LOGISTIC REGRESSION WITH SURVEY DATA

Alastair Scott

Department of Mathematics and Statistics, University of Auckland
and Department of Statistics, University of Wisconsin

ABSTRACT

Statistical methods for analyzing cross-classified categorical data based on log-linear and logistic models under more complex sampling schemes than the standard multinomial or product-multinomial models have been discussed widely in recent years. In a series of papers Rao and Scott (1979, 1981, 1984, 1986) have discussed approximate adjustments to the output of standard log-linear programs using information likely to be available from well-conducted surveys. Roberts (1985) and Roberts, Rao and Kumar (1986) have looked at similar results for logistic regression models. We generalize their results and look at the impact in a practical example. We also look at the loss of efficiency from using ad hoc adaptations of multinomial-based methods in a situation for which fully efficient maximum likelihood methods have been developed.

1. INTRODUCTION

This paper deals with the analysis of tables of counts or proportions which are derived from a sample survey rather than from a designed experiment. A typical example is shown in Table 1, which is based on interviews with 9918 women in the Canada Health Survey (1981), a complex stratified multistage survey covering about 12,000 Canadian households.

TABLE 1

Proportion of women who have never carried out a breast self-examination

Education	Age		
	15-24	25-44	45+
Secondary or less	.45	.41	.40
Some post-secondary	.28	.23	.22

A natural way of analyzing such a table is by fitting a logistic regression model, i.e. a model of the form

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

for the population proportions corresponding to the cell estimates. Any good quantitative social science journal is full of illustrations. For example, in a recent issue of the American Journal of Sociology, McLanahan (1985) fits models of the form

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 FA + \ldots,$$

where p is the probability of still attending school at age 17 and FA is an indicator of the father's absence, to compare the effect of parental absence on the educational achievements of black and white children. The source of the data used to fit the models is the Panel Study of Income Dynamics, which is a complex survey of about 5000 U.S. families conducted by the Survey Research Centre at the Univerity of Michigan.

Almost all of these studies fit their logistic models using a standard computer package, such as SAS or GLIM, which implement methods based on the assumption that the proportions are estimated from independent random samples in each cell (or some equivalent scheme such as random sampling from the whole population etc). Good accounts of this standard methodology can be found in Cox (1970), Bishop et al (1975) or McCullagh and Nelder (1983). Any large-scale survey, however, has a much more complicated structure with stratification and several stages of sampling within each stratum, and the estimated proportions may well be weighted to reflect selection probabilities and involve post-stratification, ratio estimation and so on. The resulting covariance structure will often be a long way from that obtained under the assumption of independent binomial samples in each of the cells. The modifications necessary to allow for this complicated structure are straightforward in principle, but their implementation in practice may not be straightforward at all.

We give a brief outline of the relevant theory in the next section and apply the results to the data in Table 1. In most cases a full likelihood analysis is too complicated to be practicable but it is possible for some special designs. In these cases we can quantify the loss of efficiency from using an ad hoc adaptation of the standard analysis. We look at one such special case in the final section.

2. BASIC THEORY

Suppose we have a vector, $\hat{\underset{\sim}{p}}$ say, of T estimated cell proportions and there is a Central Limit theorem of some sort available so that we are willing to assume that $\sqrt{n}(\hat{\underset{\sim}{p}} - \underset{\sim}{p})$ converges in distribution to a T-variate normal random variable random variable with mean vector $\underset{\sim}{0}$ and covariance matrix V_p, where $\underset{\sim}{p}$ is the corresponding vector of population proportions and n is the sample size (or at least an increasing function of the sample size). Let $\underset{\sim}{\ell}$ denote the vector with ith component $\ell_i = \log \frac{p_i}{1-p_i}$. We are interested in estimating $\underset{\sim}{\beta}$, and perhaps testing the hypothesis that $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$, in the model

$$\underset{\sim}{\ell} = X\underset{\sim}{\beta}$$
$$= X_0\beta_0 + X_1\beta_1 \qquad (1)$$

