Results on point and interval estimation for log-linear models with non-ignorable non-response

Paul S. Clarke and Peter W. F. Smith

Abstract

It is common that log-linear models for multi-way contingency tables with one variable subject to non-ignorable non-response will yield non-response boundary solutions, where the probability of non-respondents being classified in certain cells of the table is estimated to be zero, resulting in infinite estimates for some of the log-linear parameters. This paper investigates the effect of such non-standard behaviour on the maximum likelihood estimator. Provided that the model parameters are identifiable from infinite samples, it is demonstrated that: 1) existence and uniqueness of the maximum likelihood estimates is assured under weak conditions; and 2) the maximum likelihood estimator is consistent and asymptotically normal. However, boundary solutions do result in a singular information matrix, which prevents calculating confidence intervals based on a normal approximation to the maximum likelihood estimator; it is shown that these singularities can be removed by a simple transformation of the log-linear parameters.

S^{3}RI Methodology Working Paper M03/23

Results on point and interval estimation for log-linear models with non-ignorable non-response

Paul S. Clarke

Department of Infectious Disease Epidemiology, Imperial College London, St. Mary's Campus, Norfolk Place, London W2 1PG, U.K.

Peter W.F. Smith

Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, U.K.

Abstract:

It is common that log-linear models for multi-way contingency tables with one variable subject to non-ignorable non-response will yield non-response boundary solutions, where the probability of non-respondents being classified in certain cells of the table is estimated to be zero, resulting in infinite estimates for some of the log-linear parameters. This paper investigates the effect of such non-standard behaviour on the maximum likelihood estimator. Provided that the model parameters are identifiable from infinite samples, it is demonstrated that: 1) existence and uniqueness of the maximum likelihood estimates is assured under weak conditions; and 2) the maximum likelihood estimator is consistent and asymptotically normal. However, boundary solutions do result in a singular information matrix, which prevents calculating confidence intervals based on a normal approximation to the maximum likelihood estimator; it is shown that these singularities can be removed by a simple transformation of the log-linear parameters.

Keywords: Boundary solutions; Categorical data; Consistency; Informative non-response; Normal approximation.

1 Introduction

Consider a multi-way contingency table cross-classified by k variables, $X_1,\ldots, X_k$, each with nominal categorical outcomes. If $Y = X_k$ is subject to non-response then the observed data are incompletely classified. Baker and Laird (1988) introduced a family of log-linear models for incomplete tables of this type, which have been considered by other authors (Baker et al. 1992; Park and Brown 1994; Clarke 1998; Forster and Smith 1998; Park 1998; Smith et al. 1999; Clarke 2002; Clarke and Smith 2004). Such 'Baker and Laird' models are important because they permit estimation of the coefficients in the regression of Y on $X_1,\ldots, X_{k-1}$ under non-ignorable non-response. In Baker and Laird models, non-response is ignorable if the probability of missing Y is independent of Y given $X_1,\ldots, X_{k-1}$; hence, it follows that non-response is non-ignorable if the probability of Y being missing depends on Y itself, even after controlling for $X_1,\ldots, X_{k-1}$. (The conditions required for valid inference under ignorable non-response are delineated by Rubin (1976).)

Baker and Laird models can produce extreme and implausible point estimates called non-response boundary solutions. A non-response boundary solution occurs when the probability of non-respondents being classified in certain cells of the table is estimated to be zero. For example, non-ignorable models were used by Smith et al. (1999) to estimate the proportion voting for each party in a pre-election survey where voting intention was missing for a subset of sampled individuals; the maximum likelihood (ML) estimate of the proportion of non-respondents intending to vote for the Labour Party was estimated to be zero, despite the model fitting the data well. A simulation study by Clarke (1998) showed that the probability of a boundary solution could be non-zero and non-negligible in finite samples, even when the Baker and Laird model is the true population model.

A consequence of such non-standard behaviour is that point estimates of certain log-linear parameters can be $\pm\infty$. Infinite estimates can also occur when fitting log-linear models to completely classified but sparse contingency tables; that is, observed tables with one or more cells having frequency zero. Difficulties on determining the existence, uniqueness and calculation of ML estimates for sparse tables were resolved by developing a theory of 'extended' ML estimation (Haberman 1974, Wedderburn 1976, Clarkson and Jennrich 1991). In this paper, these issues are considered for non-response boundary solutions in a simple non-ignorable Baker and Laird model. The findings for non-response boundary solutions and sparse contingency tables are compared in the discussion.

Also in this paper, the usual properties of consistency and asymptotic normality of the ML estimator are demonstrated using geometric arguments. The asymptotic normality property permits the use of approximate confidence intervals based on the normal distribution. Clarke and Smith (2004) considered the use of normal intervals for a simple Baker and Laird model, and found that boundary solutions preclude calculation of 'normal intervals' for log-linear parameters because the

information matrix becomes singular. In the last part of this paper, it is demonstrated why an exponential transformation obviates this problem.

2 Preliminaries

2.1 Notation

Recall that $X_1,\ldots, X_{k-1}$ and Y are nominal categorical variables and that Y is incompletely observed. Let Y have levels numbered from 1 to q, and let X_j have levels numbered 1 to s_j ($j = 1,\ldots, k-1$) with $s = \prod_j s_j$. Further define the response indicator variable as R, equal to 1 if Y is observed and 2 otherwise. The resulting $(k+1)$-way contingency table of $X_1,\ldots, X_{k-1}$, Y and R is referred to as the complete data table; that is, the hypothetical multi-way table which would have been be observed if the missing outcomes of Y were known. Denote the cell frequencies of the complete data table by the column vector $\mathbf{z} = \{z_{\mathbf{x}yr}\}$, where $z_{\mathbf{x}yr}$ is the (possibly unknown) sample frequency classified by $\mathbf{X} = \mathbf{x}$, $Y = y$ and $R = r$; $\mathbf{X} = (X_1,\ldots, X_{k-1})^{\mathrm{T}}$ and $\mathbf{x} = (x_1, \ldots, x_{k-1})^{\mathrm{T}}$. It is assumed that $\mathbf{z}$ is a realisation of the multinomial random vector $\mathbf{Z} = \{Z_{\mathbf{x}yr}\} \sim \mathrm{MN}(n, \boldsymbol{\pi})$, where $n = z_{+++}$ and $\boldsymbol{\pi} = \{\pi_{\mathbf{x}yr}\}$ are the cell frequencies.

