Logistic Regression Analysis of Sample Survey Data
Author(s): G. Roberts, J. N. K. Rao and S. Kumar
Source: *Biometrika*, Vol. 74, No. 1 (Mar., 1987), pp. 1-12
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2336016

Biometrika (1987), **74**, 1, *pp.* 1-12
Printed in Great Britain

Logistic regression analysis of sample survey data

By G. ROBERTS
Social Survey Methods Division, Statistics Canada, Ottawa K1A 0T6, Canada

J. N. K. RAO
Department of Mathematics and Statistics, Carleton University, Ottawa K1S 5B6, Canada

AND S. KUMAR
Social Survey Methods Division, Statistics Canada, Ottawa K1A 0T6, Canada

Summary

Standard chi-squared, X^2, or likelihood ratio, G^2, test statistics for logistic regression analysis, involving a binary response variable, are adjusted to take account of the survey design. These adjustments are based on certain generalized design effects. Logistic regression diagnostics to detect any outlying cell proportions in the table and influential points in the factor space are also developed, taking account of the survey design. Finally, the results are used to analyse some data from the October 1980 Canadian Labour Force Survey.

Some key words: Binary response data; Chi-squared test statistic; Design effect; Diagnostic; Satterthwaite's approximation.

1. Introduction

Logistic regression models are extensively used for the analysis of variation in the estimated proportions associated with a binary response variable; see, for example, Cox (1970). However, the standard statistical methods for binomial proportions are often inappropriate for analysing sample survey data due to clustering and stratification used in the survey design. For instance, the standard chi-squared, X^2, and likelihood ratio, G^2, test statistics greatly inflate the type I error rate when a strong, positive intra-cluster correlation is present. As a result, some adjustments to the classical methods that take account of the survey design are necessary in order to make valid inferences from survey data. Section 2 provides adjustments, based on certain generalized design effects, to standard statistics for testing goodness of fit of a model and for testing subhypotheses given a model. A valid estimate of the asymptotic covariance matrix of fitted cell proportions is also obtained. Derivations of asymptotic variances and covariances and of adjustments to test statistics are sketched in the Appendix; details are given in G. Roberts's 1985 Ph.D. thesis at Carleton University.

In addition to formal statistical tests, it is essential to develop diagnostic procedures to detect any outlying cell proportions and influential points in the factor space. Pregibon (1981) developed diagnostic methods for logistic regression with binomial proportions. In § 3, some of these methods have been modified, by making necessary adjustments to account for the survey design. Finally, the results are used in § 4 to analyse some data from the October 1980 Canadian Labour Force Survey.

The methods developed in this paper require the estimated covariance matrix of cell response proportions.

2. TEST STATISTICS

2·1. *Pseudo maximum likelihood estimates*

Suppose that the population of interest is partitioned into I cells or domains according to the levels of one or more factors. Let $\hat{N}_i$ denote the survey estimate of the ith domain size N_i $(i=1, \ldots, I; \sum N_i = N)$. The corresponding estimate of the ith domain total, N_{i1}, of a binary (0, 1) response variable is denoted by $\hat{N}_{i1}$. The ratio estimate $p_i = \hat{N}_{i1}/\hat{N}_i$ is often used to estimate the population proportion $\pi_i = N_{i1}/N_i$. Standard sampling theory provides an estimate of the covariance matrix of the p_i.

A logistic regression model for the proportions π_i is given by $\pi_i = f_i(\beta)$, where

$$\nu_i = \log[f_i(\beta)/\{1-f_i(\beta)\}] = x_i'\beta \quad (i=1, \ldots, I). \tag{2·1}$$

In (2·1), x_i is an s-vector of known constants derived from the factor levels and β is an s-vector of unknown parameters. Under independent binomial sampling in each domain, the maximum likelihood estimates $\hat{\beta}$ and $\hat{f} = f(\hat{\beta}) = (\hat{f}_1, \ldots, \hat{f}_I)'$ are obtained from the following likelihood equations through iterative calculations:

$$X'D(n)\hat{f} = X'D(n)q, \tag{2·2}$$

where $X' = (x_1, \ldots, x_I)$ is an $s \times I$ matrix of rank s, $D(n) = \operatorname{diag}(n_1, \ldots, n_I)$, q is the vector of sample proportions $q_i = n_{i1}/n_i$, n_i is the sample size from the ith domain, $\sum n_i = n$, and n_{i1} is the ith sample domain total. For general sample designs, we do not have maximum likelihood estimates due to difficulties in obtaining appropriate likelihood functions. Hence, it is a common practice to use a 'pseudo' maximum likelihood estimate of β obtained from (2·2) by replacing n_i/n by the estimated domain relative size $w_i = \hat{N}_i/\hat{N}$, and q_i by the ratio estimate p_i:

$$X'D(w)\hat{f} = X'D(w)p, \tag{2·3}$$

where $D(w) = \operatorname{diag}(w_1, \ldots, w_I)$ and $p = (p_1, \ldots, p_I)'$. The resulting estimates, $\hat{\beta}$ and $\hat{f} = f(\hat{\beta})$, are asymptotically consistent.

2·2. *Estimated asymptotic variances and covariances*

Let $n^{-1}\hat{V}$ denote the survey estimate of the covariance matrix of p. Then the estimated asymptotic covariance matrix of $\hat{\beta}$ is given by

$$\hat{V}_\beta = n^{-1}(X'\hat{\Delta}X)^{-1}\{X'D(w)\hat{V}D(w)X\}(X'\hat{\Delta}X)^{-1}, \tag{2·4}$$

where $\hat{\Delta} = \operatorname{diag}\{w_1\hat{f}_1(1-\hat{f}_1), \ldots, w_I\hat{f}_I(1-\hat{f}_I)\}$; see Appendix 1. In the binomial case, (2·4) reduces to the standard formula $(X'\hat{\Delta}_b X)^{-1}$, where

$$\hat{\Delta}_b = \operatorname{diag}\{n_1^{-1}\hat{f}_1(1-\hat{f}_1), \ldots, n_I^{-1}\hat{f}_I(1-\hat{f}_I)\}.$$

The estimated asymptotic covariance matrix of the fitted cell proportions $\hat{f}$ is

$$\hat{V}_f = D(w)^{-1}\hat{\Delta}X\hat{V}_\beta X'\hat{\Delta}D(w)^{-1}; \tag{2·5}$$

see Appendix 1. The smoothed estimates $\hat{f}$ can be considerably more efficient than the survey estimates p, especially for cells with a small sample, if the model (2·1) provides an adequate fit to p; see § 3·3. The estimates $\hat{f}_i$ are similar to the so-called synthetic estimates employed in small area estimation.

