LOGISTIC REGRESSION ANALYSIS OF LABOUR FORCE SURVEY DATA

S. Kumar and J.N.K. Rao[1]

Standard chisquared (X^2) or likelihood ratio (G^2) tests for logistic regression analysis, involving a binary response variable, are adjusted to take account of the survey design. The adjustments are based on certain generalized design effects. The adjusted statistics are utilized to analyse some data from the October 1980 Canadian Labour Force Survey (LFS). The Wald statistic , which also takes the survey design into account, is also examined for goodness-of-fit of the model and for testing hypotheses on the parameters of the assumed model. Logistic regression diagnostics to detect any outlying cell proportions in the table and influential points in the factor space are applied to the LFS data, after making necessary adjustments to account for the survey design.

1. INTRODUCTION

Logistic regression models have been extensively used by researchers in social, behavioural and health sciences to analyse the variation in binomial proportions (see, for example, the books by Cox (1970) and McCullagh and Nelder (1983)). Due to clustering and stratification used in the survey design the statistical methods for binomial proportions, however, are often inappropriate for analysing sample survey data. For instance, the standard chisquared (X^2) or the likelihood ratio (G^2) tests greatly inflate the type I error rate (significance level). Hence, some adjustments to the classical methods that take account of the survey design are necessary in order to make valid inferences from survey data. In this article, we have utilized two simple adjustments to X^2 or G^2, based on certain generalized design effects (deffs) to analyse some data from the October 1980 Canadian Labour Force Survey (LFS) (Section 3). The Wald statistic, which also takes the survey design into account, is also examined.

[1] S. Kumar, Census and Household Survey Methods Division, Statistics Canada, and J.N.K. Rao, Department of Mathematics and Statistics, 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. Regression diagnostics for the standard linear model have been extensively investigated in the literature (see the recent book by Cook and Weisberg (1982)). Pregibon (1981) recently developed similar methods for the logistic regression with binomial proportions. In Section 4 some of these methods have been applied to the October 1980 LFS data, after making necessary adjustments to account for the survey design.

2. THEORETICAL RESULTS

Suppose that the population of interest is partitioned into I cells (domains) according to the levels of one or more factors, and $\hat{N}_i$ denotes the survey estimate of the i-th domain size, N_i ($i = 1, 2, \ldots, I$; $\Sigma N_i = N$). The corresponding estimate of the i-th domain total, N_{i1}, of a binary (0, 1) response variable is denoted by $\hat{N}_{i1}$. The ratio estimate, $\hat{p}_i = \hat{N}_{i1}/\hat{N}_i$, is used to estimate the population proportion $\pi_i = N_{i1}/N_i$.

A logit model on the proportions π_i is given by $\pi_i = f_i(\underset{\sim}{\beta})$, where

$$\ln\{f_i/(1 - f_i)\} = \text{logit } f_i = \underset{\sim}{x}_i'\underset{\sim}{\beta}, \quad i = 1, \ldots, I. \tag{1}$$

In (1), $\underset{\sim}{x}_i$ is an s-vector of known constants derived from the factor levels and $\underset{\sim}{\beta}$ is the s-vector of unknown parameters. Under independent binomial sampling in each domain, the maximum likelihood estimates (m.l.e.) are obtained from the following likelihood equations:

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

where $X' = (\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_I)$, $D(\underset{\sim}{n}/n) = \text{diag}(n_1/n, \ldots, n_I/n)$, $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}}) = (\hat{f}_1, \ldots, \hat{f}_I)'$, and $\hat{\underset{\sim}{q}}$ is the vector of sample proportion $q_i = n_{i1}/n_i$, where n_i is the sample size from i-th domain ($\Sigma n_i = n$). For general sample designs, we do not have m.l.e. due to difficulties in obtaining appropriate likelihood functions. Hence, it is a common practice to use a "pseudo m.l.e." of $\underset{\sim}{\beta}$ or $\underset{\sim}{f}$

obtained from (2) by replacing n_i/n by the estimated domain relative size, $w_i = \hat{N}_i/\hat{N}$, and $\hat{q}_i$ by the survey estimate $\hat{p}_i$:

$$X'D(\underset{\sim}{w})\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{w})\hat{\underset{\sim}{p}}. \tag{3}$$

The resulting estimates, $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{f}} = \underset{\sim}{f}(\hat{\underset{\sim}{\beta}})$, are asymptotically (i.e., in large samples) consistent. The equations (3) may also be written as

$$X'\hat{\underset{\sim}{N}}_1(m) = X'\hat{\underset{\sim}{N}}_1, \tag{4}$$

where $\hat{\underset{\sim}{N}}_1$ is the vector of estimated counts $\hat{N}_{i1}$, and $\hat{\underset{\sim}{N}}_1(m)$ is the vector of pseudo m.l.e., $\hat{N}_{i1}(m) = \hat{N}_i\hat{f}_i$, of the totals N_{i1}. The estimates $\hat{\underset{\sim}{\beta}}$, and hence $\hat{\underset{\sim}{f}}$ and $\hat{\underset{\sim}{N}}_1(m)$, are obtained from (3) or (4) by iterative calculations.

2.1 Estimated Variances and Covariances

Let $\hat{V}$ denote the estimated covariance matrix of $\hat{\underset{\sim}{p}}$, then the estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$ is given by

$$\hat{D}(\hat{\underset{\sim}{\beta}}) = (X'\hat{\Delta}X)^{-1}(X'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})X)(X'\hat{\Delta}X)^{-1} \tag{5}$$

in large samples, where $\hat{\Delta} = diag(w_1\hat{f}_1(1 - \hat{f}_1), \ldots, w_I\hat{f}_I(1 - \hat{f}_I))$. The diagonal elements of (5) provide the estimated variances of the estimates $\hat{\beta}_i$. Similarly, the estimated covariance matrix of the residual vector $\underset{\sim}{r} = \hat{\underset{\sim}{p}} - \hat{\underset{\sim}{f}}$ is given by

$$\hat{D}(\underset{\sim}{r}) = A\hat{V}A', \tag{6}$$

where

$$A = I - D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})X(X'\hat{\Delta}X)^{-1}X'D(\underset{\sim}{w}). \tag{7}$$

The diagonal elements $\hat{V}_{ii}(r)$ of (6) lead to standardized residuals $r_i/s.e.(r_i)$ which are useful in detecting outlying cell proportions.

2.2 Goodness-of-Fit Tests

The standard chi-squared test of goodness-of-fit of the model (1) is given by

$$X^2 = n \sum_{i=1}^{I} \frac{(\hat{p}_i - \hat{f}_i)^2 w_i}{\hat{f}_i(1 - \hat{f}_i)} = \sum_{i=1}^{I} X_i^2. \tag{8}$$

The likelihood ratio test statistic is given by

$$G^2 = 2n \sum_{i=1}^{I} w_i \left\{ \hat{p}_i \ln \frac{\hat{p}_i}{\hat{f}_i} + (1 - \hat{p}_i) \ln \frac{(1 - \hat{p}_i)}{(1 - \hat{f}_i)} \right\} = \sum_{i=1}^{I} G_i^2 \tag{9}$$

Note that G_i^2 is also defined at $\hat{p}_i = 0$ and 1 as given by $-2nw_i \ln(1 - \hat{f}_i)$ and $-2nw_i \ln \hat{f}_i$ respectively. 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, but for general designs this result is no longer valid. In fact, X^2 (or G^2) is asymptotically distributed as a weighted sum $\Sigma \delta_i Z_i$, of independent χ^2 variables, Z_i, each with 1 d.f. where the weights δ_i ($i = 1, \ldots, I - s$) are the eigenvalues of a "generalized design effects" matrix given by $\Sigma_0^{-1} \Sigma_\phi$, where

$$\Sigma_\phi = G'D(\hat{\underset{\sim}{f}})^{-1}D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})^{-1}\hat{V}D(\hat{\underset{\sim}{f}})^{-1}D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})^{-1}G, \tag{10}$$

$$\Sigma_0 = \frac{1}{n} G'\hat{\Delta}^{-1}G \tag{11}$$

and G is any $I \times (I - s)$ matrix of rank $I - s$ such that $G'X = 0$, i.e., G is orthogonal to X. Under binomial sampling, $\Sigma_0^{-1} \Sigma_\phi$ reduces to I, the identity matrix