2.2 Model specification

In general, Baker and Laird models are selection models defined by two sub-models called the 'margin' model and the 'non-response' model. The margin model is log-linear for the joint distribution of $\mathbf{X}$ and Y. For $k = 2$ the margin model is defined as

$$\log(\pi_{xy+}) = \nu + \lambda^X(x) + \lambda^Y(y) + \lambda^{XY}(x, y), \tag{1}$$

where $X = X_1$, corner-point constraints fix $\lambda^X(1) = \lambda^Y(1) = \lambda^{XY}(1, y) = \lambda^{XY}(x, 1) = 0$ for all x, y without loss of generality, and ν is the normalising constant ensuring $\sum_{x,y}\pi_{xy+} = 1$. The interest parameters are the sq non-zero log-linear parameters of the margin model. The non-response model is a logistic regression of R on $\mathbf{X}$ and Y, which specifies how Y came to be missing or observed. Returning again to the $k = 2$ example, the regression model can be written

$$\log(\pi_{2|xy} / \pi_{1|xy}) = \lambda^R(2) + \lambda^{XR}(x, 2) + \lambda^{YR}(y, 2) + \lambda^{XYR}(x, y, 2), \tag{2}$$

where $\pi_{r|xy} = \Pr(R = r|X = x, Y = y)$, $\lambda^R(2)$ is the intercept term and $\lambda^{XR}(x, 2)$ is the main-effect of X in the regression, and so on. The non-response parameters are the non-zero parameters in (2), and have

the same corner-point constraints as the interest parameters. Together, (1) and (2) define the cell probabilities for the complete table as $\pi_{xyr} = \pi_{xy+}\pi_{r|xy}$. Generalising (1) and (2) to $k > 2$ is straightforward.

The focus of this paper is on the case where the probability of non-response depends only on Y and the margin model is saturated; that is, all interactions between $\mathbf{X}$ and Y are included in the margin model. For notational simplicity, the multivariate $\mathbf{X}$ is replaced by the single covariate, X, each level of which corresponds to one of the s cross-classifications of $\mathbf{X}$. Using this notation, if non-response depends only on Y, this model is defined by (1) and (2) under the further constraints that $\lambda^{XR}(x, 2) = \lambda^{XYR}(x, y, 2) = 0$ for all x, y.

It was shown by Baker and Laird (1988) that, when $\lambda^{XR}(x, 2) = \lambda^{XYR}(x, y, 2) = 0$, inferences based on the log-linear model

$$\log(\pi_{xyr}) = \nu + \lambda^{X}(x) + \lambda^{Y}(y) + \lambda^{R}(r) + \lambda^{XY}(x, y) + \lambda^{YR}(y, r) \qquad (3)$$

are equivalent to (1) and (2), where ν now ensures that $\sum_{x,y,r}\pi_{xyr} = 1$ and corner-point constraints are used as before. This result holds in general provided that the margin model is saturated. Model (3) is an important special case that has been the focus of previous work. The primary reason for this is that (3) is the most parsimonious non-ignorable model allowing estimation of the unrestricted cell probabilities of the marginal table. The number of free parameters in model (3) is $p = q(s + 1) - 1$, the sum of the non-zero parameters in the margin and non-response models less one (ν is a function of the remaining non-zero parameters). Denote the p free log-linear parameters of model (3) by the vector $\boldsymbol{\lambda} = \{\lambda^{X}(2),\ldots, \lambda^{YR}(q, 2)\}^{\mathrm{T}}$ with corresponding parameter space $\Lambda = \Re^{p}$.

2.3 Maximum likelihood estimation

For cases where $R = 2$, Y is missing and so $\mathbf{z}$ is incompletely observed. Only the cells of the observed data table $\mathbf{z}_{\mathrm{obs}} = (z_{111},\ldots, z_{sq1}, z_{1+2},\ldots, z_{s+2})^{\mathrm{T}}$ are known, which are outcomes from random vector $\mathbf{Z}_{\mathrm{obs}}$. It follows that the log-likelihood for $\boldsymbol{\lambda}$ is (ignoring the additive constant)

$$l(\boldsymbol{\lambda}; \mathbf{z}_{\mathrm{obs}}) = \sum_{x,y} z_{xy1}\log(\pi_{xy1}) + \sum_{x} z_{x+2}\log(\sum_{y}\pi_{\mathrm{xy2}}), \qquad (4)$$

where $\boldsymbol{\pi}$ is defined by (3) (c.f. Baker and Laird 1988). Throughout this paper, $\hat{\boldsymbol{\lambda}} = \{\hat{\lambda}^{X}(2),\ldots,\hat{\lambda}^{YR}(q,2)\}^{\mathrm{T}}$ will be used to denote the ML estimate minimising (4), and also the ML estimator of $\boldsymbol{\lambda}$, depending on the context. Under multinomial sampling, the observed data have $s(q + 1) - 1$ degrees of freedom. Hence, it is necessary for identifiability that $p \le s(q + 1) - 1$, which simplifies to the inequality $s \ge q$; this condition is assumed to hold throughout this paper. It shall

henceforth be assumed that the complete data are realisations from model (3) with true parameter $\boldsymbol{\lambda} \in \Lambda$ and sample size n.

2.4 Non-response boundary solutions

Definition 1: A non-response boundary solution for model (3) is a ML estimate such that $\hat{\pi}_{xy2} = 0$ for at least one combination of x and y (Baker and Laird 1988).

