Multinomial logit bias reduction via the Poisson log-linear model

Ioannis Kosmidis
Department of Statistical Science, University College
London, WC1E 6BT, UK
ioannis@stats.ucl.ac.uk

and

David Firth
Department of Statistics, University of Warwick
Coventry CV4 7AL, UK
d.firth@warwick.ac.uk

June 10, 2011

Abstract

For the parameters of a multinomial logistic regression it is shown how to obtain the bias-reducing penalized maximum likelihood estimator by using the equivalent Poisson log-linear model. The calculation needed is not simply an application of the Jeffreys- prior penalty to the Poisson model. The development allows a simple and computationally efficient implementation of the reduced-bias estimator, using standard software for generalized linear models.
Keywords: Jeffreys prior; leverage; logistic linear regression; Poisson trick.

1 Introduction

Use of the Jeffreys-prior penalty to remove the $O(n^{-1})$ asymptotic bias of the maximum likelihood estimator in full exponential family models was developed in Firth (1993) and has been found to be particularly effective in binomial and multinomial logistic regressions (e.g., Heinze & Schemper, 2002; Bull et al., 2002, 2007). Implementation of the method in binomial and other univariate-response models is by means of a simple, iterative data-adjustment scheme (Firth, 1992). In this paper we extend such simplicity of implementation to multinomial models.

In what follows, the Kronecker function δ_{sk} is equal to 1 when $s = k$ and zero otherwise. Suppose that observed k-vectors $y_1, \ldots, y_n$ of counts are realizations of independent multinomial random vectors $Y_1, \ldots, Y_n$. Let $m_r = \sum_{s=1}^{k} y_{rs}$ be the multinomial total and let π_{rs} be the probability of the sth category for the multinomial vector Y_r, with $\sum_{s=1}^{k} \pi_{rs} = 1$ $(r = 1, \ldots, n; s = 1, \ldots, k)$. In multinomial logistic regression the log-odds of category s versus category k, say, for the rth multinomial vector is

$$\log\left(\frac{\pi_{rs}}{\pi_{rk}}\right) = x_r^T \beta_s \quad (r = 1, \ldots, n; s = 1, \ldots, k-1)\,. \tag{1}$$

Here x_r is a vector of p covariate values, with first component unity if a constant is included in the model; and $\beta_s \in \Re^p$ is a vector of parameters for the sth category $(s = 1, \ldots, k-1)$.

The multinomial model (1) can be embedded conveniently into a Poisson log-linear model. If Y_{rs} $(r = 1, \ldots, n;\ s = 1, \ldots, k)$ are independently Poisson with means

$$\mu_{rs} = \exp\{\phi_r + (1 - \delta_{sk})x_r^T \beta_s\}, \tag{2}$$

then the Poisson likelihood factorizes: with M_r denoting $\sum_{s=1}^{k} Y_{rs}$, the conditional distribution of Y_r given M_r is the multinomial model of interest, while the totals M_r are Poisson-distributed with means $\tau_r = \sum_{s=1}^{k} \mu_{rs}$ $(r = 1, \ldots, n)$. Maximum likelihood inferences for $\beta = (\beta_1^T, \ldots, \beta_{k-1}^T)^T$ obtained from the full, unconditional Poisson likelihood are thus identical to those based directly on the multinomial likelihood. This equivalence was noted in Birch (1963), and Palmgren (1981) showed that the inverse of the expected information on $\beta_1, \ldots, \beta_{k-1}$ is the same in both representations under the restriction $\tau_r = m_r$ $(r = 1, \ldots, n)$ on the parameter space of the Poisson log-linear model. That restriction is automatically satisfied at the maximum likelihood estimate because if $l(\beta, \phi_1, \ldots, \phi_n)$ is the log-likelihood for the model (2) then $\partial l / \partial \phi_r = m_r - \tau_r$.

The multinomial logit model (1) and the Poisson log-linear model (2) are both full exponential families, and so in either case the bias-reducing penalty of Firth (1993) to the likelihood is simply the Jeffreys (1946) invariant prior for the model. However, in the (β, ϕ) parameterization, the penalized Poisson likelihood cannot in general be factorized as the product of the required penalized multinomial likelihood and a factor free of β. As a result, naive computation of reduced-bias estimates for the full parameter vector (β, ϕ) in the Poisson log-linear model does not deliver reduced-bias estimates for the parameters β of the multinomial model, as might be hoped.

The solution is to work with a restricted version of the Poisson model, in which the constraints $\tau_r = m_r$ $(r = 1, \ldots, n)$ are explicitly imposed. This Poisson model is then a generalized nonlinear model. This might at first sight appear to complicate what is intended to be a simplifying computational device; however, the general results of Kosmidis & Firth (2009) apply, and yield a useful representation of the adjusted score vector which in turn suggests a simple iterative algorithm.