The estimated asymptotic covariance matrix of the residual vector $r = p - \hat{f}$ is

$$\hat{V}_r = n^{-1} A \hat{V} A', \tag{2·6}$$

where

$$A = \mathscr{I} - D(w)^{-1} \hat{\Delta} X (X' \hat{\Delta} X)^{-1} X' D(w) \tag{2·7}$$

and $\mathscr{I}$ is the $I \times I$ identity matrix; see Appendix 1. The diagonal elements, $\hat{V}_{ii,r}$, of (2·6) are needed to calculate the standardized residuals $r_i / \hat{V}^{\frac{1}{2}}_{ii,r}$ which are useful in detecting outlying cell proportions; see § 2·5.

2·3. *Goodness of fit of the model*

The standard X^2 and G^2 tests of goodness-of-fit of the model (2·1) are given by

$$X^2 = n \sum_{i=1}^{I} (p_i - \hat{f}_i)^2 w_i / \{\hat{f}_i (1 - \hat{f}_i)\} = \sum_{i=1}^{I} X_i^2, \tag{2·8}$$

say, and

$$\begin{aligned} G^2 &= 2n \sum_{i=1}^{I} w_i [p_i \log (p_i / \hat{f}_i) + (1 - p_i) \log \{(1 - p_i)/(1 - \hat{f}_i)\}] \\ &= \sum_{i=1}^{I} G_i^2, \end{aligned} \tag{2·9}$$

say. Note that G_i^2 is defined at $p_i = 0$ and 1, respectively, by the quantities $-2nw_i \log (1 - \hat{f}_i)$ and $-2nw_i \log \hat{f}_i$. Under independent binomial sampling, it is well known that both X^2 and G^2 are asymptotically distributed as a χ^2 variable with $I - s$ degrees of freedom when the model (2·1) holds, but for general sample designs this result is no longer valid. In fact, X^2 or G^2 is asymptotically distributed as a weighted sum $\sum \delta_i Z_i$ of independent χ^2 variables Z_i, each with 1 degree of freedom; see Appendix 2. Here, the weights δ_i $(i = 1, \ldots, I - s)$ are estimated by $\hat{\delta}_i$, the eigenvalues of $\hat{V}_{0\phi}^{-1} \hat{V}_\phi$, where

$$\hat{V}_\phi = n^{-1} H' \hat{\Delta}^{-1} D(w) \hat{V} D(w) \hat{\Delta}^{-1} H, \quad \hat{V}_{0\phi} = n^{-1} H' \hat{\Delta}^{-1} H \tag{2·10}$$

and H is any $I \times (I - s)$ matrix of rank $I - s$ such that $H'X = 0$. The eigenvalues are invariant to the choice of H. The matrix $\hat{V}_{0\phi}^{-1} \hat{V}_\phi$ and $\hat{\delta}_i$ are termed a 'generalized design effect matrix' and a 'generalized design effect' respectively since they reduce to I and 1 respectively under binomial sampling.

An adjustment to X^2 or G^2 is obtained by treating $X_c^2 = X^2 / \hat{\delta}_.$ or $G_c^2 = G^2 / \hat{\delta}_.$ as a χ^2 variable with $I - s$ degrees of freedom, where

$$(I - s) \hat{\delta}_. = \sum \hat{\delta}_i = n \sum_{i=1}^{I} \hat{V}_{ii,r} w_i / \{\hat{f}_i (1 - \hat{f}_i)\}. \tag{2·11}$$

The adjusted statistics X_c^2 and G_c^2 should be satisfactory if the coefficient of variation of the δ_i is small. A better adjustment, based on the well-known Satterthwaite approximation, treats $X_S^2 = X_c^2 / (1 + \hat{a}^2)$ or $G_S^2 = G_c^2 / (1 + \hat{a}^2)$ as a χ^2 variable with $(I - s)/(1 + \hat{a}^2)$ degrees of freedom, where

$$\hat{a}^2 = \sum_{i=1}^{I-s} (\hat{\delta}_i - \hat{\delta}_.)^2 / \{(I - s) \hat{\delta}_.^2\}, \tag{2·12}$$

and $\sum \hat{\delta}_i^2$ is obtained from

$$\sum_{i=1}^{I-s} \hat{\delta}_i^2 = \sum_{i=1}^{I} \sum_{j=1}^{I} \hat{V}_{ij,r}^2 (nw_i)(nw_j)/\{\hat{f}_i \hat{f}_j (1-\hat{f}_i)(1-\hat{f}_j)\}, \tag{2·13}$$

where $\hat{V}_{ij,r}$ is the (i, j)th element of $\hat{V}_r$. The test statistics X_S^2 and G_S^2 take account of the variation in the δ_i unlike X_c^2 and G_c^2.

A Wald statistic, which also takes the survey design into account, is given by

$$X_W^2 = \hat{\nu}' H \hat{V}_\phi^{-1} H' \hat{\nu} = \hat{\phi}' \hat{V}_\phi^{-1} \hat{\phi}, \tag{2·14}$$

where $\hat{\nu}$ is the vector of logits $\hat{\nu}_i = \{p_i/(1-p_i)\}$. It is invariant to the choice of H. The statistic X_W^2 is asymptotically distributed as a χ^2 variable with $I-s$ degrees of freedom when the model (2·1) holds. This result follows from the fact that testing the fit of the model (2·1) is equivalent to testing the hypothesis $H\nu = 0$, where $\nu = (\nu_1, \ldots, \nu_I)'$. The statistic X_W^2, however, is not defined if $p_i = 0$ or 1 for some i, as in the case of Labour Force Survey data analysed in § 4. Moreover, it can become unstable when any p_i is close to 1 as shown in § 4, or when the number of degrees of freedom for $\hat{V}$ is not large relative to $I-s$ (Fay, 1985).