In terms of the log-linear parameters, a non-response boundary solution for model (3) implies that $\hat{\lambda}^{R}(2) + \hat{\lambda}^{YR}(y,2) = -\infty$, and hence that $\hat{\pi}_{+y2} = 0$. If the non-response boundary solution is such that $1 \notin \{y : \hat{\pi}_{+y2} = 0\}$ and $\hat{\pi}_{xy1} > 0$, then this implies that $\hat{\lambda}^{YR}(y,2) = -\infty$ because $\hat{\pi}_{xy1} > 0$ implies $\hat{\lambda}^{R}(2) \in \Re$. However, if $1 \in \{y : \hat{\pi}_{+y2} = 0\}$ and $\hat{\pi}_{xy1} > 0$ then $\hat{\lambda}^{R}(2) + \hat{\lambda}^{YR}(y,2) \in \Re$, but $\hat{\lambda}^{R}(2) = -\infty$ and $\hat{\lambda}^{YR}(y,2) = +\infty$. In other words, neither $\hat{\lambda}^{R}(2)$ nor $\hat{\lambda}^{YR}(y,2)$ is finite, but their sum is; this is an artefact of the corner-point constraints used to identify the log-linear parameters and so not of interest in this paper. Hence, only non-response boundary solutions where $1 \notin \{y : \hat{\pi}_{+y2} = 0\}$ will be considered in this paper; this is done without loss of generality because the levels of Y are nominal and so can be recoded, changing a $\hat{\pi}_{+12} = 0$ solution to an equivalent one where $\hat{\pi}_{+12} \neq 0$.

3 Alternative Model Parameterisation

Under model (3), the cell probabilities can be factorised as

$$\pi_{xyr} = \alpha_r \beta_{y|r} \gamma_{x|y}, \tag{5}$$

where $\alpha_r = \Pr(R = r; \boldsymbol{\lambda})$, $\beta_{y|r} = \Pr(Y = y \mid R = r; \boldsymbol{\lambda})$ and $\gamma_{x|y} = \Pr(X = x \mid Y = y; \boldsymbol{\lambda})$, for all x, y and r. Then, using this parameterisation, the log-likelihood function can be re-written as

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{z}_{\text{obs}}) = \sum_{x,y} z_{xy1} \log(\alpha_1 \beta_{y|1} \gamma_{x|y}) + \sum_{x} z_{x+2} \log(\sum_{y} \alpha_2 \beta_{y|2} \gamma_{x|y}), \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^{\mathrm{T}}$, $\boldsymbol{\beta}_r = (\beta_{1|r}, \ldots, \beta_{q|r})^{\mathrm{T}}$, $\boldsymbol{\gamma}_y = (\gamma_{1|y}, \ldots, \gamma_{s|y})^{\mathrm{T}}$, $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})$, and $\boldsymbol{\gamma}^{\mathrm{T}} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_q^{\mathrm{T}})$. This function can be split into two additive components, one depending on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_1$, and the other depending on $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$, which can be maximised separately. Maximising the first component gives $\hat{\alpha}_r = z_{++r} / n$ and $\hat{\beta}_{y|1} = z_{+y1} / z_{++1}$. However, the ML estimate from the second component does not always

have a simple analytical expression, although it can be interpreted geometrically. The perfect-fit log-likelihood is obtained by substituting $\hat{\pi}_{xy1} = z_{xy1} / n$ and $\hat{\pi}_{x+2} = z_{x+2} / n$ into (4). The deviance is twice the difference between the perfect-fit log-likelihood and (6), and simplifies to

$$\mathrm{dev}(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\gamma}}) = 2\sum_y z_{+y1} D(\mathbf{p}_{y1}, \hat{\boldsymbol{\gamma}}_y) + 2z_{++2} D(\mathbf{p}_2, \hat{\boldsymbol{\mu}}), \tag{7}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^{\mathrm{T}} = \sum_y \boldsymbol{\gamma}_y \beta_{y|2}$, $\mu_x = \sum_y \gamma_{x|y} \beta_{y|2} = \Pr(X = x \mid R = 2; \boldsymbol{\lambda})$, $\mathbf{p}_{y1} = (p_{1|y1}, \ldots, p_{s|y1})^{\mathrm{T}}$, $\mathbf{p}_2 = (p_{1|2}, \ldots, p_{s|2})^{\mathrm{T}}$, $p_{x|y1} = z_{xy1}/z_{+y1}$, and $p_{x|2} = z_{x+2}/z_{++2}$. The function $D(\mathbf{a}, \mathbf{b}) = \sum_i a_i \log(a_i / b_i)$ is the Kullback-Leibler distance between $\mathbf{a}, \mathbf{b} \in S$, where S is the s-dimensional simplex, the space of discrete probability distributions with s mass points. It follows from (7) that the ML estimate is the point $(\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})$ such that the weighted distance between a) $\hat{\boldsymbol{\mu}}$ and $\mathbf{p}_2$ and b) each of $\hat{\boldsymbol{\gamma}}_y$ and $\mathbf{p}_{y1}$, is minimised. This permits a geometrical interpretation of maximum likelihood estimation, which has been considered graphically by Smith et al. (1999) and Clarke (2002).

4 Results on Maximum Likelihood Esitmation

To investigate the consequences of the ML estimator's non-standard behaviour, it is first necessary to impose the condition that $\mathbf{z}_{obs} > 0$ (where $\mathbf{a} > b$ denotes the condition that every element of vector $\mathbf{a}$ is greater than scalar b). Thus the behaviour of non-response boundary solutions can be considered in isolation from those due to sparseness in $\mathbf{z}_{obs}$. Now consider the following lemma.

Lemma 1: Suppose complete data $\mathbf{Z}$ follows a multinomial sampling model with the cell probabilities determined by model (3), and that $\mathbf{z}_{\mathrm{obs}} > 0$. Then $\hat{\pi}_{xy1} > 0$ and $\hat{\pi}_{x+2} > 0$ for all x, y and all $\hat{\boldsymbol{\lambda}} \in \Lambda \cup \mathrm{bd}(\Lambda)$.