2 Bias reduction via the Poisson model

2.1 Reduction of the bias for ϕ and β under Poisson sampling

The incorrect, naive approach, which simply applies the Jeffreys prior to the Poisson-log-linear model (2), is briefly reviewed here. This establishes notation, and will be useful for the iteration developed in § 3.

Let $q = k - 1$. In Firth (1992) it is shown that the bias-reducing adjusted score functions for the model (2) can be written in the form

$$U_t^* = \sum_{r=1}^{n} \sum_{s=1}^{k} \left(y_{rs} + \frac{1}{2} h_{rss} - \mu_{rs} \right) z_{rst} \quad (t = 1, \ldots, n + pq)\,. \tag{3}$$

Here z_{rst} is the (s, t)th component of the $k \times (n + pq)$ matrix

$$Z_r = \left[\begin{array}{c|c} G_r & 1_q \otimes e_r^T \\ \hline 0_{pq}^T & e_r^T \end{array} \right] \quad (r = 1, \ldots, n)\,,$$

where $G_r = I_q \otimes x_r^T$ $(r = 1, \ldots, n)$, I_q is the $q \times q$ identity matrix, 0_{pq} is a pq-vector of zeros, 1_q is a q-vector of ones, and e_r is a n-vector of zeros with one as its rth element. The quantity h_{rss}

is the sth diagonal element of the $k \times k$ matrix $H_r = Z_r F^{-1} Z_r^T W_r$, where F is the expected information for θ and $W_r = \text{diag}\{\mu_{r1}, \ldots, \mu_{rk}\}$ $(r = 1, \ldots, n)$. The matrix H_r is the $k \times k$, rth diagonal block of the asymmetric hat matrix for the Poisson log-linear model. Expression (3) directly suggests an iterative procedure for solving the adjusted score equations: at the jth iteration, (i) calculate $h_{rss}^{(j)}$ $(r = 1, \ldots, n; s = 1, \ldots, k)$, where the superscript (j) denotes evaluation at the candidate estimate $\theta^{(j)}$ of the previous iteration, and then (ii) fit model (2) by maximum likelihood but using adjusted responses $y_{rs} + h_{rss}^{(j)}/2$ in place of y_{rs}, to get new estimates $\theta^{(j+1)}$.

However, as noted in § 1, solving $U_t^* = 0$ $(r = 1, \ldots, n)$ would not result in the reduced-bias estimates of β for the multinomial model, because of the presence of the technical nuisance parameters $\phi_1, \ldots, \phi_n$. For example, from (3) the adjusted score equation for ϕ_r is $\tau_r = m_r + \text{tr}(H_r)/2$; this is in contrast to maximum likelihood, where the essential restriction $\hat{\tau}_r = m_r$ $(r = 1, \ldots, n)$ is automatic.

2.2 Adjusted score functions in the restricted parameter space

If the Poisson log-linear model (2) is parameterized in terms of $\theta^\dagger = (\beta^T, \tau^T)^T$, then the restriction $\tau_r = m_r$ $(r = 1, \ldots, n)$ can be applied directly by fixing components of the parameter vector $\theta^\dagger$. Furthermore, the parameters τ and β are orthogonal (Palmgren, 1981), which simplifies the derivations. Model (2) is then re-written as a canonically-linked generalized nonlinear model,

$$\log \mu_{rs} = \log \frac{\tau_r}{1 + \sum_{u=1}^{q} \exp(x_r^T \beta_u)} + (1 - \delta_{sk}) x_r^T \beta_s \quad (r = 1, \ldots, n; s = 1, \ldots, k)\,. \tag{4}$$

The variance and the third cumulant of Y_{rs} under the Poisson assumption are equal to μ_{rs} and the leverages h_{rss} are parameterization invariant. Hence, expression (13) in Kosmidis & Firth (2009) gives that the bias-reducing adjusted score equations using adjustments based on the expected information matrix take the form

$$U_t^\dagger = \sum_{r=1}^{n} \sum_{s=1}^{k} \left[y_{rs} + \frac{1}{2} h_{rss} + \frac{1}{2} \mu_{rs} \text{tr}\left\{ (F^\dagger)^{-1} \mathcal{D}^2\left(\zeta_{rs}; \theta^\dagger\right) \right\} - \mu_{rs} \right] z_{rst}^\dagger \quad (t = 1, \ldots, n + pq)\,,$$

where $F^\dagger$ is the expected information on $\theta^\dagger$, $\mathcal{D}^2\left(\zeta_{rs}; \theta^\dagger\right)$ denotes the $(n+pq) \times (n+pq)$ Hessian matrix of ζ_{rs} with respect to $\theta^\dagger$, and $z_{rst}^\dagger$ is the (s,t)th component of the $k \times (n+pq)$ matrix

$$Z_r^\dagger = \left[\begin{array}{c:c} G_r - 1_q \otimes \left(\pi_r^T G_r\right) & 1_q \otimes \left(\tau_r^{-1} e_r^T\right) \\ \hdashline -\pi_r^T G_r & \tau_r^{-1} e_r^T \end{array} \right] \quad (r = 1, \ldots, n)\,,$$

with $\pi_r = (\pi_{r1}, \ldots, \pi_{rq})^T$ and $\pi_{rs} = \mu_{rs}/\tau_r$ $(s = 1, \ldots, k)$.