where $X = (X_0, X_1)$ is a known $T \times p$ matrix of rank p ($p \leq T$) derived from the factor levels, $\underset{\sim}{\beta}$ is the p-vector of unknown parameters, X_1 is a $T \times p_1$ matrix and $\underset{\sim}{\beta}_1$ is $p_1 \times 1$. For example, in Table 1 we might be interested in fitting a model of the form

$$\ell_{ij} = \log \frac{p_{ij}}{1-p_{ij}} = \beta_0 + \beta_1 E_i + \beta_2 A_j,$$

for the proportion p_{ij} where $E_i = 1$ for women with some post-secondary education and zero otherwise and A_j is the median age for the women in the jth column.

If we have an estimate, $\hat{V}_p$ say, of the covariance matrix V_p we can obtain a generalized least squares estimate of $\underset{\sim}{\beta}$ based on the empirical logits $\hat{\ell}_i = \log \frac{\hat{p}_i}{1-\hat{p}_i}$. It follows from standard asymptotic theory that $\sqrt{n}(\hat{\underset{\sim}{\ell}} - \underset{\sim}{\ell})$ converges in distribution to a T-variate normal with mean vector $\underset{\sim}{0}$ and covariance matrix $V_\ell = D^{-1}V_pD^{-1}$ with $D = \mathrm{diag}(p_i(1-p_i))$. The generalized least squares estimator is

$$\hat{\underset{\sim}{\beta}}_G = (X^T\hat{V}_\ell^{-1}X)^{-1}X^T\hat{V}_\ell^{-1}\hat{\underset{\sim}{\ell}}$$

with estimated covariance matrix $(X^T\hat{V}_\ell^{-1}X)^{-1}$. Asymptotic tests for linear hypotheses about $\underset{\sim}{\beta}$ can be produced immediately from this. Good illustrations of this approach are given in Koch et al (1975).

All this requires a good estimate of the covariance matrix V_p and, unfortunately, such estimates are still rarely available. Even when an estimate is available, it will usually be obtained using a random group method (see Wolter (1985)) or a design with a small number of p.s.u.'s per stratum. In either case the degrees of freedom of the estimate will be relatively low and, for a table of any complexity, V_ℓ^{-1} will either not exist or at best be rather unstable. For these reasons researchers often simply run their data through the logistic regression program in a standard computer package. Typically these packages produce the maximum likelihood estimate of $\underset{\sim}{\beta}$ along with its estimated covariance matrix and the likelihood ratio test statistic for the hypothesis that $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$ in the model specified by (1) under the assumption of an independent binomial sample of n_i observations in the ith cell ($i=1,\ldots,T$). Let $\tilde{\underset{\sim}{\beta}}$ be the pseudo maximum-likelihood estimate of $\underset{\sim}{\beta}$ obtained by running the observed vector of proportions, together with a vector of pseudo sample sizes $\tilde{\underset{\sim}{n}} = (\tilde{n}_1,\ldots,\tilde{n}_T)^T$ through a standard package. Asymptotic properties of $\hat{\underset{\sim}{\beta}}$ when $\tilde{n}_i = n_i\hat{\Pi}_i$, where $\hat{\Pi}_i$ is the estimated proportion of the whole population falling in the ith cell, have been developed by Roberts (1985) and Roberts, Rao and Kumar (1986) using the methods developed in Rao and Scott (1984) for general log-linear models. Exactly the same methods carry through for more general choices of $\tilde{n}_i$ provided $\tilde{n}_i/n \to \rho_i$ with $0 < \rho_i < 1$ as $n \to \infty$. The resulting pseudo m - ℓ estimator $\hat{\underset{\sim}{\beta}}$ is a consistent estimator of $\underset{\sim}{\beta}$ with asymptotic covariance matrix

$$D(\hat{\underset{\sim}{\beta}}) = (X^TRV_0RX)^{-1}(X^TRV_pRX)(X^TRV_0RX)^{-1}/n. \qquad (2)$$

where $R = \mathrm{diag}(\rho_i)$ and $V_0 = \mathrm{diag}(p_i(1-p_i)/\rho_i)$ (i.e. V_0/n is the covariance matrix of $\hat{\underset{\sim}{p}}$ that would be appropriate with independent binomial samples from the cells of the table). The final factor in (2) is the asymptotic covariance matrix of $\hat{\underset{\sim}{\beta}}$ under the standard assumptions so that the product of the first two factors is the adjustment that needs to be applied to the output from a standard package to allow for the complexity of the design.