Proof: Note that, for a given choice of $\{\hat{\boldsymbol{\gamma}}_y\}$, the space of possible $\hat{\boldsymbol{\mu}}$ is the convex hull of $\{\hat{\boldsymbol{\gamma}}_y\}$, denoted by $C(\hat{\boldsymbol{\gamma}}_y) \subseteq S$; the co-ordinates of $\hat{\boldsymbol{\mu}}$ in $C(\hat{\boldsymbol{\gamma}}_y)$ - taken with respect to spanning set $\{\hat{\boldsymbol{\gamma}}_y\}$ - determine $\{\hat{\beta}_{y|2}\}$. Similarly, $\{\mathbf{p}_{y1}\}$ is the spanning set for convex hull $C(\mathbf{p}_{y1}) \subseteq S$. Hence, the ML estimate can be visualised as the choice resulting in the vertices of $C(\hat{\boldsymbol{\gamma}}_y)$ and $C(\mathbf{p}_{y1})$ lying close to each other, such that the distance between $\hat{\boldsymbol{\mu}} \in C(\hat{\boldsymbol{\gamma}}_y)$ and $\mathbf{p}_2$ is minimised. Smith et al. (1999) showed that there are two generic scenarios: (i) $\mathbf{p}_2 \in C(\mathbf{p}_{y1})$ which implies $C(\hat{\boldsymbol{\gamma}}_y) = C(\mathbf{p}_{y1})$ and $\hat{\boldsymbol{\mu}} = \mathbf{p}_2$

- a perfect-fit solution; and (ii) $\mathbf{p}_2 \notin C(\mathbf{p}_{y1})$ which implies that $C(\hat{\boldsymbol{\gamma}}_y) \neq C(\mathbf{p}_{y1})$ such that (7) is minimised.

In the perfect-fit case (i), then $\hat{\boldsymbol{\gamma}}_y = \mathbf{p}_{y1}$ and $\hat{\boldsymbol{\mu}} = \mathbf{p}_2$ is the unique solution lying in the interior of S ($\mathbf{z}_{obs} > 0$ ensures that none of $\{\mathbf{p}_{y1}\}$ or $\mathbf{p}_2$ lie on the boundary). Therefore, as $\hat{\alpha}_r = z_{++r} / n > 0$ and $\hat{\beta}_{y|1} = z_{+y1} / z_{++1} > 0$, it necessarily follows from (5) that $\hat{\pi}_{xy1} > 0$ and $\hat{\pi}_{x+2} = \hat{\alpha}_2 \hat{\mu}_x > 0$.

For case (ii), it is only required to show that a choice of $(\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})$ can always be made such that $|\mathrm{dev}(\hat{\beta}_2, \hat{\gamma})| < \infty$, even if this choice does not necessarily correspond to the minimum deviance possible. The conditon $\mathbf{z}_{obs} > 0$ ensures that the deviance cannot be $-\infty$; and a choice $C(\hat{\boldsymbol{\gamma}}_y)$ can always be found that corresponds to a finite deviance. Hence, whatever the position of $\mathbf{p}_2$ relative to $C(\mathbf{p}_{y1})$, no choice of $C(\hat{\boldsymbol{\gamma}}_y)$ exists that both minimises (7) and lies on the boundary of S, which implies that $\hat{\boldsymbol{\mu}} > 0$. This completes the proof.

Note: Clarke and Smith (2004) state that there are between 1 and $q - 1$ values of y such that $\hat{\pi}_{+y2} = 0$ for an arbitrary boundary solution. This can be seen from Lemma 1 because $\hat{\pi}_{x+2} > 0$ implies $\hat{\pi}_{xy2} > 0$ for at least one value of y, which would not be possible if all q values of y were such that $\hat{\pi}_{+y2} = 0$.

4.1 Existence and uniqueness

If follows from Lemma 1 and (5) that the ML estimate of 'natural' parameter $\boldsymbol{\omega} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\mu}^{\mathrm{T}})^{\mathrm{T}}$ will be strictly positive provided $\mathbf{z}_{obs} > 0$. This gives us the following result regarding existence.

Result 1: Suppose that $\{\mathbf{p}_{y1}\}$ and $\mathbf{p}_2$ are distinct points in S and that conditions in Lemma 1 hold. Then $\hat{\boldsymbol{\lambda}}$ will always exist and be unique.

Proof: If $\mathbf{z}_{obs} > 0$ then $\hat{\boldsymbol{\omega}} > 0$ (Lemma 1). Thus a minimum must exist in the interior of the natural parameter space because (7) is continuous, defined almost everywhere, and bounded below by zero (the perfect-fit solution). Furthermore, given $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}$, it is always possible to determine $\hat{\boldsymbol{\beta}}_2$ as the co-ordinate of $\hat{\boldsymbol{\mu}} \in C(\hat{\boldsymbol{\gamma}}_y)$, provided that $q \geq s$ (which is taken to hold throughout) and $\{\mathbf{p}_{y1}\}$ and $\mathbf{p}_2$ are distinct points in S. Hence, $\hat{\boldsymbol{\lambda}}$ can be determined for each $\hat{\boldsymbol{\omega}}$: $\hat{\alpha}_r = z_{++r} / n$, $\hat{\beta}_{y|1} = z_{+y1} / z_{++1}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}$ determine $\hat{\boldsymbol{\pi}}$ through equation (5), and thus $\hat{\boldsymbol{\lambda}}$ can be determined. Uniqueness of $\hat{\boldsymbol{\omega}}$ follows from

noting that (7) is convex because it is a weighted sum of convex $D(\mathbf{a}, \mathbf{b})$ (for fixed $\mathbf{a}$); uniqueness of $\hat{\boldsymbol{\lambda}}$ follows from noting that the mapping from $\log(\hat{\boldsymbol{\pi}})$ to is $\hat{\boldsymbol{\lambda}}$ one-to-one.

4.2 Consistency

Result 2: Suppose complete data $\mathbf{Z}$ follow a multinomial sampling model with the cell probabilities determined by (3), and that the conditions in Result 1 hold. Then $\hat{\boldsymbol{\lambda}}$ is consistent for all $\boldsymbol{\lambda} \in \Lambda$.