A simple adjustment to X^2 (or G^2) is obtained (Roberts, 1984) by treating $X_c^2 = X^2/\delta$, or $G_c^2 = G^2/\delta$, as χ^2 with $I - s$ degrees of freedom (d.f.) under the hypothesis that the model is true, where

$$(I - s)\delta_{\cdot} = n \sum_{i=1}^{I} \hat{V}_{ii}(r) w_i / [\hat{f}_i(1 - \hat{f}_i)]. \tag{12}$$

The adjusted statistic X_c^2 (or G_c^2) should be satisfactory excepting in those cases with a large coefficient of variation (C.V.) of the δ_i's. A better adjustment, based on the Satterthwaite approximation, treats $X_S^2 = X_c^2/(1 + a^2)$ or $G_S^2 = G_c^2/(1 + a^2)$ as χ^2 with $(I - s)/(1 + a^2)$ d.f., where

$$a^2 = \Sigma (\delta_i - \delta_{\cdot})^2/[(I - s)\delta_{\cdot}^2] \tag{13}$$

is the $(C.V.)^2$ of the δ_i's and

$$\Sigma \delta_i^2 = \sum_{i=1}^{I} \sum_{j=1}^{I} \hat{V}_{ij}^2(r)(nw_i)(nw_j) \; / \; [\hat{f}_i \hat{f}_j (1 - \hat{f}_i)(1 - \hat{f}_j)] \; , \tag{14}$$

where $\hat{V}_{ij}(r)$ is the (i, j)-th element of $\hat{D}(r)$. The statistics X_S^2 and G_S^2 take account of the variation in δ_i's.

A Wald statistic for goodness-of fit of the model (1) is given by

$$X_W^2 = \hat{\underset{\sim}{\nu}}' G \Sigma_{\hat{\phi}}^{-1} G' \hat{\underset{\sim}{\nu}}, \tag{15}$$

where $\hat{\underset{\sim}{\nu}}$ is the vector of logits $\hat{\nu}_i = \text{logit } \hat{p}_i$. The statistic X_W^2 is distributed as χ^2 with I - s d.f., in large samples. The statistic X_W^2 is not defined if $\hat{p}_i = 0$ or 1 for some i. Moreover, it becomes unstable when any $\hat{p}_i$ is close to 1 (see Section 3), or when the degrees of freedom for $\hat{V}$ is not large compared to I - s (Fay, 1983).

2.3 Nested Hypothesis

Suppose the matrix X is partitioned as (X_1, X_2) where X_1 is I×r and X_2 is I×u (r + u = s), then the model (1) may be written as

$$\underset{\sim}{\nu} = X\underset{\sim}{\beta} = X_1\underset{\sim}{\beta}_1 + X_2\underset{\sim}{\beta}_2, \tag{16}$$

where $\underset{\sim}{\beta}_1$ is r×1 and $\underset{\sim}{\beta}_2$ is u×1. We are often interested in testing the null hypothesis H: $\underset{\sim}{\beta}_2 = 0$ given the model (16). The "pseudo m.l.e." under H can be obtained from the equations

$$X_1'D(w)\hat{\hat{\underset{\sim}{f}}} = X_1'D(w)\hat{\underset{\sim}{p}} \tag{17}$$

again by iterative calculations, where $\hat{\hat{\underset{\sim}{f}}} = f(\hat{\hat{\underset{\sim}{\beta}}})$. The standard chisquared and likelihood ratio tests of H: $\underset{\sim}{\beta}_2 = 0$ are given by

$$X^2(2|1) = n \sum_{i=1}^{I} \frac{w_i(\hat{f}_i - \hat{\hat{f}}_i)^2}{\hat{\hat{f}}_i(1 - \hat{\hat{f}}_i)} \tag{18}$$

and

$$G^2(2|1) = 2n \sum_{i=1}^{I} w_i \left\{ \hat{f}_i \ln \frac{\hat{f}_i}{\hat{\hat{f}}_i} + (1 - \hat{f}_i)\ln \frac{(1 - \hat{f}_i)}{(1 - \hat{\hat{f}}_i)} \right\} \tag{19}$$

respectively. Under binomial sampling, both $X^2(2|1)$ and $G^2(2|1)$ are asymptotically distributed as χ^2 with u d.f. when H is true, but for general designs this result is no longer valid. In fact $X^2(2|1)$ or $G^2(2|1)$ is asymptotically distributed as a weighted sum, $\Sigma\delta_i(H)Z_i$, of independent χ_1^2 variables Z_i, where the weights $\delta_i(H)$ (i =1, ..., u) are the eigenvalues of the design effects matrix.

$$(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}(\tilde{X}_2'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\tilde{X}_2), \tag{20}$$

where

$$\tilde{X}_2 = [I - X_1(X_1'\Delta X_1)^{-1}X_1'\Delta]X_2, \tag{21}$$

(Roberts, 1984). In the binomial case, the design effects matrix (20) reduces to I, as in the previous case of goodness-of-fit.

A simple adjustment to $X^2(2|1)$ or $G^2(2|1)$ is obtained by treating $X_c^2(2|1) = X^2(2|1)/\delta_.(H)$ or $G_c^2(2|1)/\delta_.(H)$ as χ^2 with u d.f. under H, where

$$u\,\delta_{\cdot}(H) = n \sum_{i=1}^{I} \tilde{V}_{ii}(r) w_i / \hat{\hat{f}}_i (1 - \hat{\hat{f}}_i) \tag{22}$$

and $\tilde{V}_{ii}(r)$ is the i-th diagonal element of the covariance matrix of residuals, $r_i(H) = \hat{f}_i - \hat{\hat{f}}_i$, given by

$$\tilde{V}(r) = D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}})\tilde{X}_2 A \tilde{X}_2' D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1} - \hat{\underset{\sim}{f}}) \tag{23}$$

where

$$A = (\tilde{X}_2'\Delta\tilde{X}_2)^{-1}[\tilde{X}_2' D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})\tilde{X}_2](\tilde{X}_2'\Delta\tilde{X}_2)^{-1} \tag{24}$$

The standardized residuals $(\hat{f}_i - \hat{\hat{f}}_i)/[\tilde{V}_{ii}(r)]^{\frac{1}{2}}$ can also be computed. As in the case of goodness-of-fit, improved approximation, based on Satterthwaite's method, can also be obtained.

A Wald statistic of $H: \underset{\sim}{\beta}_2 = 0$ is given by

$$X_W^2(2|1) = \hat{\beta}_2'[\hat{D}(\hat{\underset{\sim}{\beta}}_2)]^{-1}\hat{\underset{\sim}{\beta}}_2. \tag{25}$$

where $\hat{D}(\hat{\underset{\sim}{\beta}}_2)$ is the principal submatrix in (5) corresponding to $\hat{\underset{\sim}{\beta}}_2$. Under H, $X^2(2|1)$ is asymptotically distributed as χ^2 with u d.f. In particular if β_2 is a scalar, we can treat $\hat{\beta}_2/\text{s.e.}(\hat{\beta}_2)$ as N(0,1)-variate under the hypothesis $H: \beta_2 = 0$ or $\hat{\beta}_2^2/\text{var}(\hat{\beta}_2)$ as χ^2 with 1 d.f.

2.4 Diagnostics

It is desirable to make a critical assessment of the logit fit by identifying any outlying cell proportions and influential points in the factor space. For this purpose, the vector of residuals and a projection matrix in the factor space provide useful tools. However, unlike in the case of the standard linear model, the residuals can be defined on different scales. The natural choice that takes account of the survey design is the vector of standardized residuals $e_i = r_i/[\hat{V}_{ii}(r)]^{\frac{1}{2}}$ given in section 2.1. Since the e_i's are

approximately N(0, 1) under the model (1), the expected numbers of residuals e_i exceeding 1.96, 2.33 and 2.58 in magnitude are 0.05I, 0.02I and 0.01I respectively, where I is the number of residuals (cells). These expected numbers provide a rough guide to identify any outlying cells. Ignoring the design and hence using standardized residuals under binomial sampling could lead to misleading conclusions.