The choice of $\tilde{n}_i$ can have considerable impact on the properties of the resulting estimator. Common choices for $\tilde{n}_i$ are the actual sample size in the ith cell or, if this is not known, $\tilde{n}_i = n\hat{\Pi}_i$ as above. It is possible to do better than this if more is known about the covariance structure of $\hat{\underset{\sim}{p}}$. For example, if we have estimates of the cell variances, say $\hat{V}_i$, then we could take $\tilde{n}_i = \hat{p}_i(1-\hat{p}_i)/\hat{V}_i$ to make diagonal elements of $\hat{R}\hat{V}_p\hat{R}$ identical to those under the

assumed binomial model. In most of the examples we have tried, this choice of $\tilde{n}_i$ has worked so well that the standard output needs very little modification.

Now turn to the problem of testing the hypothesis H_0: $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$ in model (1). The standard likelihood ratio test is based on the statistic

$$G^2(2|1) = 2 \sum_1^T \tilde{n}_i \left[p_i(\hat{\underset{\sim}{\beta}}) \log \frac{p_i(\hat{\underset{\sim}{\beta}})}{p_i(\hat{\underset{\sim}{\beta}}_0)} + (1-p_i(\hat{\underset{\sim}{\beta}})) \log \frac{(1-p_i(\hat{\underset{\sim}{\beta}}))}{(1-p_i(\hat{\underset{\sim}{\beta}}_0))} \right] \qquad (3)$$

where $p_i(\hat{\underset{\sim}{\beta}})$ is the solution of (1) corresponding to the pseudo maximum likelihood estimate $\hat{\underset{\sim}{\beta}}$ and $p_i(\hat{\underset{\sim}{\beta}}_0)$ is the corresponding value under the restriction that $\underset{\sim}{\beta}_1 = \underset{\sim}{0}$. Under stratified binomial sampling, $G^2(2|1)$ has an asymptotic chi-squared distribution with p_1 degrees of freedom under H_0 but this does not remain valid when we have a more complex design. If $n_i \to \rho_i > 0$, then, again following the argument in Roberts (1985) and Roberts, Rao and Kumar (1986) exactly, it can be shown that the asymptotic null distribution of $G^2(2|1)$ is a weighted sum,

$$G^2(2|1) \sim \sum_1^{p_1} \delta_i W_i, \qquad (4)$$

where $W_1,\ldots,W_p$ are independent χ^2_1 random variables and $\delta_1,\ldots,\delta_{p_i}$ are the eigenvalues of $(\tilde{X}_1^T R V_0 R \tilde{X}_1)^{-1}(\tilde{X}_1^T R V_p R \tilde{X}_1)$ with

$$\tilde{X}_1 = X_1 - X_0(X_0^T R V_0 R X_0)^{-1} X_0^T R V_0 R X_1. \qquad (5)$$

The choice of $\tilde{n}_i$ can again have a considerable impact on the quality of the output. If we choose $\tilde{n}_i$ to be $\hat{p}_i(1-\hat{p}_i)/\hat{V}_i$ then $G^2(2|1)$ needs no modification at all if V_p is a diagonal matrix, since $G^2(2|1) \sim \chi^2_{p_1}$ under H_0, and needs little modification if the off-diagonal elements are small.