2·4. *Nested hypotheses*

Suppose that the matrix X is partitioned as (X_1, X_2), where X_1 is $I \times r$ and X_2 is $I \times u$ $(r+u=s)$. The logistic regression model (2·1), say M, may then be written as

$$\nu = X\beta = X_1\beta_1 + X_2\beta_2,$$

where β_1 is $r \times 1$ and β_2 is $u \times 1$. We are often interested in testing the null hypothesis $H_{2.1}$: $\beta_2 = 0$, given M. Denote the reduced model under $H_{2.1}$ as M_1. The pseudo maximum likelihood estimate $\tilde{\beta}_1$ of β_1 under M_1 can be obtained from the equations

$$X_1' D(w) \tilde{f} = X_1' D(w) p \tag{2·15}$$

again by iterative calculations, where $\tilde{f} = f(\tilde{\beta}_1)$. The standard X^2 and G^2 tests of $H_{2.1}$ are given by

$$X^2(2|1) = n \sum_{i=1}^{I} (\hat{f}_i - \tilde{f}_i)^2 w_i / \{\tilde{f}_i (1-\tilde{f}_i)\}, \tag{2·16}$$

$$G^2(2|1) = 2n \sum_{i=1}^{I} w_i [\hat{f}_i \log\{\hat{f}_i/\tilde{f}_i\} + (1-\hat{f}_i) \log\{(1-\hat{f}_i)/(1-\tilde{f}_i)\}] \tag{2·17}$$

respectively. Under $H_{2.1}$, $X^2(2|1)$ or $G^2(2|1)$ is asymptotically distributed as a weighted sum, $\sum \delta_i(2|1) Z_i$, of independent χ^2 variables Z_i, each with 1 degree of freedom. Here the weights $\delta_i(2|1)$ $(i = 1, \ldots, u)$ are estimated by $\hat{\delta}_i(2|1)$, the eigenvalues of

$$(\tilde{X}_2' \hat{\Delta} \tilde{X}_2)^{-1} (\tilde{X}_2' D(w) \hat{V} D(w) \tilde{X}_2), \tag{2·18}$$

where $\tilde{X}_2 = \{\mathcal{I} - X_1(X_1' \hat{\Delta} X_1)^{-1} X_1' \hat{\Delta}\} X_2$; see Appendix 2.

An adjustment to $G^2(2|1)$ or $X^2(2|1)$ is obtained by treating $G^2(2|1)/\hat{\delta}_.(2|1)$ or $X^2(2|1)/\hat{\delta}_.(2|1)$ as χ^2 with u degrees of freedom under $H_{2.1}$, where $\hat{\delta}_.(2|1) = u^{-1} \sum \hat{\delta}_i(2|1)$ may be computed from

$$u\hat{\delta}_.(2|1) = n \sum_{i=1}^{I} \tilde{V}_{ii,r} w_i / \{\tilde{f}_i (1-\tilde{f}_i)\} \tag{2·19}$$

and $\tilde{V}_{ii,r}$ is the ith diagonal element of the estimated covariance matrix of residuals, $r_i(2|1)=\hat{f}_i-\tilde{f}_i$, given by

$$\tilde{V}_r=n^{-1}D(w)^{-1}\hat{\Delta}\tilde{X}_2\tilde{A}\tilde{X}_2'\hat{\Delta}D(w)^{-1}, \tag{2·20}$$

$$\tilde{A}=(\tilde{X}_2'\hat{\Delta}\tilde{X}_2)^{-1}\{\tilde{X}_2'D(w)\hat{V}D(w)\tilde{X}_2\}(\tilde{X}_2'\hat{\Delta}\tilde{X}_2)^{-1}; \tag{2·21}$$

see equations (A·10) and (A·12) in Appendix 2. The standardized residuals $r_i(2|1)/\tilde{V}_{ii,r}^{\frac{1}{2}}$ can also be computed. As in the case of goodness of fit, a better adjustment based on the Satterthwaite approximation can be obtained, using the elements of $\tilde{V}_r$.

A Wald statistic for testing $H_{2.1}$ is given by

$$X_W^2(2|1)=\hat{\beta}_2'\hat{V}_{2\beta}^{-1}\hat{\beta}_2, \tag{2·22}$$

where $\hat{V}_{2\beta}$ is the principal submatrix of (2·4) corresponding to β_2. Under $H_{2.1}$, the statistic $X_W^2(2|1)$ is asymptotically distributed as a χ^2 with u degrees of freedom. In particular, if β_2 is a scalar, then we can treat $\hat{\beta}_2/\{\text{var}(\hat{\beta}_2)\}^{\frac{1}{2}}$ as $N(0,1)$ or $\hat{\beta}_2^2/\text{var}(\hat{\beta}_2)$ as χ^2 with 1 degree of freedom, under $H_{2.1}$. Note that $X_W^2(2|1)$ is well defined even if $p_i=0$ or 1 for some i, unlike X_W^2. The Wald statistic (2·22) is computationally simpler than the adjusted $X^2(2|1)$ and $G^2(2|1)$ statistics.

The F statistic used in GLIM accounts for extra binomial variation. To test $H_{2.1}$ given M, the statistic

$$F=\{G^2(2|1)/u\}\{G^2/(I-s)\}^{-1} \tag{2·23}$$

is treated as an F variable with degrees of freedom u and $I-s$ respectively. Rao & Scott (1987) have shown that F works well if $\delta_. \simeq \delta_.(2|1)$ and $I-s$ is large. The GLIM method does not require the knowledge of any design effect, but it cannot provide an overall goodness-of-fit test of the model M.

2·5. *Diagnostics*

It is desirable to make a critical assessment of the logistic regression fit by identifying any outlying cell proportions and influential points in the factor space. For identifying outliers, a natural choice that takes account of the survey design is the vector of standardized residuals $e_i=r_i/\hat{V}_{ii,r}^{\frac{1}{2}}$ $(i=1,\ldots,I)$. Since the e_i are approximately $N(0,1)$ under the model, the expected numbers of $|e_i|$ exceeding 1·96, 2·33 and 2·58 are roughly equal to $0{\cdot}5I$, $0{\cdot}02I$ and $0{\cdot}01I$ respectively. These expected numbers provide a rough guide for identifying any outlying cells. Ignoring the design, and hence using standardized residuals under binomial sampling, could lead to erroneous diagnostics. The standardized residuals e_i, however, become unreliable for those cells with $p_i=1$ or close to 1. To circumvent this difficulty, we suggest the use of components of X_c^2 or G_c^2, $\tilde{X}_i=X_i/\hat{\delta}_.^{\frac{1}{2}}$ or $\tilde{G}_i=G_i/\hat{\delta}_.^{\frac{1}{2}}$ $(i=1,\ldots,I)$, for residual analysis. Pregibon (1981) used X_i or G_i in the binomial context. Large individual components $\tilde{X}_i$ or $\tilde{G}_i$ should roughly indicate cells poorly accounted for by the model. Index plots of $\tilde{X}_i$ versus i and $\tilde{G}_i$ versus i are useful for displaying these components. A normal probability plot of $\tilde{X}_i$ or $\tilde{G}_i$ is also useful for detecting deviations from the model.