The standardized residuals e_i, however, become unreliable for those cells with $\hat{p}_i = 1$ or close to 1. Following Pregibon (1981), we suggest the use of components of X_c^2 or G_c^2, viz., $\tilde{X}_i = X_i/\delta_\cdot^{\frac{1}{2}}$ or $\tilde{G}_i = G_i/\delta_\cdot^{\frac{1}{2}}$, i = 1, ..., I, for residual analysis in order to circumvent this difficulty. In either case, large individual components should roughly indicate cells poorly accounted for by the model. Index plots (i.e., plots of $\tilde{X}_i$ vs i and $\tilde{G}_i$ vs i) are useful for displaying these components. Normal probabilities plot of $\tilde{X}_i$ or $\tilde{G}_i$ (i.e., the ordered values plotted against standard normal quantiles) is also useful to detect deviations from the model (i.e., deviations from a straight-line configuration).

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

$$M = I - \hat{V}_b^{\frac{1}{2}} X(X'\hat{V}_b X)^{-1} X'\hat{V}_b^{\frac{1}{2}}$$

$$= I - H \text{ (say)} \qquad (26)$$

to detect influential points, where $\hat{V}_b$ is the estimated covariance matrix under binomial sampling, viz., $\text{diag}[\hat{p}_1(1 - \hat{p}_1)/(nw_1), \ldots, \hat{p}_I(1 - \hat{p}_I)/(nw_I)]$ in the context of survey data. The matrix M arises naturally in solving likelihood equations (4) by iteratively reweighted least squares, and small values of m_{ii} call attention to extreme points in the factor space. Again, an index plot (m_{ii} vs i) would provide a useful display. It may be noted that the design effect does not come into picture with m_{ii} since we are using "pseudo m.l.e." based on binomial sampling. Another useful plot which effectively summarizes the information in the index plots $\tilde{X}_i$ vs i and m_{ii} vs i is given by the scatter plot of $\tilde{X}_i^2/X_c^2 = X_i^2/X^2$ vs h_{ii}, where h_{ii} is the i-th diagonal element of H given by (26) (see Pregibon, 1981).

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, $\hat{\underset{\sim}{\beta}}$, fitted values, $\hat{\underset{\sim}{f}}$, and goodness-of-fit measures $X^2/\delta_.$ or $G^2/\delta_.$ or others. Following Pregibon (1981) we suggest three measures which quantify the effect of extreme cells (points) on the fit.

(1) Coefficient sensitivity: Let $\hat{\beta}_j(-\ell)$ denote the pseudo m.l.e. of β_j obtained after deleting the ℓ-th cell data. Then the quantity $\Delta_j(\ell) = [\hat{\beta}_j - \hat{\beta}_j(-\ell)]/\text{s.e.}(\hat{\beta}_j)$ provides a measure of the j-th coefficient sensitivity to ℓ-th point. The index plots $\Delta_j(\ell)$ vs ℓ for each j provide useful displays but the task of looking at the index plots could become unmanageable if the number of coefficients in the model is large.

(2) Sensitivity of fitted values: Significant changes in coefficient estimates when ℓ-th point (cell) deleted does not necessarily imply that the fitted values $\hat{\underset{\sim}{f}}$ also vary significantly from $\hat{\underset{\sim}{f}}(-\ell)$, the vector of fitted values obtained after deleting the ℓ-th cell, i.e., $\|\hat{\underset{\sim}{f}} - \hat{\underset{\sim}{f}}(-\ell)\|$ could be small. We therefore use $[G^2 - \tilde{G}^2(-\ell)]/\delta_.$ or $[X^2 - \tilde{X}^2(-\ell)]/\delta_.$ to assess the impact of the ℓ-th point on the fitted values. where $\tilde{G}^2(-\ell)$ and $\tilde{X}^2(-\ell)$ are given by (9) and (8) respectively when $\hat{f}_i = f_i(\hat{\underset{\sim}{\beta}})$ is replaced by $\hat{f}_i(-\ell) = f_i(\beta(-\ell))$.

(3) Goodness-of-fit: A measure of goodness-of-fit sensitivity is given by $[G^2 - G^2(-\ell)]/\delta_.$ or $[X^2 - X^2(-\ell)]\delta_.$, where $G^2(-\ell)$ and $X^2(-\ell)$ are the likelihood ratio and chisquared statistics obtained after deleting the ℓ-th cell. (Note that $G^2(-\ell) \neq \tilde{G}^2(-\ell)$).

3. APPLICATION TO LFS

We have applied the previous methods to some data from the October 1980 Canadian Labour Force Survey (LFS). The sample consisted of males aged 15-64 who were in the labour force and not full-time students. We have chosen two factors, age and education, to explain the variation in unemployment rates via logit models. Age-group levels were formed by dividing the interval [15, 64] into ten groups with the j-th age group being the interval $[10 + 5j, 14 + 5j]$, $j = 1, 2, \ldots, 10$, and then using the mid-point of each interval, A_j, as the value of the 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. Thus the age by education cross-classification provided a two-way table of I = 60 cell proportions, π_{jk}.

The LFS design employed stratified multi-stage cluster sampling with two stages in the self-representing (SR) urban areas and three or four stages in non-self-representing (NSR) areas in each province. The survey estimates, $\hat{p}_{jk}$, were adjusted for post-stratification, using the projected census age-sex distribution at the provincial level. The estimated covariance matrix $\hat{V}$ of the estimates $\hat{p}_{jk}$ is based on more than 450 first-stage units (psu's) so that the degrees of freedom for $\hat{V}$ are large compared to I = 60.

3.1 Formal Tests of Hypotheses.

Scatter plot of the logits $\hat{\nu}_{jk}$ vs age levels A_j at each education level E_k indicated that $\hat{\nu}_{jk}$ for given k generally increases with age to a maximum and then decreases (i.e., the graph is convex and upward to a maximum). Hence, the following model might be suitable to explain the variation in π_{jk}'s.

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k + \beta_4 E_k^2,$$

$$j = 1, \ldots, 10; \quad k = 1, \ldots, 6. \qquad (27)$$

Some previous work in sociological literature also supports such a model (Bloch and Smith, 1977). Applying the results of Section 2 we obtained the following values for goodness-of-fit statistics

$$X^2 = 98.9 \qquad G^2 = 101.2$$
$$X^2/\delta_* = 52.5 \qquad G^2/\delta_* = 53.7, \quad \delta_* = 1.88.$$

Since X^2 or G^2 is larger than $\chi^2_{0.05}(55) = 73.3$, the upper 5% point of χ^2 with I - s = 55 d.f., we would reject the model if the survey design is ignored. On the other hand, the value of X^2/δ_* or G^2/δ_* indicate that the model is adequate, the significance level (or P-value) being approximately equal

to 0.52. The value of X_S^2 when adjusted to refer to $\chi^2_{0.05}(55)$ is equal to 47.7 which is also not significant. Moreover, in the present context with s(= 5) relatively small compared to I(= 60), the simple correction $\bar{d}$, the average cell deff, (see Fellegi, 1980), is very close to $\bar{\delta}$: $\bar{d}$ = 1.905 compared to $\bar{\delta}$ = 1.88: see Rao and Scott (1984) for a theoretical explanation.

The Wald statistic X_W^2 is not defined here since two of the cells have $\hat{p}_{jk} = 1$, but we made minor perturbations to the estimated counts to ensure that $\hat{p}_{jk} < 1$ for all cells and then computed X_W^2. The resulting values of X_W^2 are all large compared to $X^2/\delta_.$ (at least 30 times larger than $X^2/\delta_.$) and vary considerably (1715 to 3061). Hence, the Wald statistic is very unstable for goodness-of-fit test in the present context. If the two cells having $\hat{p}_{jk} = 1$ are deleted, then $X_W^2 = 68.4 < \chi^2_{0.05}(53) = 71.0$, indicating that the model (27) is adequate. However, it is not a good practice to delete cells just to accomodate a chosen test statistic. The other problem with X_W^2, noted by Fay (1983), does not arise here since d.f. for $\hat{V}$ is large compared to the number of cells in the table.