In principle, we could use the results in (2) and (4) to correct the output of a standard package. This is a useful approach when $\hat{V}_p$ has low degrees of freedom since the corrections do not involve V_p^{-1}. In many cases, however, an estimate of the full covariance matrix is not available and we have to make do with partial information such as estimates of the cell variances. The GLIM package has a procedure for adjusting the estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$ and the likelihood ratio statistic that needs no external information about V_p at all. Let $G^2(1)$ be the standard goodness of fit statistic for the full model (1) (i.e. the likelihood ratio statistic for testing model (1) against a completely saturated model), and let $\tilde{\sigma}^2 = G^2(1)/(T-p)$. The adjustments are simple; the estimated covariance matrix is multiplied by $\tilde{\sigma}^2$ and $G^2(2|1)$ is replaced by $F = G^2(2\ 1)/\tilde{\sigma}^2$ (see McCullagh and Nelder (1983) for details). If $V_p = \delta V_0$ for some constant δ, then F has an asymptotic F distribution with p_1 and T-p degrees of freedom. Such a structure is rather special but arises, for example, with the Dirichlet-multinomial model for cluster sampling developed by Brier (1979). One implication of this structure is that all the estimated proportions have a common design effect (i.e. ratio of the actual variance to the variance for a simple random sample of the same size). If the estimated cell design effects differ widely then it is likely that the GLIM correction will not be completely effective.

There has been a great deal of recent work on producing approximations for the likelihood-ratio test for log-linear models based on partial information about V_p. Details can be found in Bedrick (1983) Rao and Scott (1984), Kumar and Rao (1984), Nathan (1984), Gross (1985), and Scott and Styan (1985). Although the logistic model is formally a special case of the log-linear model the approximations do not work well for the logistic in general. Rao and Scott (1987) consider approximations based on the eigenvalues, $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_T$ say, of $nV_0^{-1}V_p$. Using standard results for eigenvalues, it follows that λ_1 gives an upper bound for the design effect of $\hat{\beta}_i$ and that $\lambda_i \geq \delta_i \geq \lambda_{T-p+i}$ if the δ_i's are arranged in increasing order. It is often adequate to approximate the null distribu-

tion of $G^2(2(1)$ by $\overline{\delta}\chi^2_p$, where $\overline{\delta} = \Sigma\delta_i/p_1$. It follows that $\overline{\delta} \leq T\overline{\lambda}/p_1$ where $\overline{\lambda}$ only requires the diagonal terms of V_p. This gives a good bound if p_1 is large compared to T as when checking goodness-of-fit against a saturated model. If $\tilde{n}_i$ is taken to be $p_i(1-p_i)/V_i$, then $\overline{\lambda} = 1$.

3. EXAMPLE

A good estimate of the full covariance matrix is available for the Canada Health Survey Data in Table 1 (see Hidiroglou and Rao (1983)) so that it is possible to make reasonably precise comparisons. Suppose we fit a model of the form

$$\log \frac{p_{ij}}{1-p_{ij}} = \beta_0 + \beta_1 E_i + \beta_2 A_j \qquad (6)$$

where p_{ij} is the population proportion for the (i,j)th cell and E_i and A_j are as in the previous section. Tables 2 and 3 give values of $\hat{\beta}_i$, the pseudo maximum likelihood estimate of β_i, along with its estimated true standard error and the nominal standard error under the assumption of independent binomial sampling, for i = 0,1,2. In Table 2 $\tilde{n}_{ij}$ was taken to be $n\hat{\Pi}_{ij}$, where $\hat{\Pi}_{ij}$ is the estimated proportion of the target population falling in the (i,j)th cell, and in Table 3 $\tilde{n}_{ij}$ was taken to be $\hat{p}_{ij}(1-\hat{p}_{ij})/\hat{V}_{ij}$, where $\hat{V}_{ij}$ is the estimated variance of $\hat{p}_{ij}$.

TABLE 2

Estimates for Canada Health Survey data with $\tilde{n}_{ij} = n\pi_{ij}$.

i	$\hat{\beta}_i$	Estimated Standard Errors Nominal	True
0	-0.090	.052	.057
1	-0.803	.052	.100
2	-0.115	.024	.031

Clearly the naive estimates of the standard errors are all far too small in Table 2 and need to be inflated by a substantial amount. In this case $G^2(1) = 4.20$ giving a value of $\tilde{\sigma}^2 = 1.40$ so the GLIM correction (based on only 3 degrees of freedom) works well for β_0, reasonably well for β_2, and is inadequate for β_1. Of course, since the inflation factor needed varies from 1.2 for $\hat{\beta}_0$ to 3.7 for $\hat{\beta}_1$ no single correcting factor could possibly be satisfactory here. The fact that the design effects of the cell proportions vary from 1.27 to 3.23 gives us prior warning that this is likely to happen.

