I}}_{21} \tilde{\mathbf{I}}_{11}^{-1} = \sum_{i \in S} w_{is} x_i / \sum_{i \in S} w_{is} = \hat{X} \text{ say} \quad (21)$$

and

$$\tilde{z}_i = \mathbf{u}_{2i}(\tilde{\beta}) - \tilde{\mathbf{A}} \mathbf{u}_{1i}(\tilde{\beta}) = (y_i - \tilde{p})(x_i - \hat{X}). \quad (22)$$

The linearization estimator of $\text{var}(\tilde{S}_2)$ is then the standard variance estimator for the total of the "synthetic" variable \tilde{z}_i under the specified design, denoted by $\hat{V}(\sum_s w_{is} \tilde{z}_i)$.

The quasi-score test statistic X_S^2 based on the linearization variance estimator reduces to

$$X_S^2 = \left[\sum_{i \in S} w_{is} x_i (y_i - \tilde{p}) \right]^2 / \hat{V}(\sum_{i \in S} w_{is} \tilde{z}_i) \quad (23)$$

which is distributed as a χ_1^2 variable under the null hypothesis. If the jackknife method is used under stratified multi-stage sampling, then we replace $\hat{V}(\sum_s w_{is} \tilde{z}_i)$ by

$$\hat{V}_J(\tilde{S}_2) = \sum_h \frac{n_h - 1}{n_h} \sum_i (\tilde{S}_{2(hi)} - \tilde{S}_2)^2. \quad (24)$$

5. EXAMPLE

Consider the data shown in Table 1. Here the response variable, Y , is binary and takes the value 1 if the person has leprosy (case) and 0 otherwise (control), and the explanatory variables are Age and Scar, where Scar takes value 1 if a person has a BCG vaccination scar and 0 otherwise. The sample design is stratified random sampling with seven strata. All cases are sampled and the sampling fractions for the six control age strata can be found from the table. For reasons outside the scope of this illustration, we consider a logistic regression model with

$$\log[\mu_i / (1 - \mu_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad (25)$$

where $x_1 = (\text{Age} + 7.5)^{-2}$ and $x_2 = \text{Scar}$. We take the Bernoulli variance, $V_{0i} = \mu_i(1 - \mu_i)$, as our working model variance. Interest here centers on whether the BCG vaccination has any impact on the incidence of leprosy, *i.e.*, in testing the hypothesis that $\beta_2 = 0$. Thus we set $\beta_1 = (\beta_0, \beta_1)^T$ and $\beta_2 = \beta_2$ here. We obtain $\tilde{\beta}_1$ by solving

$$\hat{S}_1(\tilde{\beta}) = \sum_{i \in S} w_{is} \begin{bmatrix} y_i - \tilde{\mu}_i \\ x_{1i} (y_i - \tilde{\mu}_i) \end{bmatrix} = 0 \quad (26)$$

where $\log[\tilde{\mu}_i / (1 - \tilde{\mu}_i)] = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i}$. This gives $\tilde{\beta}_0 = -4.6$ and $\tilde{\beta}_1 = -427.0$. Then

$$\tilde{S}_2 = \sum_{i \in S} w_{is} x_{2i} (y_i - \tilde{\mu}_i) = -32.61.$$

and the linearization variance estimate is $\tilde{V}_{2S} = 99.10$, leading to a value for the score statistic of $X_S^2 = 10.73$.

Thus there does seem to be a strong association between BCG vaccination status and the odds of contracting leprosy (P -value = $Pr(\chi_1^2 < 10.73) < .001$), although as with an observational study, great care needs to be taken with any interpretation of this result.

REFERENCES

- Binder, D.A. (1983). "On the variance of asymptotically normal estimators from complex surveys", *International Statistical Review*, 51, 279-292.
- Binder, D.A., and Patak, Z. (1994). "Use of estimating functions for estimation from complex surveys", *Journal of the American Statistical Association*, 89, 1035-1043.
- Boos, D.D. (1992). "On generalized score tests", *The American Statistician*, 46, 327-333.
- Clayton, D., and Hills, M. (1993). *Statistical Methods in Epidemiology*, Oxford: Oxford University Press.
- Fuller, W.A. (1975). "Regression analysis for sample surveys", *Sankyā C*, 37, 117-32.
- Godambe, V.P. (1991). "Orthogonality of estimating functions and nuisance parameters", *Biometrika*, 78, 143-151.
- Godambe, V.P., and Thompson, M.E. (1986). "Parameters of superpopulation and survey population: their relationship and estimation", *International Statistical Review*, 54, 127-138.
- Liang, K.Y., and Zeger, S. (1986). "Longitudinal data analysis using generalized linear models", *Biometrika*, 73, 13-22.
- Rao, J.N.K. (1996). "Developments in sample survey theory: an appraisal", Technical Report No. 291, Laboratory for Research in Statistics and Probability, Carleton University, Canada.
- Rao, J.N.K., and Scott, A.J. (1991). "A simple method for the analysis of clustered binary data", *Biometrics*, 48, 577-585.
- Rotnitzky, A., and Jewell, N.P. (1990). "Hypothesis testing in semiparametric generalized linear models for cluster correlated data", *Biometrika*, 77, 485-497.
- Scott, A.J., and Wild, C.J. (1986). "Fitting logistic models under case-control or choice-based sampling", *Journal of the Royal Statistical Society B*, 48, 170-182.
- Thomas, D.R., and Rao, J.N.K. (1987). "Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling", *Journal of the American Statistical Association*, 82, 630-636.

Table 1. A Stratified Case-Control Sample

Age ²	Scar = 0		Scar = 1		Total		Popn
	Case	Control	Case	Control	Case	Control	Control
7.5	11	10	14	15	25	25	17327
12.5	28	19	22	31	50	50	13172
17.5	16	6	28	38	44	44	10325
22.5	20	13	19	26	39	39	8026
27.5	36	35	11	12	47	47	4981
32.5	47	49	6	4	53	53	6479
Total					258	258	61310

² Age is group midpoint