The pseudo m.l.e., their s.e. and the corresponding s.e. under binomial sampling, all obtained under the model (27), are given in Table 1 along with Wald statistic $X_W^2(2|1)$ and G^2 statistic $G^2(2/1)/\delta_.(H)$ for the hypotheses H_i: $\beta_i = 0$, i=1, 2, 3, 4 given the model (27). As expected, the true s.e.'s are larger than the corresponding binomial s.e.'s. The hypothesis H_4: $\beta_4 = 0$ (i.e., coefficient of E_i^2 is zero) is not rejected at the 5% level either by the Wald statistic or G^2 statistic. On the other hand, the coefficient, β_2, of A_i^2 is highly significant. In testing the significance of individual coefficients we compare the values of $X_W^2(2|1)$ or $G^2(2|1)/\delta_.(H)$ to $\chi^2_{0.05}(1) = 3.84$, the upper 5% point of χ^2 - variate with 1 d.f.

We have also tested the following nested hypotheses given model (27): H_{34}: $\beta_3 = \beta_4 = 0$ (i.e., no education effect); H_{24}: $\beta_2 = \beta_4 = 0$ (i.e., no quadratic effects). Both H_{34} and H_{24} are highly significant:

$$G^2(2|1)/\delta_.(H_{34}) = 282.2/1.64 = 172.1,\ X_W^2(2|1) = 165.6 \text{ for } H_{34}:$$

$$G^2(2|1)/\delta_.(H_{24}) = 242.2/2.28 = 106.3,\ X_W^2(2|1) = 162.1 \text{ for } H_{24} \text{ compared to } \chi^2_{0.05}(2) = 5.99.$$

Table 1: Pseudo m.l.e. $\hat{\beta}_i$, s.e. $(\hat{\beta}_i)$, $X_W^2(2|1) = \hat{\beta}_i^2/\text{var}\,(\hat{\beta}_i)$ and $G^2(2|1)/\delta_\cdot(H_i)$ Values for the LFS Data under Model (27).

	$\hat{\beta}_i$	s.e.$(\hat{\beta}_i)$		$X_W^2(2\|1)$	$G^2(2\|1)/\delta_\cdot(H_i)$
		True	Binomial		
0	-2.76	0.557		24.6	
1	0.209	0.0132	0.012	250.6	168.4
2	-0.00217	0.000173	0.000136	157.3	102.1
3	0.0913	0.0891	0.068	1.04	1.01
4	0.00276	0.00411	0.0030	0.45	0.46

Unlike in the case of goodness-of-fit, the Wald statistics is stable for testing nested hypotheses and leads to values close to the corresponding $G^2(2|1)/\delta_\cdot(H)$ values.

By the above test of goodness-of-fit and tests of nested hypotheses we have arrived at the following simple model involving only four parameters:

$$\nu_{jk} = \ln \frac{\pi_{jk}}{1 - \pi_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k, \qquad (28)$$

with $\hat{\beta}_0 = -3.10$, $\hat{\beta}_1 = 0.211$, $\hat{\beta}_2 = -0.00218$ and $\hat{\beta}_3 = 0.1509$ and corresponding standard errors are 0.247, 0.0130, 0.000172, and 0.0115. We will use the model (28) in Section 3.2 to develop logistic regression diagnostics.

3.2 Diagnostics

We now illustrate the use of diagnostics developed in Section 2.4.

(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 (28) arrived through

formal testing of hypotheses. Among the sixty e_i, cells numbered 6 and 54 with $\hat{p}_{jk} = 1$ lead to very large e_i values: 166.6 and 6.2 respectively. Among the remaining e_i, the residuals numbers 7, 27 and 59 have values 3.84, 2.73 and 2.52 respectively, whereas the expected number of $|e_i|$ exceeding 2.33 under model (28) is roughly $0.02 \times 60 = 1.2$. Hence, there is some indication that cells 7 and 27 could correspond to outlying cell proportions.

The normal probability plot of $\tilde{G}_i$ is displayed in FIG. 1; the plot of $\tilde{\chi}_i$ is not given to save space since it is similar to the plot of $\tilde{G}_i$. Figure 1 indicates no strong deviations from a straight line configuration. The index plot of $\tilde{G}_i$, Figure 2, is consistent with Figure 1. Hence, there is no evidence of outlying cell proportions when the components $\tilde{G}_i$ of G_c^2 are used for residual analysis.

(ii) Detection of Influential Cells.

The index plot of m_{ii} is displayed in Figure 3 which clearly points to cells 1 and 6. Figure 4 displays the plot of $\tilde{\chi}_i^2/\chi_c^2 = \chi_i^2/\chi^2$ vs h_{ii}, where the line with slope - 1 is given by $\chi_i^2/\chi^2 + h_{ii} = 3\text{ave}(h^*_{ii})$. Here $h^*_{ii} = h_{ii} + \chi_i^2/\chi^2$, and the values of h^*_{ii} near unity corresponds to cells which are outlying or influential or both (Pregibon, 1981) and appear above the line in Figure 3. It is clear that cells 1 and 6, and to a lesser extent cells 7 and 58, warrant further examination.

(iii) Coefficient Sensitivity.

The index plots for measuring coefficient sensitivity ($\Delta_j(\ell)$ vs ℓ) are displayed in Figures 5, 6, 7, and 8 for β_0, β_1, β_2 and β_3 respectively. It is clear from the plots that cells 2 and 3 cause instability in $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, whereas $\hat{\beta}_3$ is affected by cell 7.

(iv) Sensitivity of Fitted Values

Figure 9 displays the plot of $[G^2 - \tilde{G}^2(-\ell)]/\delta_\cdot = c$ vs ℓ for assessing the impact of individual cells on fitted values. Significant peaks in this figure correspond to cells 2 and 3 and to a lesser extent to cell 7. Following Cook (1977) and Pregibon (1981), it may be noted that the comparison of c to the percentage point of $\chi^2(s)$ (s = 4 in model (28)) gives a rough guide as to which contour of the confidence region the pseudo m.l.e. is displaced due to deletion of the ℓ-th cell. The value c = 2.1 for cell 2 roughly corresponds to 78% contour of the confidence region.

(v) Goodness-of-fit Sensitivity

Figure 10 displays the plot of $[G^2 - G^2(-\ell)]/\delta_.$ vs ℓ: the plot of $[X^2 - X^2(-\ell)]/\delta_.$ is similar and hence not displayed but the former plot is preferred (Pregibon, 1981). Significant peaks in this figure corresponds to cells 2, 3, 7, 27, 39 and 54 (values $\geq$ 3), the most significant being cell 7 with the value 5.4. By deleting cell 7 and recomputing the adjusted statistic $G^2_c(-\ell) = G^2(-\ell)/\delta_.(-\ell)$ where $\delta_.(-\ell)$ is the corresponding value of $\delta_.$, we get a value of 48.43 with 55 d.f. compared to $G^2/\delta_. = 55.3$ with 56 d.f.

Our investigation on the whole indicated that cells 7, 2 and 3 are possible candidates for deletion, but we feel that their impact is not significant enough to warrant their deletion - one would like to explain the variation among all cell proportions unless certain cells contribute heavily to the disagreement between the data and the fitted model.

ACKNOWLEDGEMENT

We wish to thank M. Gratton of Statistics Canada for producing the graphs included in the paper.

REFERENCES

[1] Bloch, F.E., and Smith, S.P. (1977). Human capital and labour market employment. J. Human Resources, 12, pp. 550-559.

[2] Cook, R.D. (1977). Detection of influential observations in linear regression. J. American Statistical Association, 72, pp. 169-174.

[3] Cook, R.D., and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.

[4] Cox, D.R. (1970). Analysis of Binary Data. Chapman and Hall, London.

[5] Fay. R.E. (1983). Replication approaches to the log-linear analysis of data from complex samples. Unpublished manuscript (courtesy of the author).

[6] Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. J. American Statistical Association, 75, pp. 261-268.

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

[8] Pregibon, D. (1981). Logistics regression diagnostics. Ann. Statist., 9, pp. 705-724.

[9] Rao, J.N.K., and Scott, A.J. (1984). On simple adjustments to chisquared tests with survey data: log-linear and logit models. Unpublished manuscript.