Following Pregibon (1981), we suggest the use of diagonal elements, m_{ii}, of the projection matrix

$$M=\mathscr{I}-\hat{\Delta}^{\frac{1}{2}}X(X'\hat{\Delta}X)^{-1}X'\hat{\Delta}^{\frac{1}{2}}=\mathscr{I}-T, \tag{2·24}$$

say, to detect influential points. The matrix M arises naturally in solving the 'pseudo' likelihood equations (2·3) by the method of iteratively reweighted least squares (Pregibon, 1981), and small values of m_{ii} call attention to extreme points in the factor space. The index plot of m_{ii} versus i provides a useful display. It may be noted that the design effect does not come into the picture with m_{ii} since we are using 'pseudo' maximum likelihood estimates.

Another useful plot which effectively summarizes the information in the index plots of $\tilde{X}_i$ versus i and m_{ii} versus i is given by the scatter plot of $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ versus t_{ii}, where t_{ii} is the ith diagonal element of T given by (2·24). Again, the design effect does not come into the picture.

The diagnostic measures e_i, $\tilde{X}_i$ or $\tilde{G}_i$, and m_{ii} are useful for detecting extreme points, but not for assessing their impact on various aspects of the fit, including parameter estimates, $\hat{\beta}$, fitted values, $\hat{f}$, and goodness-of-fit measures $X^2/\hat{\delta}_.$ and $G^2/\hat{\delta}_.$ or others. Following Pregibon (1981), we suggest three measures which quantify the effect of extreme cells on the fit. These measures take account of the design effect.

(i) *Coefficient sensitivity.* Let $\hat{\beta}_j(-l)$ denote the pseudo maximum likelihood estimate of β_j obtained after deleting the lth cell from the data. Then the quantity $\Delta_j(l) = \{\hat{\beta}_j - \hat{\beta}_j(-l)\}/\{\text{est var}\,(\hat{\beta}_j)\}^{\frac{1}{2}}$ provides a measure of the jth coefficient sensitivity to the lth cell. The index plots of $\Delta_j(l)$ versus l for each j provide useful displays, but the task of looking at the index plots could become unmanageable unless the number of coefficients in the model is small.

(ii) *Sensitivity of fitted values.* Significant changes in coefficient estimates when the lth point is deleted from the data set does not necessarily imply that the fitted values $\hat{f}$ also vary significantly from $\hat{f}(-l) = f\{\hat{\beta}(-l)\}$, where $\hat{\beta}(-l)$ is the s-vector of estimates $\hat{\beta}_j(-l)$; that is, $\|\hat{f} - \hat{f}(-l)\|$ could be small. The measure $\{G^2 - \tilde{G}^2(-l)\}/\hat{\delta}_.$ or $\{X^2 - \tilde{X}^2(-l)\}/\hat{\delta}_.$ may be used to assess the impact of the lth point on the fitted values $\hat{f}$, where $\tilde{G}^2(-l)$ and $\tilde{X}^2(-l)$ are given by (2·9) and (2·8) respectively when $\hat{f} = f(\hat{\beta})$ is replaced by $\hat{f}(-l)$.

(iii) *Goodness-of-fit sensitivity.* A measure of goodness-of-fit sensitivity is given by $\{G^2 - G^2(-l)\}/\hat{\delta}_.$ or $\{X^2 - X^2(-l)\}/\hat{\delta}_.$, where

$$X^2(-l) = n \sum_{i \neq l} \{p_i - \hat{f}_i(-l)\}^2 w_i/[\hat{f}_i(-l)\{1 - \hat{f}_i(-l)\}]$$

and $G^2(-l)$ is similarly defined using (2·9). Note that $X^2(-l) \neq \tilde{X}^2(-l)$ and $G^2(-l) \neq \tilde{G}^2(-l)$.

3. Analysis of labour force survey data

3·1. *Description of data*

The methods in § 2 were applied to some data from the October 1980 Canadian Labour Force Survey. The sample consisted of males aged 15–64 who were in the labour force and not full-time students. Two factors, age and education, were chosen to explain the variation in unemployment rates via logistic regression models. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the jth age group being the interval $[10+5j, 14+5j)$, for $j = 1, \ldots, 10$, and then using the midpoint of each interval, $A_j = 12 + 5j$, as the value of age for all persons in that age group. Similarly, the levels of education, E_k, were formed by assigning to each person a value based on the median years of schooling resulting in the following six levels: 7, 10, 12, 13, 14 and 16. The resultant age by education cross-classification provided a two-way table of $I = 60$ cell proportions or employment rates, π_{jk}.

The Labour Force Survey design employed stratified multi-stage cluster sampling with two stages in the self-representing urban areas and three or four stages in the non-self-representing areas in each province. The survey estimates, p_{jk}, of π_{jk} were adjusted for post-stratification using the projected census age–sex distribution at the provincial level. The estimated covariance matrix, $\hat{V}/n$, of the estimates p_{jk} was based on more than 450 first-stage units so that the degrees of freedom for $\hat{V}$ was large compared to $I=60$. A detailed description of the sampling plan and associated estimation procedures for the Labour Force Survey is given in Statistics Canada (1977).