Proof: Consider the random variables $P_r = Z_{++r} / n$ and $P_{y|1} = Z_{+y1} / Z_{++1}$. From Section 3, $\hat{\alpha}_r = P_r$ and $\hat{\beta}_{y|1} = P_{y|1}$, with $\mathbf{z}_{obs} > 0$ ensuring that $\hat{\alpha}_r$, $\hat{\beta}_{y|1} \in (0, 1)$. The weak law of large numbers (WLLN) gives that $\hat{\alpha}_r = Z_{++r} / n \to \pi_{++r}$ and $Z_{+y1} / n \to \pi_{+y1}$ as $n \to \infty$, which implies that $\hat{\alpha}_r \to \alpha_r$ and (via Slutsky's theorem) $\hat{\beta}_{y|1} = P_{y|1} = (Z_{+y1} / n) / (Z_{++1} / n) \to \pi_{+y1} / \pi_{++1} = \beta_{y|1}$ in probability.

Now consider the sequence of functions $\{\text{dev}(\boldsymbol{\beta}_2, \boldsymbol{\gamma} | n, \mathbf{z}_{obs}) : n = 1,2,\ldots\}$ and the corresponding sequence of ML estimates obtained by maximising each deviance function by $\{\hat{\boldsymbol{\beta}}_2^n, \hat{\boldsymbol{\gamma}}^n : n = 1,2,\ldots\}$. Result 1 ensures existence and uniqueness of these estimates. Using the same arguments as above, it follows that $P_{x|y1} \to \chi_{x|y}$ and $P_{x|2} \to \mu_x$ and so $\lim_{n\to\infty}\{\text{dev}(\boldsymbol{\beta}_2, \boldsymbol{\gamma} | n, \mathbf{z}_{obs})\} = \sum_y z_{+y1} D(\boldsymbol{\gamma}_y, \boldsymbol{\gamma}_y) + z_{++2} D(\boldsymbol{\mu}, \boldsymbol{\mu})$. Hence it follows that $\lim_{n\to\infty} \hat{\boldsymbol{\beta}}_2^n = \boldsymbol{\beta}_2$ and $\lim_{n\to\infty} \hat{\boldsymbol{\gamma}}^n = \boldsymbol{\gamma}$, which implies $\hat{\boldsymbol{\beta}}_{y|2} \to \boldsymbol{\beta}_{y|2}$ and $\hat{\boldsymbol{\gamma}}_y \to \hat{\boldsymbol{\gamma}}_y$ in probability.

Together with Result 1, these results ensure that $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}$ as $\hat{\boldsymbol{\omega}} \to \boldsymbol{\omega}$, which completes the proof.

4.3 Asymptotic normality

Result 3: Suppose complete data $\mathbf{Z}$ follow a multinomial sampling model with the cell probabilities determined by (3), and that the conditions in Result 1 hold. Then $n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{d} \text{N}(\mathbf{0}, \Gamma^{-1})$, where Γ is the $p \times p$ expected single-observation information matrix.

Proof: Asymptotic normality follows from noting that, for large samples, $\hat{\boldsymbol{\lambda}}$ will lie in a neighbourhood of $\boldsymbol{\lambda}$ because the estimator is consistent (Result 2). Furthermore, as (4) is a smooth function of $\boldsymbol{\lambda}$ with finite first, second and third derivatives, and taking expectations with respect to $\mathbf{Z}_{obs}$ involves summation over a finite range of discrete values, the conditions given by Cramér (1946, Sec. 33.3) are satisfied. Therefore, the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ will be normal with a zero mean

and a $p \times p$ covariance matrix equal to the inverse of the expected single-observation information matrix.

5 Interval Estimation

It is known from Results 2 and 3 that the ML estimator is consistent and

$$n^{1/2}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Gamma^{-1}), \tag{8}$$

where $\mathbf{0}$ is a vector of p zeros and Γ is the $p \times p$ single-observation information matrix for (4). A general method for calculating Γ^{-1} has been proposed by Oakes (1999) and the performance of these intervals has been assessed by Clarke and Smith (2004), where it was noted that Γ is singular for boundary solutions. In this section, the approach of Oakes (1999) is implemented for model (3), and it is demonstrated why Γ is singular for boundary solutions. Details of a simple transformation to remove singularities from Γ` are then given.

5.1 Calculating the expected observed data information matrix.

The score equations for model (3) are obtained by differentiating (4) with respect to the free parameters and setting the resultant system of p equations to zero to give:

$$n\hat{\pi}_{x++} - z_{x++} = 0 \qquad \text{for } x = 2,\ldots, s, \tag{9a}$$

$$n\hat{\pi}_{++2} - z_{++2} = 0 \tag{9b}$$

$$n\hat{\pi}_{+y+} - z_{+y1} - \sum_x z_{x+2}\hat{\phi}_{y|x} = 0 \qquad y = 2,\ldots, q, \tag{9c}$$

$$n\hat{\pi}_{xy+} - z_{xy1} - z_{x+2}\hat{\phi}_{y|x} = 0 \qquad x = 2,\ldots, s, y = 2,\ldots, q, \tag{9d}$$

$$n\hat{\pi}_{+y2} - \sum_x z_{x+2}\hat{\phi}_{y|x} = 0, \qquad y = 2,\ldots, q, \tag{9e}$$

where $\phi_{y|x} = \Pr(Y = y|X = x, R = 2)$

$$= \exp\{\lambda^{Y}(y) + \lambda^{XY}(x, y) + \lambda^{YR}(y,2)\} \,/\, \textstyle\sum_y \exp\{\lambda^{Y}(y) + \lambda^{XY}(x, y) + \lambda^{YR}(y,2)\}, \tag{10}$$

for $y = 1,\ldots, q$.