[10] Roberts, G. (1984). On chi-squared tests for logit models with cell proportions estimated from survey data. Unpublished manuscript. Carleton University.

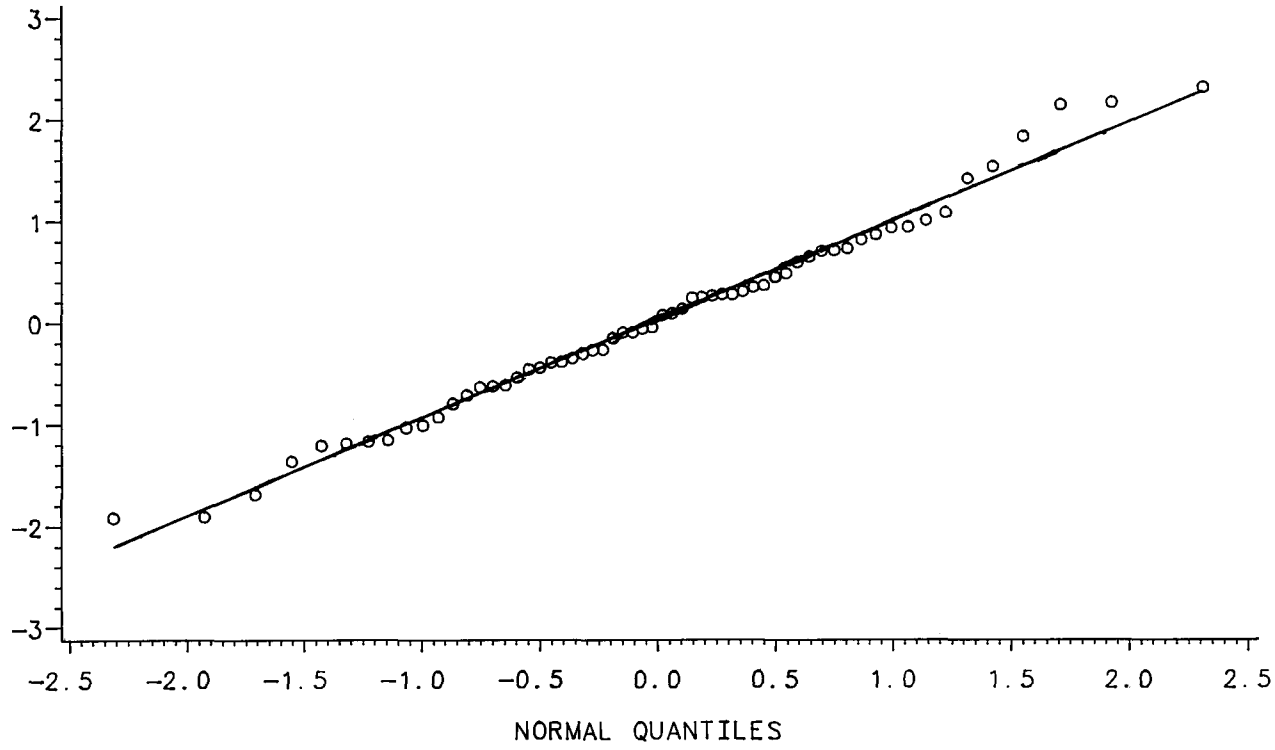

Figure 1: Normal Probability Plot of $\widetilde{G}_i$

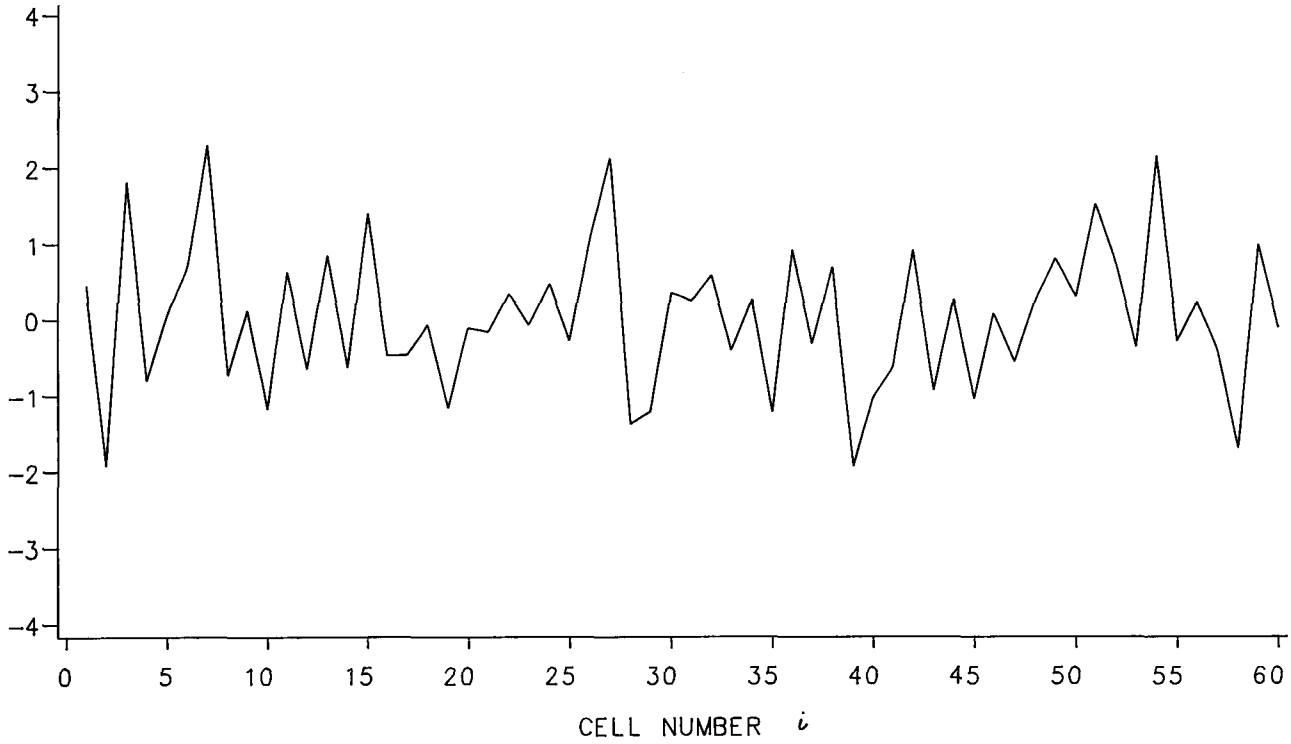

Figure 2: Index Plot of $\widetilde{G}_i$