3·2. *Formal tests of hypotheses*

Scatter plots of the logits, $\hat{\nu}_{jk}=\log\{p_{jk}/(1-p_{jk})\}$, against age levels A_j, at each education level E_k, indicate that $\hat{\nu}_{jk}$ increases with age to a maximum and then decreases. Hence, the following model might be suitable to explain the variation in the π_{jk}:

$$\nu_{jk}=\log\{\pi_{jk}/(1-\pi_{jk})\}=\beta_0+\beta_1 A_j+\beta_2 A_j^2+\beta_3 E_k+\beta_4 E_k^2 \quad (j=1,\ldots,10;\ k=1,\ldots,6). \tag{3·1}$$

Some previous work in the sociological literature also supports such a model (Bloch & Smith, 1977). Applying the results of § 2, the following values were obtained for testing the goodness-of-fit of the model (3·1): $X^2=98·9$, $G^2=101·2$; $X^2/\hat{\delta}_.=52·5$, $G^2/\hat{\delta}_.=53·7$, $\hat{\delta}_.=1·88$.

Since the value of X^2 or G^2 is larger than $\chi^2_{0·05}(55)=77·3$, the upper 5% point of χ^2 with $I-s=55$ degrees of freedom, the model (3·1) would be rejected if the sample design is ignored. On the other hand, the values of $X^2/\hat{\delta}_.$ or $G^2/\hat{\delta}_.$ indicate that the model is adequate, the significance level being approximately equal to 0·52. The value of Satterthwaite's statistic X^2_S when adjusted to refer to $\chi^2_{0·05}(55)$ is equal to 47·7 which is also not significant at the 5% level. Moreover, in the present context with $s(=5)$ relatively small compared with $I(=60)$, the simple correction $\hat{d}_.=\sum\sum \hat{d}_{jk}/60$, depending only on the cell design effects $\hat{d}_{jk}$, is very close to $\hat{\delta}_.:\hat{d}_.=1·905$ compared with $\hat{\delta}_.=1·88$, where $\hat{d}_{jk}=\operatorname{var}(p_{jk})/\{(nw_{jk})^{-1}p_{jk}(1-p_{jk})\}$ and w_{jk} is the estimated relative size for the (j,k)th cell. Rao & Scott (1987) have shown that $\hat{\delta}_.\leqslant\{I/(I-s)\}\hat{d}_.$ so that $\hat{\delta}_.\doteq\hat{d}_.$ when $I/(I-s)\doteq 1$. The Wald statistic X^2_W is not defined here since two of the cells have $p_{jk}=1$; that is, all employed. To circumvent this problem, a few minor perturbations were made to the estimated counts to ensure that $p_{jk}<1$ for all cells and then X^2_W was computed. The resulting values of X^2_W were all large compared with $X^2/\hat{\delta}_.$, at least 30 times larger than $X^2/\hat{\delta}_.$ and varied considerably: 1715 to 3061. It thus appears that the Wald statistic is very unstable for testing goodness of fit in the present context. Alternatively, if the two cells having $p_{jk}=1$ are deleted, then $X^2_W=68·4<\chi^2_{0·05}(53)=71·0$, indicating that the model (3·1) is adequate. However, it is not a good practice to delete cells just to accommodate a chosen statistic. The other problem with X^2_W, noted by Fay (1985), does not arise here since the degrees of freedom for $\hat{V}$ is large compared with the number of cells in the table.

The pseudo maximum likelihood estimates of the β_i, their standard errors and the corresponding standard errors under binomial sampling, all obtained under the model (3·1), are given in Table 1. The Wald statistic $X^2_W(2|1)$ and the G^2 statistic $G^2(2|1)/\hat{\delta}(2|1)$ for the hypotheses $H_{2·1}$: $\beta_2=0$ and $H_{2·1}$: $\beta_4=0$ conditional on model (3·1), are also given in Table 1. As expected, the true standard errors are larger than the corresponding binomial standard errors. The hypothesis $\beta_4=0$, that is, no quadratic education effect,

Table 1. *Pseudo maximum likelihood estimates $\hat{\beta}_i$ and corresponding standard errors for the labour force survey data under model* (3·1). *Also,* $X^2_W(2|1)=\hat{\beta}^2_i/\mathrm{var}(\hat{\beta}_i)$ *and* $G^2(2|1)/\hat{\delta}_.(2|1)$ *for the nested hypotheses* $H_{2\cdot 1}$: $\beta_2=0$ *and* $H_{2\cdot 1}$: $\beta_4=0$

i	$\hat{\beta}_i$	True st. err. ($\hat{\beta}_i$)	Binomial st. err. ($\hat{\beta}_i$)	$X^2_W(2\|1)$	$G^2(2\|1)/\hat{\delta}_.(2\|1)$
0	−2·76				
1	0·209	0·013	0·012		
2	−0·00217	0·000173	0·000136	157·3	102·1
3	0·0913	0·089	0·068		
4	0·00276	0·0041	0·0030	0·45	0·46

is not rejected at the 5% level either by the Wald statistic or the G^2 statistic: $\chi^2_{0\cdot 05}(1)=3\cdot 84$. On the other hand, the coefficient β_2 of A_j^2 is highly significant, indicating a quadratic age effect.

Two more nested hypotheses given the model (3·1) are also of interest: $H_{2\cdot 1}$: $\beta_3=\beta_4=0$ or no education effect; $H_{2\cdot 1}$: $\beta_2=\beta_4=0$ or no quadratic effect. Both hypotheses are rejected at the 1% level:

$$G^2(2|1)/\hat{\delta}_.(2|1)=282\cdot 2/1\cdot 64=172\cdot 1,\quad X^2_W(2|1)=165\cdot 6\quad \text{for } H_{2\cdot 1}\text{: } \beta_3=\beta_4=0,$$

$$G^2(2|1)/\hat{\delta}_.(2|1)=242\cdot 2/2\cdot 28=106\cdot 3,\quad X^2_W(2|1)=162\cdot 1\quad \text{for } H_{2\cdot 1}\text{: } \beta_2=\beta_4=0,$$

as compared with $\chi^2_{0\cdot 01}(2)=9\cdot 21$. Note that the Wald statistic for $H_{2\cdot 1}$ leads to values close to the corresponding values of $G^2(2|1)/\hat{\delta}_.(2|1)$.

The GLIM overdispersion correction amounts to dividing $G^2(2|1)/u$ by $G^2/(I-s)=101\cdot 2/55=1\cdot 84$ $(u=1, 2)$ and treating the ratio as an F variable with u and 55 degrees of freedom respectively. The GLIM results are in broad agreement with $G^2(2|1)/\hat{\delta}_.(2|1)$.