Now consider the complete data information matrix, namely, that for model (3) in the hypothetical situation where the complete data table $\mathbf{Z} = \mathbf{z}$ is known. Denote this symmetric matrix by

J, the definition of which is given by Fienberg (1980, p. 170). By differentiating (4) twice with respect to $\boldsymbol{\lambda}$ and taking expectations with respect to $\mathbf{Z}_{obs}$, it can be shown that

$$\Gamma = \begin{pmatrix} J_{\sigma\sigma} & J_{\sigma\eta} \\ J_{\eta\sigma} & J_{\eta\eta} - \Psi \end{pmatrix}, \tag{11}$$

where Ψ is a singular symmetric matrix, $J_{\sigma\sigma}$, $J_{\sigma\eta}$, $J_{\eta\sigma}$ and $J_{\eta\eta}$ are sub-matrices of J corresponding to the partition $\boldsymbol{\lambda}^{\mathrm{T}} = (\boldsymbol{\sigma}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})$; $\boldsymbol{\sigma} = \{\lambda^X(2),\ldots, \lambda^X(q), \lambda^R(2)\}^{\mathrm{T}}$ and $\boldsymbol{\eta} = \{\lambda^Y(2),\ldots, \lambda^Y(q), \lambda^{XY}(2,2),\ldots, \lambda^{XY}(s,q), \lambda^{YR}(2,2),\ldots, \lambda^{YR}(q,2)\}^{\mathrm{T}}$. Similarly, the sub-matrices of the inverse of J are

$$J^{-1} = \begin{pmatrix} V_{\sigma\sigma} & V_{\sigma\eta} \\ V_{\eta\sigma} & V_{\eta\eta} \end{pmatrix}. \tag{12}$$

Let ψ_{ij} be element (i, j) of Ψ, corresponding to the second derivative of (4) taken with respect to i, j, which are both elements of $\boldsymbol{\eta}$, where $\boldsymbol{\lambda}^{\mathrm{T}} = (\boldsymbol{\sigma}^{\mathrm{T}}, \boldsymbol{\eta}^{\mathrm{T}})$ is the partition used in (11) and (12). The definition of $\{\psi_{ij}\}$ is given in Table 1.

Table 1: The elements of Ψ for equation (11).

i	j	$\psi_{ij} = \psi_{ji}$
$\lambda^Y(y), \lambda^{YR}(y,2)$	$\lambda^Y(y), \lambda^{YR}(y,2)$	$\sum_x \pi_{x+2}\phi_{y\|x}(1-\phi_{y\|x})$
$\lambda^Y(y), \lambda^{YR}(y,2)$	$\lambda^Y(y'), \lambda^{YR}(y',2)$	$-\sum_x \pi_{x+2}\phi_{y\|x}\phi_{y'\|x}$
$\lambda^{XY}(x,y)$	$\lambda^Y(y), \lambda^{YR}(y,2), \lambda^{XY}(x,y)$	$\pi_{x+2}\phi_{y\|x}(1-\phi_{y\|x})$
$\lambda^{XY}(x,y)$	$\lambda^Y(y'), \lambda^{YR}(y',2), \lambda^{XY}(x,y')$	$-\pi_{x+2}\phi_{y\|x}\phi_{y'\|x}$
$\lambda^{XY}(x,y)$	$\lambda^{XY}(x',y), \lambda^{XY}(x',y')$	0

Note: Indices are defined for $x,\ x' = 1,\ldots, s$; $y,\ y' = 1,\ldots, q$; $x \neq x'$; and $y \neq y'$; and $\phi_{y|x}$ is defined in (12).

By differentiating the observed data score function with respect to $\boldsymbol{\lambda}$ and taking expectations with respect to $\mathbf{Z}_{obs}$ gives the observed data information matrix. It is possible to express Γ^{-1} in terms of J and Ψ as

$$\Gamma^{-1} = J^{-1} + \begin{pmatrix} V_{\sigma\eta} \\ V_{\eta\eta} \end{pmatrix} \Psi L \begin{pmatrix} V_{\sigma\eta} \\ V_{\eta\eta} \end{pmatrix}^{\mathrm{T}}, \tag{13}$$

where $L = (I - V_{\eta\eta}\Psi)^{-1}$ and I is an identity matrix of the appropriate dimensions. The second term of the right-hand side can thus be interpreted as the additional variability resulting from the data being incomplete. The derivation of (13) follows from noting that (11) is of the form $A - B$ and $(A - B)^{-1} = A^{-1} + A^{-1}B(I - A^{-1}B)^{-1}A^{-1}$ (Mardia, et al. 1979, A 2.4f). Substituting gives

$$\begin{aligned} \Gamma^{-1} &= \begin{pmatrix} V_{\sigma\sigma} & V_{\sigma\eta} \\ V_{\eta\sigma} & V_{\eta\eta} \end{pmatrix} + \begin{pmatrix} 0 & V_{\sigma\eta}\Psi \\ 0 & V_{\eta\eta}\Psi \end{pmatrix} \begin{pmatrix} I & -V_{\sigma\eta}\Psi \\ 0 & (I - V_{\eta\eta}\Psi) \end{pmatrix}^{-1} \begin{pmatrix} V_{\sigma\sigma} & V_{\sigma\eta} \\ V_{\eta\sigma} & V_{\eta\eta} \end{pmatrix} \\ &= J^{-1} + \begin{pmatrix} V_{\sigma\eta} \\ V_{\eta\eta} \end{pmatrix} \Psi (I - V_{\eta\eta}\Psi)^{-1} \begin{pmatrix} V_{\sigma\eta} \\ V_{\eta\eta} \end{pmatrix}^{\mathrm{T}} \end{aligned}$$

which simplifies to (13). Calculating (13) requires evaluating $\hat{J}^{-1}$, $\hat{\Psi}$ and $\hat{L}$. If ML estimation is performed using the EM algorithm, as described by Baker and Laird (1988), the complete data covariance matrix will usually be available from the standard software used to implement the M step.

5.2 Calculating the information matrix for boundary solutions

Calculation of $\hat{\Gamma}^{-1}$ when $\hat{\boldsymbol{\lambda}} \in \Lambda$ is reasonably straightforward if based on (13). For boundary solutions, however, there are some complications because the information matrix becomes singular. This section considers calculation of the information matrix when a boundary solution occurs. An arbitrary b-point boundary solution is considered, where without loss of generality $\hat{\pi}_{+(q-b+1)2} = \cdots = \hat{\pi}_{+q2} = 0$ ($b > 0$), the infinite components of which can be represented by $B(\hat{\boldsymbol{\lambda}})$= $\{\lambda^{YR}(q - b + 1,2),\ldots, \lambda^{YR}(q, 2)\}$.