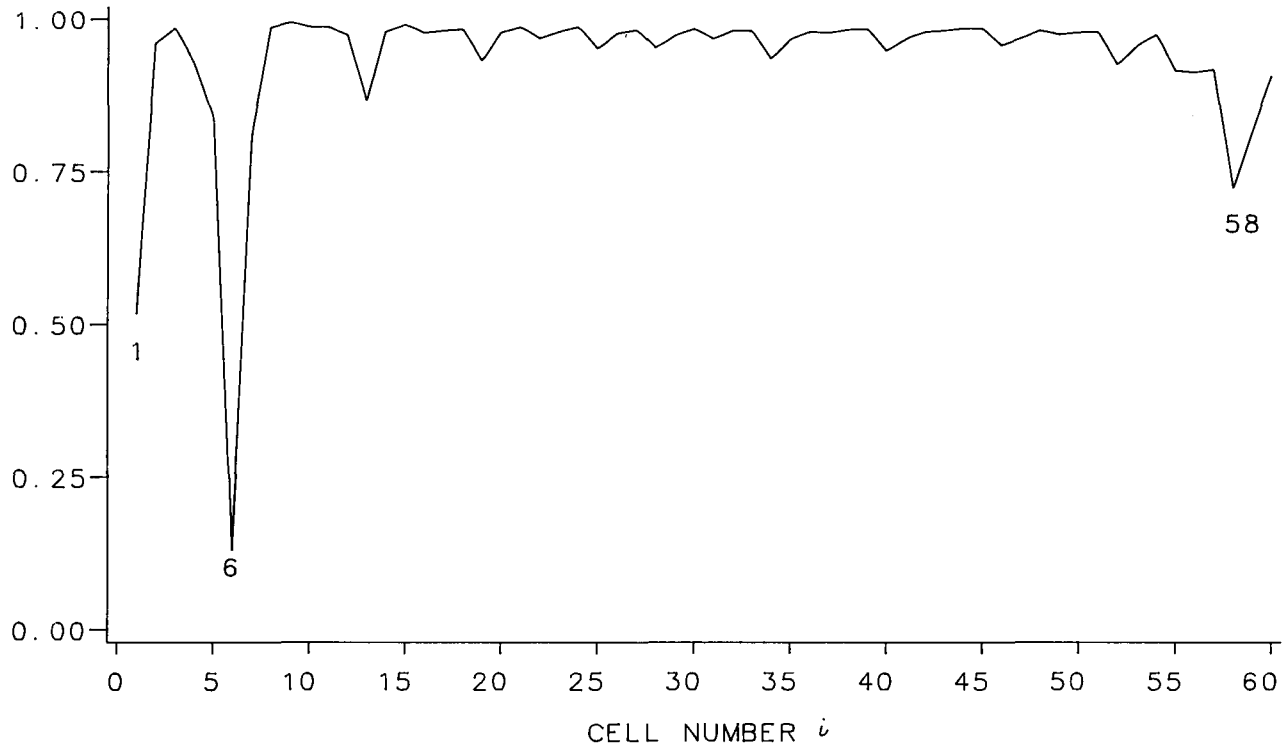

Figure 3: Index Plot of m_{ii}

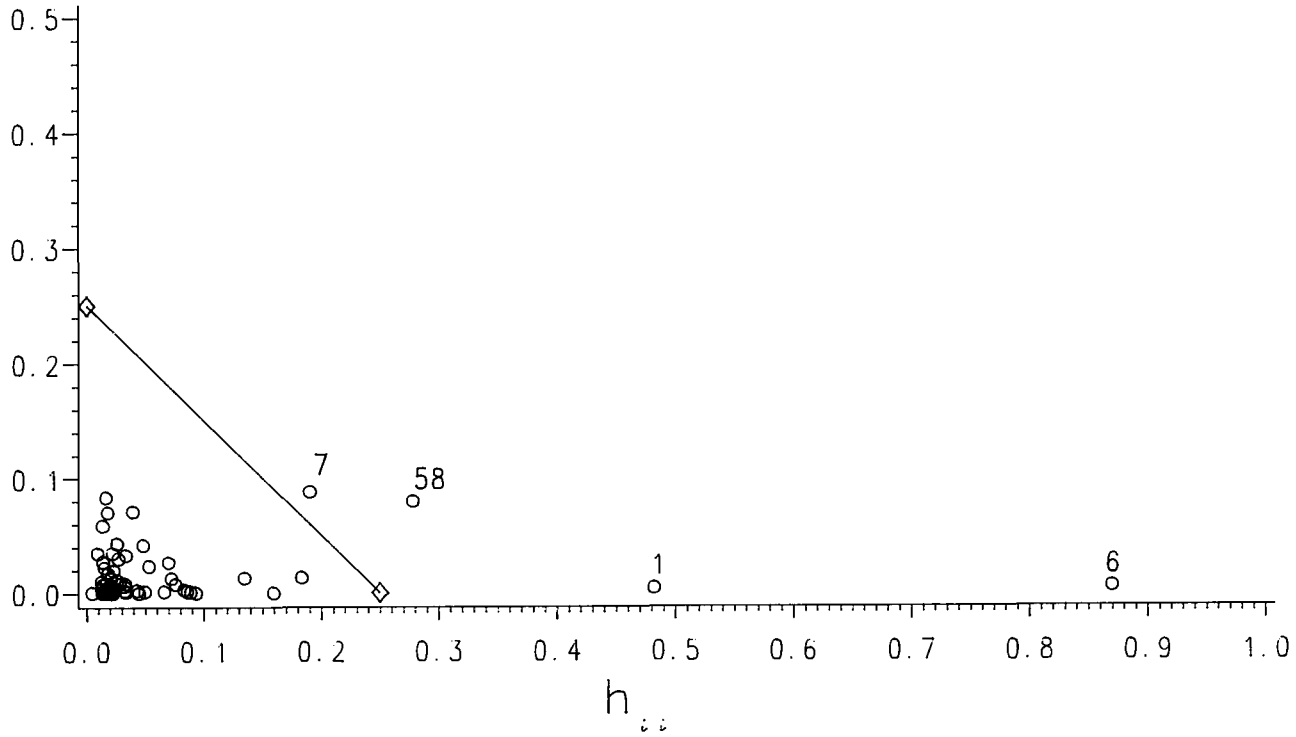

Figure 4: Scatter Plot of χ_i^2/χ^2 vs. h_{ii}

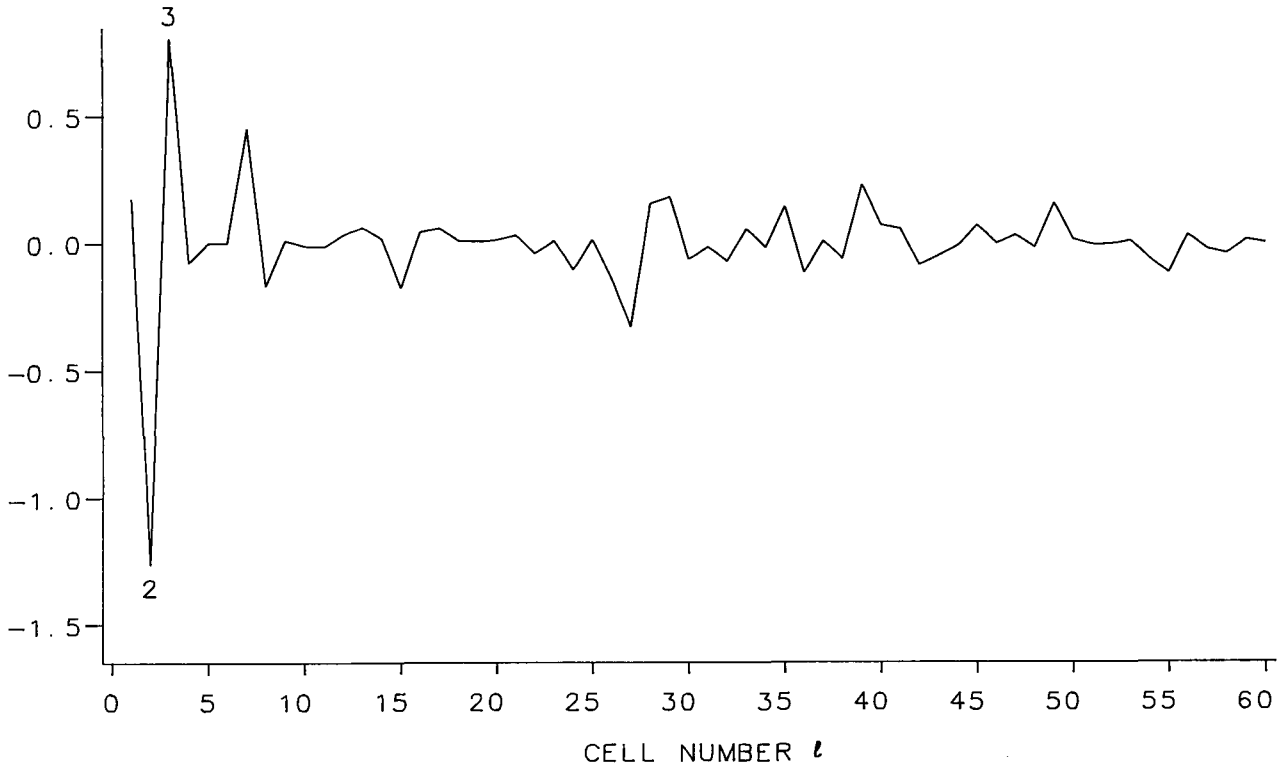

Figure 5: Index Plot of $\{\hat{\beta}_0 - \hat{\beta}_0(-\ell)\}/\text{s.e.}(\hat{\beta}_0)$