TABLE 3

Estimates for Canada Health Survey data with $\tilde{n}_{ij} = p_{ij}(1-p_{ij})/V_{ij}$.

i	$\hat{\beta}_i$	Estimated Standard Errors Nominal	True
0	-0.096	.062	.061
1	-0.817	.079	.099
2	-0.111	.031	.033

The true standard errors are roughly the same in both tables but the nominal values are very much more realistic in Table 3, although the value for $\hat{\beta}_1$ still needs adjusting. The value of $G^2(1)$ in this case is 1.522.

In most examples we have looked at, both the GLIM correction and the device of using $\tilde{n}_i = p_i(1-p_i)/\hat{V}_i$ work somewhat better than the example above. In Scott (1986), for example, we look at some data on unemployment from the Canadian Labour Force Survey quoted in Kumar and Rao (1984) and find the both corrections work extremely well.

4. SPECIAL CASE

There is one very important special case where results can be obtained explicitly. In a case-control study, independent random samples are drawn from the cases (e.g. women who have never carried out a breast self-examination in the example of the previous section) and the controls. There has been a great deal of work on fitting logistic models to such data in the medical literature where cases correspond to people with disease. A good survey of this work can be found in Breslow and Day (1980). There has been a parallel development in the econometric literature under the heading of "choice-based sampling". A survey of this literature can be found in Manski and McFadden (1981). Both approaches are put in a sampling context in Scott and Wild (1986).

Let N_{1t} and N_{0t} denote the number of cases and controls respectively in the t-th cell for the whole population and n_{1t} and n_{0t} the corresponding sample numbers. Assuming the populations are large enough to ignore the finite population correction, $\underset{\sim}{n}_\ell = (n_{\ell},\ldots,n_{\ell T})$ has a multinomial distribution with parameters $n_\ell = \Sigma_1^T n_{\ell t}$ and $\Pi_\ell = (\Pi_{\ell 1},\ldots,\Pi_{\ell t})$ where $\Pi_{\ell t} = N_{\ell t}/N_\ell$ with

$N_\ell = \Sigma_1^T N_{\ell t}$ for $\ell = 0,1$. The usual survey sampler's estimator of $P_t = N_{1t}/(N_{0t} + N_{1t})$ would be $\hat{p}_t = \hat{N}_{1t}/(\hat{N}_{0t}+\hat{N}_{1t})$ where $\hat{N}_{\ell t}$ is the Horwitz Thompson estimator $\hat{N}_{\ell t} = n_{\ell t}/w_\ell$ with $w_\ell = n_\ell/N_\ell$ being the corresponding selection probability (ℓ=0,1). The asymptotic covariance of $\hat{\underset{\sim}{p}}$ can then be derived from standard results for ratio estimators. The resulting matrix has (i,j)th element V_{ij} given by the expression

$$p_i p_j(1-p_i)(1-p_j)\left[\delta_{ij}\lambda_i - (\frac{1}{n_0} + \frac{1}{n_1})\right] \qquad (7)$$

where $\lambda_i = \frac{1}{n_0\pi_{0i}} + \frac{1}{n_1\pi_{1i}}$ and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. This enables us to evaluate $D(\hat{\beta})$ in expression (2) explicitly for any matrix X.