Now consider a column in Γ 'corresponding' to $B(\hat{\boldsymbol{\lambda}})$, that is, a column whose elements were obtained by differentiating (4) with respect to at least one element in $B(\hat{\boldsymbol{\lambda}})$. This column is equal to the difference between the equivalent column in J and that in Ψ. The columns of J corresponding to $B(\hat{\boldsymbol{\lambda}})$ are of the form

$$\begin{pmatrix} j_{\lambda^{YR}(j,2),\lambda^{X}(x)} \\ j_{\lambda^{YR}(j,2),\lambda^{R}(2)} \\ j_{\lambda^{YR}(j,2),\lambda^{Y}(j)} \\ j_{\lambda^{YR}(j,2),\lambda^{Y}(y)} \\ j_{\lambda^{YR}(j,2),\lambda_{XY}(x,j)} \\ j_{\lambda^{YR}(j,2),\lambda^{XY}(x,y)} \\ j_{\lambda^{YR}(j,2),\lambda^{YR}(j,2)} \\ j_{\lambda^{YR}(j,2),\lambda^{YR}(y,2)} \end{pmatrix} = \begin{pmatrix} \pi_{xj2} - \pi_{x++}\pi_{+j2} \\ \pi_{+j2}(1-\pi_{++2}) \\ \pi_{+j2}(1-\pi_{+j+}) \\ -\pi_{+j2}\pi_{+y+} \\ \pi_{xj2} - \pi_{+j2}\pi_{xj+} \\ -\pi_{+j2}\pi_{xy+} \\ \pi_{+j2}(1-\pi_{+j2}) \\ -\pi_{+j2}\pi_{+y2} \end{pmatrix} \text{ for } j = (q-b+1),\ldots, q,$$

where $x = 2,\ldots, s$, $y = 2, \ldots, q$ and $y \neq j$. Clearly, each element of these columns will be zero because $\hat{\pi}_{xj2} = \hat{\pi}_{+j2} = 0$, for all $\lambda^{YR}(j, 2) \in B(\hat{\boldsymbol{\lambda}})$. From Table 1, the equivalent elements of Ψ are also zero because $\hat{\phi}_{j|x} = 0$, for all x, j. Therefore $\hat{\Gamma}$ is singular because b columns (and b rows due to symmetry) have every element equal to zero.

A solution to this problem is to transform the infinite parameters to a new parameter space so that the singularity disappears. Partition $\boldsymbol{\lambda}^{\mathrm{T}} = (\boldsymbol{\omega}^{\mathrm{T}}, \boldsymbol{\lambda}^{YR})$, where $\boldsymbol{\lambda}^{YR} = \{\lambda^{YR}(2, 2), \ldots, \lambda^{YR}(q, 2)\}$ is a row vector and $\boldsymbol{\omega}$ is a column vector containing the remaining $p - q + 1$ parameters. Now consider transforming $\boldsymbol{\lambda}^{\mathrm{T}} = (\boldsymbol{\omega}^{\mathrm{T}}, \boldsymbol{\lambda}^{YR})$ to $(\boldsymbol{\omega}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})$, where

$$\boldsymbol{\theta}^{\mathrm{T}} = \exp(\boldsymbol{\lambda}^{YR}) = [\exp\{\lambda^{YR}(2,2)\},\ldots,\exp\{\lambda^{YR}(q,2)\}] = (\theta_2,\ldots,\theta_q). \tag{14}$$

If $\lambda^{YR}(y, 2) \in B(\hat{\boldsymbol{\lambda}})$ then $\hat{\lambda}^{YR}(y,2) = -\infty$ and so $\hat{\theta}_y = 0$, which is finite. Using the delta method, the information matrix for this new parameterisation can be written as

$$\tau(\boldsymbol{\theta})\, \Gamma(\boldsymbol{\omega}, \boldsymbol{\lambda}^{YR})\, \tau(\boldsymbol{\theta}), \tag{15}$$

where $\tau(\boldsymbol{\theta}) = \mathrm{diag}\{\mathbf{1}^{\mathrm{T}}, (1/\boldsymbol{\theta})^{\mathrm{T}}\}$, $\mathbf{1}$ is a unitary vector of length $p - q + 1$, and $1/\boldsymbol{\theta} = (1/\theta_2,\ldots, 1/\theta_q)^{\mathrm{T}}$. The elements of (15) that involved differentiating (4) with respect to elements of $\boldsymbol{\theta}$ can be partitioned into those differentiated by: (i) one element in $\boldsymbol{\theta}$ and one in $\boldsymbol{\omega}$; (ii) two distinct elements in $\boldsymbol{\theta}$ (empty if $\boldsymbol{\theta}$ is a scalar); and (iii) an element in $\boldsymbol{\theta}$ twice. For the first group, if differentiation was taken with respect to $\theta_j \in \boldsymbol{\theta}$ then the element in (15) is equal to the corresponding element in Γ divided by θ_j. If $\hat{\theta}_j = 0$ then the group (i) column is non-zero. To see why, consider the example where the element is associated with $\lambda^{R}(2)$ and θ_j. Dividing by θ_j gives

$$\frac{\pi_{+j2}(1-\pi_{++2})}{\theta_j} = (1-\pi_{++2})\sum_{x}\exp\{\nu + \lambda^{X}(x) + \lambda^{Y}(j) + \lambda^{R}(2) + \lambda^{XY}(x,j)\}, \tag{16}$$

which is positive when $\hat{\theta}_j = 0$. Applying this principle to the group (ii) columns, the element in Γ is divided by the product of the two elements in $\boldsymbol{\theta}$, with the same effect that the element in (15) is finite, because the elements in Ψ are also nonzero when divided by θ_j. For group (iii) the situation is a little more complicated but the same result holds. By noting that $\pi_{xy2} = \pi_{x+2}\phi_{j|x}$, the elements of Γ in group (iii) can be written as