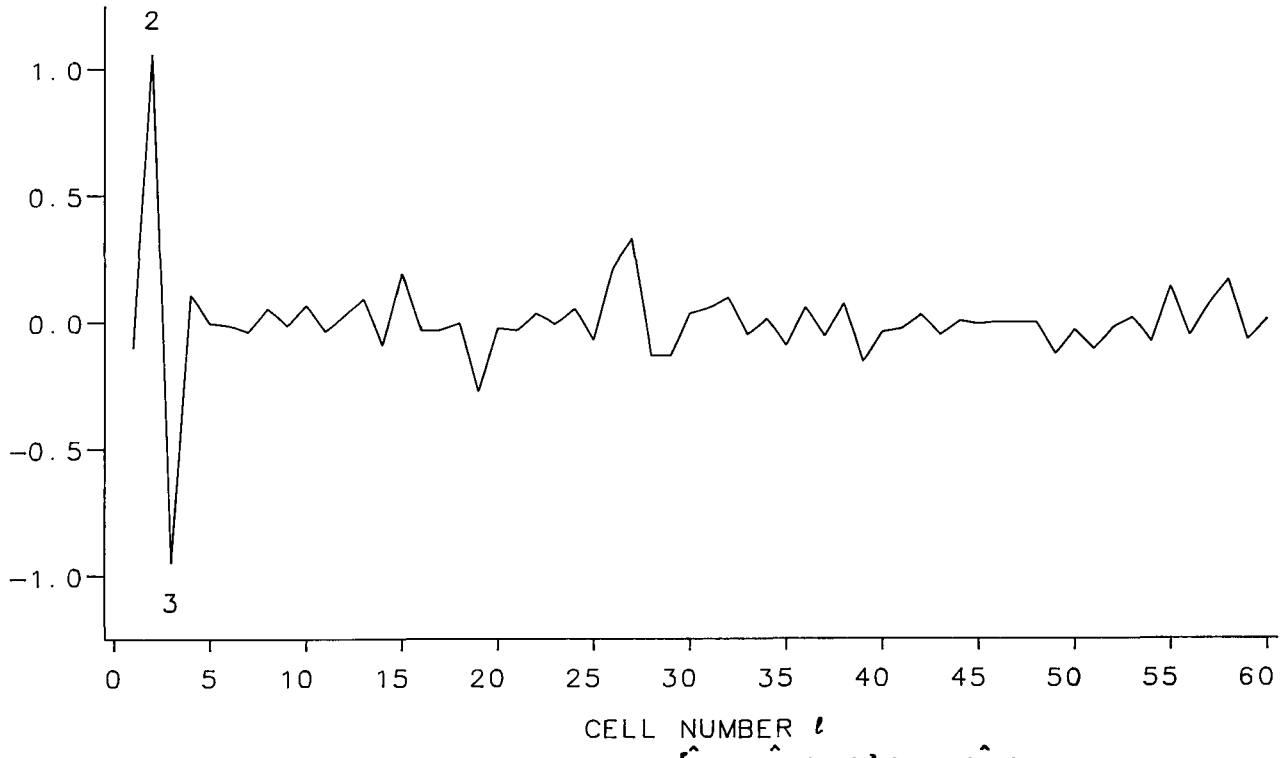

Figure 6: Index Plot of $\{\hat{\beta}_1 - \hat{\beta}_1(-\ell)\}/\text{s.e.}(\hat{\beta}_1)$

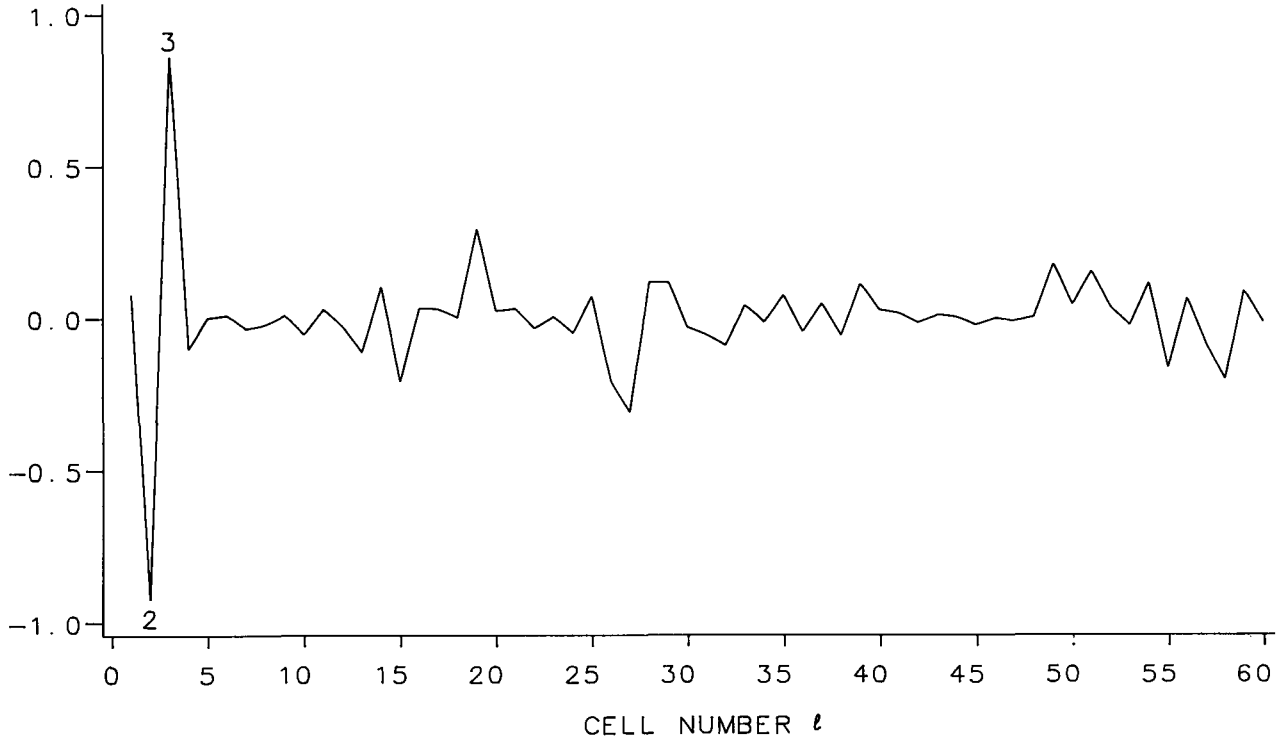

Figure 7: Index Plot of $\{\hat{\beta}_2 - \hat{\beta}_2(-\ell)\}/\text{s.e.}(\hat{\beta}_2)$