The above tests of goodness of fit and nested hypotheses lead to the following simple model involving only four parameters:

$$\log\{f_{jk}/(1-f_{jk})\}=-3\cdot 10+0\cdot 211A_j-0\cdot 00218A_j^2+0\cdot 1509E_k. \tag{3·2}$$

The standard errors of the four estimates are 0·247, 0·013, 0·000172, 0·0115, respectively. The diagnostics in § 3·3 are based on the fitted model (3·2).

3·3. *Diagnostics*

The diagnostics developed in § 2·4 are now applied to the Labour Force Survey data. Due to space limitations, only selected plots are given here.

(i) *Residual analysis.* The 60 cells in the two-way table were numbered lexicographically and the standardized residuals e_i were computed under the model (3·2). The cells numbered 6 and 54 with $p_i=1$ lead to very large values of e_i: 66·2 and 6·2 respectively, which are unreliable as noted earlier. Among the remaining e_i, the residuals numbered 7, 27 and 59 have values 3·84, 2·73 and 2·52 respectively, whereas the expected number of $|e_i|$ exceeding 2·33 is roughly $60\times 0\cdot 02=1\cdot 2$. Hence, there is some indication that cells 7 and 27 might correspond to outlying cell proportions.

The normal probability plots and index plots of $\tilde{G}_i=G_i/\hat{\delta}^{\frac{1}{2}}_.$ and $\tilde{X}_i=X_i/\hat{\delta}^{\frac{1}{2}}_.$ all indicate no evidence of outlying cell proportions.

(ii) *Influential cells.* The index plot of m_{ii} and the plot of $\tilde{X}^2_i/X^2_c=X^2_i/X^2$ versus t_{ii} both suggest that cells 2, 3 and 55 warrant further examination.

(iii) *Coefficient sensitivity.* The index plots for measuring coefficiency sensitivity: $\Delta_j(l)$ versus l are displayed in Fig. 1(a) and (b) for β_1 and β_3 respectively. These plots indicate that cells 2 and 3 might cause instability in $\hat{\beta}_1$, while $\hat{\beta}_3$ may be affected by cell 7. The plots for β_0 and β_2, not given here, show that cells 2 and 3 might also cause instability in $\hat{\beta}_0$ and $\hat{\beta}_2$. However, the values of $\Delta_j(l)$ for $l=2$, 3 and 7 are all small relative to the corresponding values of $\hat{\beta}_j/\{\text{est var}(\hat{\beta}_j)\}^{\frac{1}{2}}$. For example, $\Delta_1(2) \simeq 1\cdot 1$ compared with $\hat{\beta}_1/\{\text{est var}(\hat{\beta}_1)\}^{\frac{1}{2}} = 0\cdot 211/0\cdot 013 = 16\cdot 23$.

(iv) *Sensitivity of fitted values.* The plot of $\{G^2 - \tilde{G}^2(-l)\}/\hat{\delta}_. = c_l$ versus l for assessing the impact of individual cells on fitted values is displayed in Fig. 2(a). Significant peaks in this figure correspond to cells 2 and 3. Following Pregibon (1981), the comparison of c_l to the percentage point of χ^2 with $s=5$ degrees of freedom gives a rough guide as to which contour of the confidence region the pseudo maximum likelihood estimates are displaced due to deletion of the lth cell. The value $c_l = 2\cdot 1$ for cell 2 roughly corresponds to the 78% contour of the confidence region.

(v) *Goodness-of-fit sensitivity.* The plot of $\{G^2 - G^2(-l)\}/\hat{\delta}_.$ versus l is displayed in Fig. 2(b); the plot of $\{X^2 - X^2(-l)\}/\hat{\delta}_.$ is similar but the former plot is preferred (Pregibon, 1981). Significant peaks in this figure correspond to cells 2, 3, 7, 27, 39 and 54 with values $\geqslant 3$, the most significant being cell 7 with the value 5·4. By deleting cell 7 and recomputing the adjusted statistic $G_c^2(-7) = G^2(-7)/\hat{\delta}_.(-7)$, where $\hat{\delta}_.(-7)$ is the corresponding estimate of $\delta_.$, the value of $G_c^2(-7) = 48\cdot 43$ with 55 degrees of freedom is obtained compared with $G^2/\hat{\delta}_. = 55\cdot 3$ with 56 degrees of freedom.

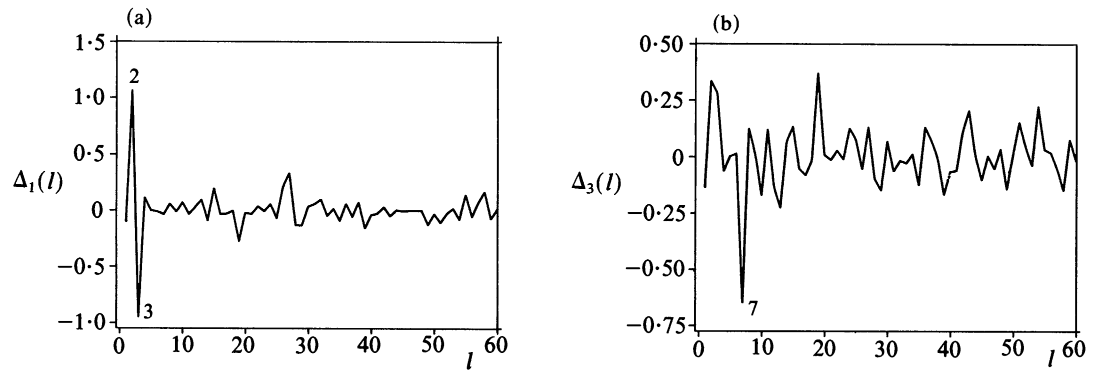

Fig. 1. Index plots for measuring coefficient sensitivity: (a) $\Delta_1(l) = \{\hat{\beta}_1 - \hat{\beta}_1(-l)\}/\{\text{est var}(\hat{\beta}_1)\}^{\frac{1}{2}}$ versus l; (b) $\Delta_3(l) = \{\hat{\beta}_3 - \hat{\beta}_3(-l)\}/\{\text{est var}(\hat{\beta}_3)\}^{\frac{1}{2}}$ versus l.