$$\frac{1}{\theta_j^2}\left(\pi_{+j2}(1-\pi_{+j2}) - \sum_x \pi_{x+2}\phi_{j|x}(1-\phi_{j|x})\right) = \sum_x \pi_{x+2}\left(\frac{\phi_{j|x}}{\theta_j}\right)^2 - \left(\sum_x \pi_{x+2}\frac{\phi_{j|x}}{\theta_j}\right)^2, \qquad (17)$$

and substituting

$$\frac{\phi_{j|x}}{\theta_j} = \frac{\exp\{\lambda^Y(j) + \lambda^{XY}(x,j)\}}{\sum_{j=1}^{q}\exp\{\lambda^Y(j) + \lambda^{YR}(j,2) + \lambda^{XY}(x,j)\}}$$

into the right hand side of (17). It follows that (17) is positive when $\hat{\theta}_j = 0$. Hence intervals for $\boldsymbol{\theta}$ can now be calculated.

6 Discussion

In this paper, it has been shown that although ML estimation of simple non-ignorable Baker and Laird models can lead to infinite estimates of certain log-linear parameters, these estimates exist and are unique under weak conditions; namely, that the standard condition that $q \geq s$ and the $\{\mathbf{p}_{y1}\}$ and $\mathbf{p}_2$ are distinct points in the s-dimensional simplex S. The reason for this is that ML estimation can be visualised as determining the values of the natural parameters $\boldsymbol{\omega}$ by minimising the deviance (7); a unique minimum exists in the interior of the natural parameter space, which can be shown to map uniquely to the space of log-linear parameters because the mapping between the expected cell frequencies and log-linear parameters is convex. Consistency is demonstrated by showing that the ML estimator of the natural parameters tends to the true value, which again maps to a unique point in the log-linear parameter space; asymptotic normality follows straightforwardly.

It is interesting to compare these findings with those on extended ML estimation of log-linear models for completely classified but sparse contingency tables. In contrast to Result 1, Wedderburn (1976) showed that existence and uniqueness of extended ML estimates for log-linear models in sparse tables can only be guaranteed if the row space of the design matrix is equal to that of the design matrix with the rows corresponding to zero-frequency cells removed. The ML estimate exists in the interior

of the natural parameter space, bounding the likelihood above and ensuring existence and uniqueness. Thus, the results outlined in this paper also mean that complicated algorithms for calculation extended ML estimates (Clarkson and Jennrich 1991) are unnecessary for non-response boundary solutions.

Extending these results to the general class of non-ignorable Baker and Laird models is relatively simple. Begin by noting that every non-ignorable model corresponds to an independence graph, and call the most complex model corresponding to a particular graph the 'factor' model for that graph. Clarke (2002) showed that only factor models of the form $\mathbf{X}_1\mathbf{X}_2Y + \mathbf{X}_1YR$ are identifiable, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is a partition of the covariates and $\mathbf{X}_1$ may be an empty vector so $\mathbf{X}_2 = \mathbf{X}$; the product notation **AB** indicates the presence of the highest-order interaction involving all variables corresponding to the elements of **A** and **B**. Model (3) is included in this definition with $\mathbf{X}_1 = 0$ and $\mathbf{X}_2 = X$. The levels of $\mathbf{X}_2$ correspond to q_2 strata within which separate $\mathbf{X}_1Y + YR$ models are fitted; the results in this paper can be applied to the $\mathbf{X}_1Y + YR$ models within each stratum to demonstrate existence, etc. It follows that all hierarchical models corresponding to the same independence graph are also identified; the log-linear model acts only to impose a convex space in S within which the solution lies, at the point in the space closest to the factor model solution. Extension to causal models for non-ignorable non-response with more complex patterns of missing values (Fay 1986) would form a basis for further research.

The results in Section 5 act to provide background for the work on interval estimation by Clarke and Smith (2004), comparing normal intervals to bootstrap and profile likelihood alternatives. It was found that, although important, the findings in this paper are only part of the story. Although consistent and asymptotically normal, the ML estimator for some log-linear parameters is highly non-normal for finite sample sizes, which can lead to very poor interval estimates using the normal approximation.

References

Baker, S.G. and Laird, N.M. (1988) Regression analysis for categorical variables with outcomes subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **83**, 62-69.

Baker, S.G., Rosenberger, W.F. and DerSimonian, R. (1992) Closed-form estimates for missing counts in two-way tables. *Statistics in Medicine*, **11**, 643-657.

Clarke, P.S. (1998) Nonignorable nonresponse models for categorical survey data. Unpublished Ph.D. thesis, Department of Social Statistics, University of Southampton, U.K.

Clarke, P.S. (2002) On boundary solutions and identifiability for categorical regression with non-ignorable non-response. *Biometrical Journal*, **44**, 701-717.

Clarke, P.S. and Smith, P.W.F. (2004) Interval estimation for log-linear models with one variable subject to non-ignorable non-response. *Journal of the Royal Statistical Society* B, **66** (in press).

Clarkson, D.B. and Jennrich, R.I. (1991) Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society* B, **53**, 417-426.

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Fay, R.E. (1986) Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, **81**, 354-365.

Fienberg, S.E. (1980) *The Analysis of Cross-classified Categorical Data*. Cambridge, MA: MIT Press.

Forster, J.J. and Smith, P.W.F. (1998) Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Journal of the Royal Statistical Society* B, **60**, 57-70.

Haberman, S.J. (1974) *The Analysis of Frequency Data.* Chicago: University of Chicago Press.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*. London: Academic Press.

Oakes, D. (1999) Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society* B, **61**, 479-482.

Park, T.S. (1998) An approach to categorical data with nonignorable nonresponse. *Biometrics*, **54**, 1579-1590.

Park, T.S. and Brown, M.B. (1994) Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, **89**, 44-52.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.

Smith, P.W.F., Skinner, C.J. and Clarke, P.S. (1999) Allowing for non-ignorable non-response in the analysis of voting intention data. *Applied Statistics*, **48**, 563-577.

Wedderburn, R.W.M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27-32.