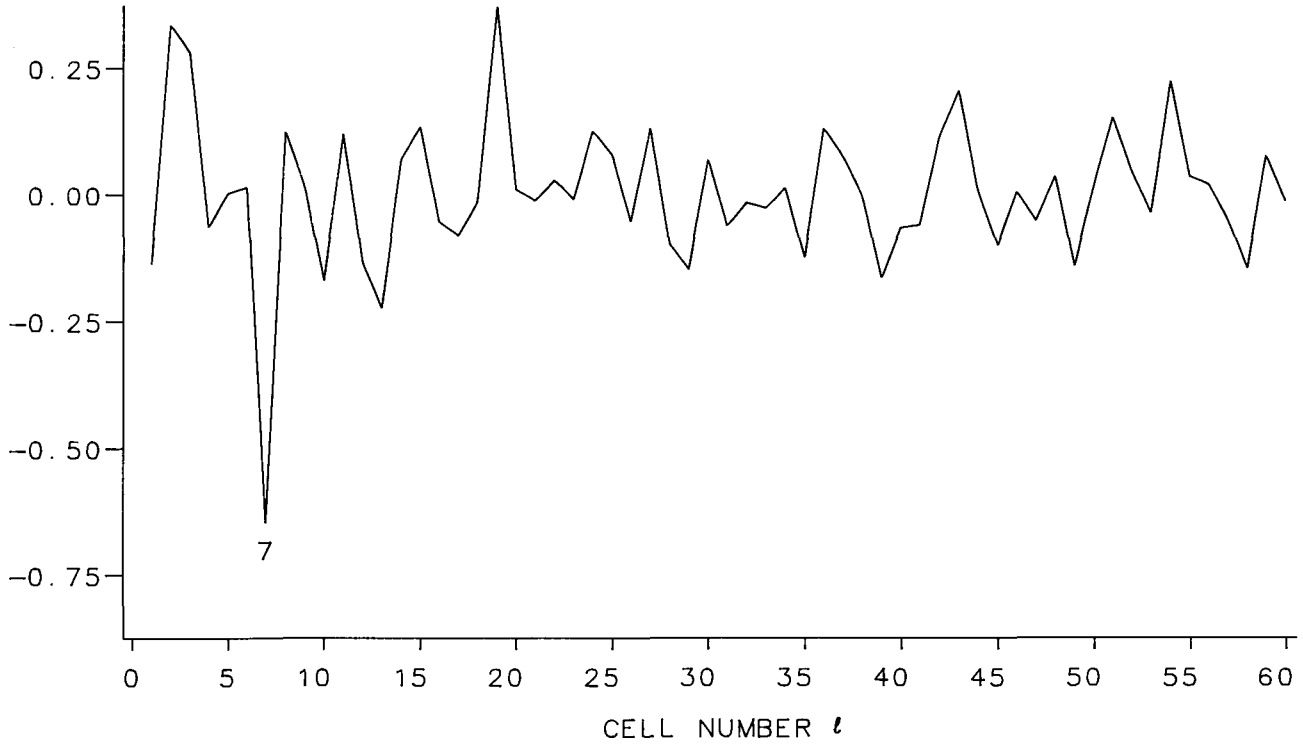

Figure 8: Index Plot of $\{\hat{\beta}_3 - \hat{\beta}_3(-\ell)\}/\text{s.e.}(\hat{\beta}_3)$

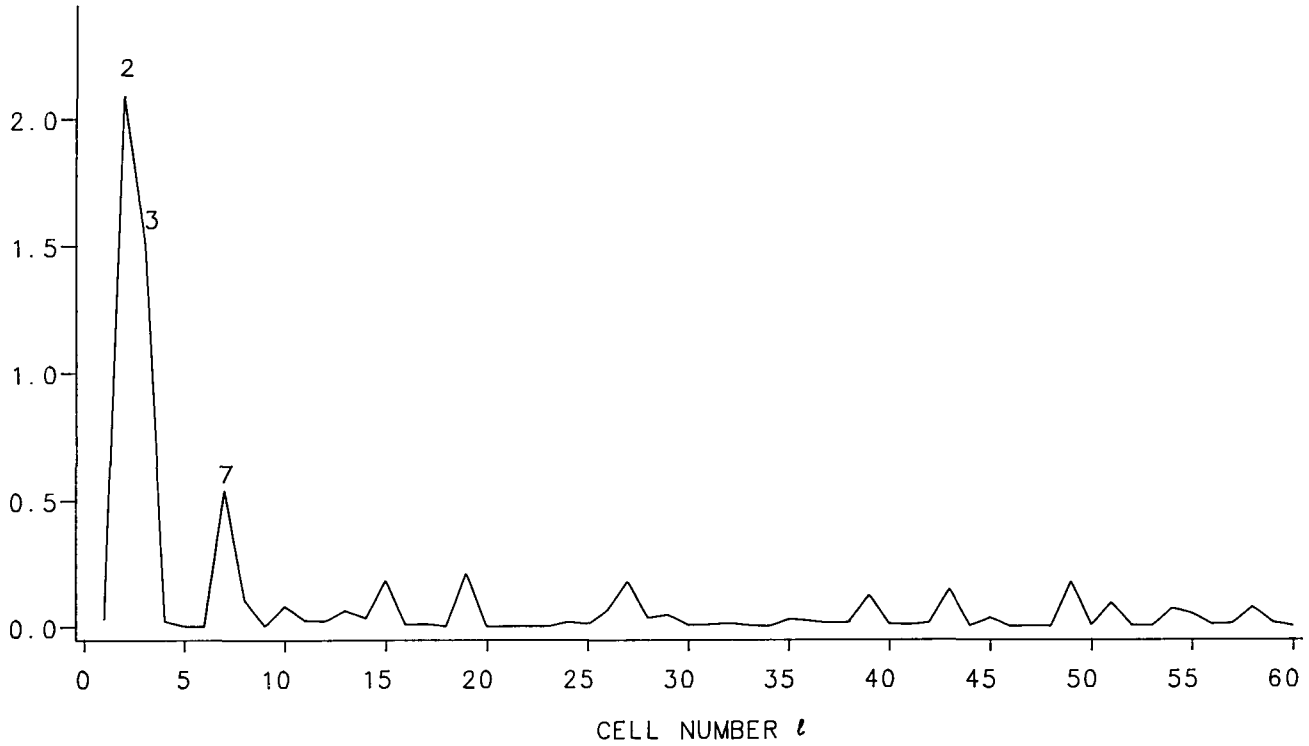

Figure 9: Index Plot of $\{G^2 - \tilde{G}^2(-\ell)\}/\hat{\delta}$

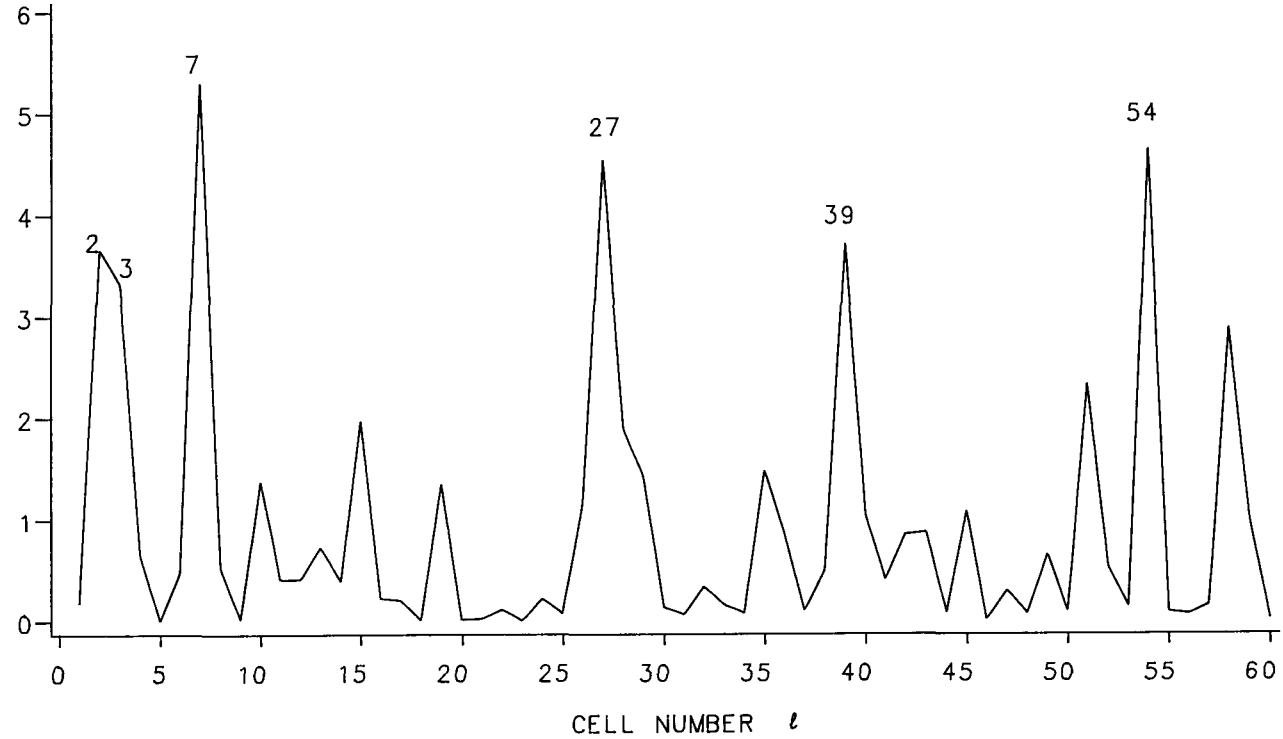

Figure 10: Index Plot of $\{G^2 - G^2(-\ell)\}/\hat{\delta}$