For simplicity, consider the case of a simple linear regression model for the logits, i.e. $\ell_i = \beta_0 + \beta_1 x_i$ $(i=1,\ldots,T)$. Combining (7) and (2) we find, after some rather tedious algebra, that the asymptotic variance of $\hat{\beta}_1$ is

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_1^T n_t p_t(1-p_t)^2\lambda_t(x_t-\tilde{x})^2}{(\sum_1^T n_t p_t(1-p_t)(x_t-\tilde{x})^2)^2} \qquad (8)$$

where $\tilde{x} = \Sigma_1^T \tilde{n}_t p_t(1-p_t)x_t / \Sigma_1^T \tilde{n}_t p_t(1-p_t)$. We can use this expression to compare the efficiencies for different choices of $\tilde{n}_t$. In particular, $\mathrm{Var}(\hat{\beta}_1)$ is minimized by choosing $\tilde{n}_t$ equal to

$$\tilde{n}_t^{(opt)} = \frac{1}{\lambda_t p_t(1-p_t)}, \qquad (9)$$

which results in a value

$$V_{opt} = \left[\sum_1^T \frac{(x_t-\tilde{x})^2}{\lambda_t}\right]^{-1} \qquad (10)$$

for $\mathrm{Var}(\hat{\beta}_1)$. This optimal choice has another very surprising property; the estimate of $\mathrm{Var}(\hat{\beta}_1)$ obtained from a standard program (given by the final term in expression (2)) is exactly right. Similarly the null distribution of the likelihood statistic for testing H_0: $\beta_1 = 0$ is exactly chi-squared with one degree of freedom. Thus choosing $\tilde{n}_t = \tilde{n}_t^{(opt)}$ is not only efficient but enables us to use the output from a standard program directly (apart from the variance of the constant term). Of course we have to estimate p_t by $\hat{p}_t$ and λ_t by $\frac{1}{n_{0t}} + \frac{1}{n_{1t}}$ when we use (9) in practice but the asymptotic properties are unaffected. Note that choosing $\tilde{n}_t = \tilde{n}_t^{(opt)}$ is not the same as taking $\tilde{n}_t = p_t(1-p_t)/V_t$ which leads to

$$\tilde{n}_t = \frac{1}{p_t(1-p_t)(\lambda_t - \frac{1}{n_0} - \frac{1}{n_1})}$$

In general, $\tilde{n}_t > \hat{n}_t^{(opt)}$ so the nominal standard errors produced will be slightly smaller than the true values on average. If T is large, however, λ_t will be much larger than $\frac{1}{n_0} + \frac{1}{n_1}$ and $\tilde{n}_t$ will be close to $\tilde{n}_t^{(opt)}$. Thus the approximation should work very well for large tables.

We can also look at the properties of the generalized least-squares estimator explicitly in this important special case. It follows from (7) that

$$V_\ell = \Lambda - (\frac{1}{n_0} + \frac{1}{n_1})\underset{\sim}{1}\underset{\sim}{1}^t, \qquad (11)$$

where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_T)$ and $\underset{\sim}{1}^T = (1,\ldots,1)$, and hence that the generalised least squares estimator of β_1 has asymptotic variance equal to V_{opt}.

It is straightforward to extend this to more general linear models provided the model includes a constant term. Suppose $X = (\underset{\sim}{1}, X_1)$ and we center the columns of X_1 so that $X_1^T\underset{\sim}{\lambda} = 0$. Then the asymptotic covariance matrix of $\hat{\underset{\sim}{\beta}}_G$, the generalized least squares estimator, is

$$D(\hat{\underset{\sim}{\beta}}_G) = (X^T V_\ell^{-1} X)^{-1} = \begin{pmatrix} \mu - \frac{1}{n_0} - \frac{1}{n_1} & \underset{\sim}{0}^T \\ \underset{\sim}{0} & (X_1^T\Lambda^{-1}X_1)^{-1} \end{pmatrix} \qquad (12)$$

with $\mu = (\Sigma_1^T \lambda_t^{-1})^{-1}$.

If we take $\tilde{n}_t = \tilde{n}_t^{(opt)}$ (as in (9)) and substitute these values in (2) we find that the

asymptotic covariance matrix of the resulting pseudo maximum likelihood estimator is identical to $D(\hat{\beta}_G)$. Thus if we take $\tilde{n}_t^{(opt)}$ as our pseudo sample size, any standard logistic regression program produces estimates that are asymptotically equivalent to the generalized least squares estimates. Moreover, the standard errors produced by the program are consistent estimates of the correct values with the exception of the constant term where the program produces a value of μ instead of $\mu - \frac{1}{n_0} - \frac{1}{n_1}$.