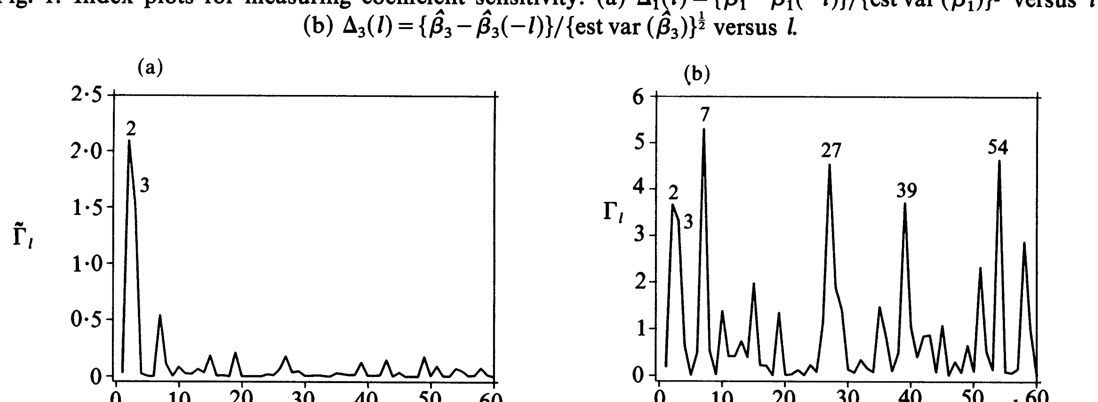

Fig. 2. Index plot for measuring (a) sensitivity of fitted values: $\tilde{\Gamma}_l = \{G^2 - \tilde{G}^2(-l)\}/\hat{\delta}_.^{\frac{1}{2}}$ versus l; (b) goodness-of-fit sensitivity: $\Gamma_l = \{G^2 - G^2(-l)\}/\hat{\delta}_.^{\frac{1}{2}}$ versus l.

The investigation suggests on the whole that the impact of cells indicated by the diagnostics is not significant enough to warrant their deletion.

3·4. *Smoothed estimates*

The coefficient of variation of survey estimates of unemployment rates, $1-p_{jk}$, is quite large for cells with small samples, ranging from 6·8% for cell 3 to 98·5% for cell 59. Because of this, the coefficient of variation of smoothed estimates, $1-\hat{f}_{jk}$, under the model (3·2) were computed using (2·5). The smoothed estimates lead to a dramatic reduction in coefficient of variation: the coefficient of variation of $1-\hat{f}_{jk}$ ranges from 3·3% for cell 8 to 12·4% for cell 60; the coefficient of variation for cell 59 is reduced from 98·5% to 11·0%. The average coefficient of variation of $1-p_{jk}$ over the 58 cells with $1-p_{jk}>0$ is 32·1% compared with 6·2%, the average coefficient of variation of $1-\hat{f}_{jk}$ over all the 60 cells. Moreover, the bias of smoothed estimates should be relatively small since model (3·2) provides an adequate fit to the data.

ACKNOWLEDGEMENTS

J. N. K. Rao's work was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. Thanks are due to A. J. Scott, P. J. Kempthorne and the referees for helpful suggestions, and to M. Gratton for producing the computer-drawn diagrams.

APPENDIX 1

Outline derivation of asymptotic variances and covariances

The pseudo maximum likelihood estimates are obtained from the binomial likelihood, $L(\beta)$ say, by replacing n_i by nw_i and n_{i1} by $(nw_i)p_i$ and then minimizing with respect to β. It is easily seen that, omitting terms not involving β,

$$-2\log L(\beta)=2nG^2\{a^*, b^*(\beta)\},$$

where

$$a^*=\{w_1p_1,\ldots,w_Ip_I,\ w_1(1-p_1),\ldots,w_I(1-p_I)\}',$$
$$b^*=b^*(\beta)=\{w_1f_1,\ldots,w_If_I,\ w_1(1-f_1),\ldots,w_I(1-f_I)\}',$$
$$G^2(a^*,b^*)=\sum a_i^*\log(a_i^*/b_i^*),\quad \sum a_i^*=\sum b_i^*=1.$$

Hence, noting that maximizing $L(\beta)$ is equivalent to minimizing $G^2\{a^*, b^*(\beta)\}$, we can use the results of Birch (1964) to get

$$n^{\frac{1}{2}}(\hat{\beta}-\beta)\sim n^{\frac{1}{2}}\{(B'B)^{-1}B'D(b)^{-\frac{1}{2}}(a-b(\beta))\},\tag{A·1}$$

where $\sim$ denotes asymptotic equivalence. Here a and $b=b(\beta)$ are derived from a^* and b^* respectively by replacing w_i with $W_i=N_{i1}/N_i$, $w_i-W_i=o_p(1)$, $D(b)=\operatorname{diag}(b_1,\ldots,b_I)$ and $B=D(b)^{-\frac{1}{2}}(\partial b/\partial\beta)$. In the case of the logistic regression model (2·1), Birch's (1964) regularity conditions are satisfied and (A·1) reduces to

$$n^{\frac{1}{2}}(\hat{\beta}-\beta)\sim(X'\Delta X)^{-1}X'D(W)\{n^{\frac{1}{2}}(p-f)\},\tag{A·2}$$

where $\Delta=\operatorname{diag}\{W_1f_1(1-f_1),\ldots,W_If_I(1-f_I)\}$ and $D(W)=\operatorname{diag}(W_1,\ldots,W_I)$. Now assuming that $n^{\frac{1}{2}}(p-f)$ converges in distribution to $N_I(0, V)$ and using (A·2), the asymptotic covariance matrix of $\hat{\beta}$ is

$$V_\beta=n^{-1}(X'\Delta X)^{-1}\{X'D(W)VD(W)X\}(X'\Delta X)^{-1}.\tag{A·3}$$

Replacing the parameters in (A·3) by their estimates, (2·4) is obtained. Similarly, noting that

$$n^{\frac{1}{2}}(\hat{f}-f) \sim \left(\frac{\partial f}{\partial \beta}\right)\{n^{\frac{1}{2}}(\hat{\beta}-\beta)\} = D(W)^{-1}\Delta X\{n^{\frac{1}{2}}(\hat{\beta}-\beta)\}, \tag{A·4}$$

$$n^{\frac{1}{2}}(p-\hat{f}) = n^{\frac{1}{2}}r \sim \{I - D(W)^{-1}\Delta X(X'\Delta X)^{-1}X'D(W)\}\{n^{\frac{1}{2}}(p-f)\}, \tag{A·5}$$

leads to (2·5) and (2·6).