The standard way of analyzing data from simple case control studies is to ignore the stratification into cases and controls completely and feed the raw counts, n_{0t} and n_{1t}, into a standard program. Provided the model contains a constant term, the resulting values are the maximum likelihood estimates of all coefficients except the constant term and, although the standard assumptions are not met, they are asymptotically efficient. (See Prentice and Pyke (1979) and Cosslett (1981) for details.) A valid estimate of the constant term can be obtained by adding $\log(W_1/W_0)$ to the estimate given by the program. The asymptotic covariance matrix of this maximum likelihood estimator is given in Scott and Wild (1986) and turns out to be identical to $D(\hat{\beta}_G)$ with categorical explanatory variables as here.

Thus the technique of constructing consistent estimates of $\underset{\sim}{p}_t$ and feeding the resulting values into a standard package can be made fully efficient in this special case by careful choice of the pseudo sample sizes. The common choice of $n_t = p_t(1-p_t)/V_t$ does not give full efficiency here but the differences are negligible if there is a large number of cells in the table. The resulting estimates of the standard errors will be slight underestimates of the true values for all coefficients except the constant. The standard error of the constant will be an overestimate.

Acknowledgement. This paper has arisen out of joint work over a number of years with J. N. K. Rao of Carleton University, T. M. F. Smith and D. Holt of the University of Southampton and C. J. Wild of Auckland University. I would like to express my thanks to all of them.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika, 59, 19-35.

Bedrick, E. J. (1983). Adjusted chi-squared tests for cross-classified tables of survey data. Biometrika, 70, 591-596.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, Mass.

Breslow, N. E. and Day, N. E. (1980). The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon.

Brier, S. E., (1980). Analysis of contingency tables under cluster sampling. Biometrika, 67, 591-596.

Cosslett, S. (1981). Maximum likelihood estimators for choice-based samples. Econometrica, 49, 1289-1316.

Cox, D. R. (1970). The Analysis of Binary Data. Methuen, London.

Gross, W. F. (1984). A note on chi-squared tests with survey data. Journal of the Royal Statistical Society B, 46, 270-272.

Hidiroglou, M. A. and Rao, J. N. K. (983). Chi-squared tests for the analysis of three-way contingency tables from the Canada Health Survey. Technical Report, Statistics Canada.

Koch, G. G., Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 43, 59-78.

Kumar, S. and Rao, J. N. K. (1984). Logistic regression analysis of labour force survey data. Survey Methodology, 10, 62-81.

Manski, C. F. and McFadden, D. (eds) (1981). Structural Analysis of Discrete Data with Applications. MIT Press, Cambridge, Mass.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.

Nathan, G. (1984). The effect of complex sample design on log-linear model analysis. Unpublished report.

Prentice, R. L. and Pyke, R. (1976). Logistic disease incidence models and case-control studies. Biometrika, 66, 403-411.

Rao, J. N. K. and Scott, A. J. (1979). Chi-squared tests for categorical data from complex surveys. Proceedings of the ASA Section on Survey Research Methods, 58-66.

Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex surveys. J. Amer. Statist. Ass., 76, 221-230.

Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. Annals of Statistics, 12, 46-60.

Rao, J. N. K. and Scott, A. J. (1987). On simple adjustments to chi-squared tests with survey data. Annals of Statistics, 15 (to appear).

Roberts, G. (1985). Contributions to chi-squared tests with survey data. Unpublished thesis, Carleton University, Ottawa, Canada.

Roberts, G. A., Rao, J. N. K. and Kumar, S. (1986). Logistic regression analysis of sample survey data. Biometrika, 73, (to appear).

Scott, A. J. and Styan, G. P. H. (1985). On generalized eigenvalues and a problem in sample survey analysis. Linear Algebra and its Applications, 70, 209-224.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice-based sampling. J. Royal Statist. Soc. B, 48, (to appear).

Wolter, K. M. (1985). Introduction to Variance Estimation. Springer-Verlag.