Appendix 2

Outline derivation of asymptotic null distribution of $X^2(2|1)$

The statistic $X^2(2|1)$, given by (2·16), for testing the nested hypothesis $H_{2\cdot 1}$: $\beta_2 = 0$ is asymptotically equivalent to

$$n(\hat{f}-\tilde{f})'D(W)\Delta^{-1}D(W)(\hat{f}-\tilde{f}) \tag{A·6}$$

under $H_{2\cdot 1}$. Now, similarly to (A·4), $n^{\frac{1}{2}}(\tilde{f}-f) \sim D(W)^{-1}\Delta X_1\{n^{\frac{1}{2}}(\tilde{\beta}_1-\beta_1)\}$, where

$$n^{\frac{1}{2}}(\tilde{\beta}_1-\beta_1) \sim (X_1'\Delta X_1)^{-1}X_1'D(W)\{n^{\frac{1}{2}}(p-f)\}. \tag{A·7}$$

Hence, from (A·4),

$$n^{\frac{1}{2}}(\hat{f}-\tilde{f}) \sim D(W)^{-1}\Delta\{X_1 n^{\frac{1}{2}}(\hat{\beta}_1-\beta_1) + X_2 n^{\frac{1}{2}}\hat{\beta}_2 - X_1 n^{\frac{1}{2}}(\tilde{\beta}_1-\beta_1)\} \tag{A·8}$$

under $H_{2\cdot 1}$. Following Rao & Scott (1984), $X'\Delta X = (X_1, X_2)'\Delta(X_1, X_2)$ may be expressed as a partitioned matrix, and then, using the standard formula for the inverse of a partitioned matrix, it follows that

$$n^{\frac{1}{2}}(\tilde{\beta}_1-\beta_1) \sim n^{\frac{1}{2}}(\hat{\beta}_1-\beta_1) + (X_1'\Delta X_1)^{-1}(X_1'\Delta X_2)n^{\frac{1}{2}}\hat{\beta}_2. \tag{A·9}$$

Substitution of (A·9) into (A·8) leads to

$$n^{\frac{1}{2}}(\hat{f}-\tilde{f}) \sim D(W)^{-1}\Delta\tilde{X}_2 n^{\frac{1}{2}}\hat{\beta}_2, \tag{A·10}$$

where $\tilde{X}_2 = X_2 - X_1(X_1'\Delta X_1)^{-1}(X_1'\Delta X_2)$. As a result, the following asymptotic representation is obtained from (A·6) and (A·10):

$$X^2(2|1) \sim n\hat{\beta}_2'(\tilde{X}_2'\Delta\tilde{X}_2)\hat{\beta}_2. \tag{A·11}$$

Also it follows from (A·3) and the formula for the inverse of a partitioned matrix that the asymptotic covariance matrix of $\hat{\beta}_2$ may be written as

$$V_{\beta_2} = n^{-1}(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}\{\tilde{X}_2'D(W)VD(W)\tilde{X}_2\}(\tilde{X}_2'\Delta\tilde{X}_2)^{-1} \tag{A·12}$$

so that $\hat{\beta}_2$ is approximately $N_u(0, V_{\beta_2})$ under $H_{2\cdot 1}$. Hence, $X^2(2|1)$ is asymptotically distributed as $\sum \delta_i(2|1)Z_i$, using a standard result on the distribution of a quadratic form in normal variables, where the $\delta_i(2|1)$ are eigenvalues of

$$(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}\{\tilde{X}_2'D(W)VD(W)\tilde{X}_2\}.$$

Replacing Δ, W and V by their estimates $\hat{\Delta}$, w and $\hat{V}$ respectively, (2·18) is obtained. It can be shown that $G^2(2|1)$ is asymptotically equivalent to $X^2(2|1)$ under $H_{2\cdot 1}$ so that the above result also holds in the case of $G^2(2|1)$.

The asymptotic null distribution of X^2 or G^2 can be obtained as a special case of the result for nested hypothesis $H_{2\cdot 1}$, by treating the model M as a saturated model. In the saturated case, $X_1 = X$, X_2 is any $I \times (I-s)$ matrix of rank $I-s$ such that (X, X_2) is $I \times I$ of rank I. Let $H = \Delta\tilde{X}_2$ so that rank H = rank $\tilde{X}_2 = I - s$ and $H'X = \tilde{X}_2'\Delta X = X_2'\{I - \Delta X(X'\Delta X)^{-1}X'\}\Delta X = 0$. Hence

$$(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}\{\tilde{X}_2'D(W)VD(W)\tilde{X}_2\} = (H'\Delta^{-1}H)^{-1}\{H'\Delta^{-1}D(W)VD(W)\Delta^{-1}H\}.$$

Therefore, X^2 or G^2 is asymptotically distributed as the weighted sum $\sum \delta_i Z_i$, where the weights δ_i are the eigenvalues of $(H'\Delta^{-1}H)^{-1}\{H'\Delta^{-1}D(W)VD(W)\Delta^{-1}H\}$. By letting $\tilde{H} = HG$, where G is a nonsingular matrix of order $I-s$, it is easily verified that the δ_i are invariant to the choice of H.

REFERENCES

BIRCH, M. W. (1964). A new proof of the Pearson–Fisher Theorem. *Ann. Math. Statist.* **35**, 818–24.

BLOCH, F. E. & SMITH, S. P. (1977). Human capital and labour market employment. *J. Hum. Resources* **12**, 550–9.

COX, D. R. (1970). *Analysis of Binary Data.* London: Chapman and Hall.

FAY, R. E. (1985). Replication approaches to the log-linear analysis of data from complex samples. In *Recent Developments in the Analysis of Large-scale Data Sets*, pp. 95–118, Luxembourg: Office for Official Publications of the European Communities.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705–24.

RAO, J. N. K. & SCOTT, A. J. (1987). On simple adjustments to chi-squared tests with survey data. *Ann. Statist.* **15**. To appear.

STATISTICS CANADA (1977) *Methodology of the Canadian Labour Force Survey, 1976.* Catalogue 71-526 Occasional. Ottawa: Statistics Canada.

[*Received July* 1985. *Revised May* 1986]