After noting that $\mathcal{D}^2\left(\zeta_{rs}; \theta^\dagger\right)$ does not depend on s and substituting for $z_{rst}^\dagger$ $(r = 1, \ldots, n; s = 1, \ldots, k)$, the adjusted score functions for β take the simple form

$$U_t^\dagger = \sum_{r=1}^{n} \sum_{s=1}^{q} \left[y_{rs} + \frac{1}{2} h_{rss} - \left\{ m_r + \frac{1}{2} \text{tr}(H_r) \right\} \pi_{rs} \right] g_{rst} \quad (t = 1, \ldots, pq)\,, \tag{5}$$

where g_{rst} is the (s,t)th component of G_r $(r = 1, \ldots, n)$.

The only quantities in expression (5) affected by the restriction $\tau_r = m_r$ $(r = 1, \ldots, n)$ are the leverages h_{rss}. The following theorem shows the effect of the restriction on the leverages by providing some identities on the relationship between the matrix H_r and the $q \times q$, rth diagonal block of the asymmetric hat matrix for the multinomial logistic regression model (1). Denote the latter matrix by V_r.

Theorem 1 *Let v_{rsu} be the (s,u)th component of the matrix V_r $(r=1,\ldots,n; s,u=1,\ldots,q)$. If the parameter space is restricted by $\tau_1 = m_1, \ldots, \tau_n = m_n$, then*

$$h_{rss} = \pi_{rs} + v_{rss} - \sum_{u=1}^{q} \pi_{ru} v_{rus} \quad (s=1,\ldots,q)\,,$$

$$h_{rkk} = \pi_{rk} + \sum_{s,u=1}^{q} \pi_{ru} v_{rus}\,,$$

where $\pi_{rs} = \mu_{rs}/m_r$ $(r=1,\ldots,n; s=1,\ldots,k)$.

The proof of Theorem 1 is in the Appendix.

Direct use of the identities in Theorem 1 yields that, under the restriction $\tau_r = m_r$ $(r = 1,\ldots,n)$, the adjusted score functions for β in (5) take the form

$$U_t^{\dagger} = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} + \frac{1}{2}v_{rss} - \left\{m_r + \frac{1}{2}\mathrm{tr}(V_r)\right\}\pi_{rs} - \frac{1}{2}\sum_{u=1}^{q}\pi_{ru}v_{rus}\right] g_{rst} \quad (t=1,\ldots,pq)\,.$$

Application of results from Kosmidis & Firth (2009, p.797) on adjusted score functions for canonical-link multivariate generalized linear models, after some simple matrix manipulation, shows that these adjusted score functions are identical to those obtained by direct penalization of the likelihood for the multinomial model (1). Hence the required reduced-bias estimates of β are reduced-bias estimates of the nonlinear Poisson model (4) under parameter constraints $\tau_r = m_r$ $(r=1,\ldots,n)$. The algebraic manipulations, which are straightforward but tedious, are in the Appendix.

3 Reduced-bias estimates for β

Expression (5) suggests the following iterative procedure: move from candidate estimates $\beta^{(j)}$ to new values $\beta^{(j+1)}$ by solving

$$0 = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} + \frac{1}{2}\tilde{h}_{rss}^{(j)} - \left\{m_r + \frac{1}{2}\mathrm{tr}\left(\tilde{H}_r^{(j)}\right)\right\}\pi_{rs}^{(j+1)}\right] g_{rst} \quad (t=1,\ldots,pq)\,, \tag{6}$$

with $\tilde{h}_{rss}^{(j)}$ calculated for the restricted parameterization. Directly from (5), the above iteration has a stationary point at the reduced-bias estimates of β.

To implement the above iteration one can take advantage of the fact that the solution of the adjusted score functions (3) for the Poisson log-linear model (2) implies the solution of $\tau_r = m_r + \mathrm{tr}(H_r)/2$ $(r=1,\ldots,n)$. Hence, iteration (6) can be implemented as:

1. set $\tilde{\phi}_r^{(j)} = \log m_r - \log\left\{1 + \sum_{s=1}^{q}\exp\left(x_r^T\beta_s^{(j)}\right)\right\}$ $\quad (r=1,\ldots,n)$,
2. use $\tilde{\theta}^{(j)} = (\beta^{(j)}, \tilde{\phi}_1^{(j)}, \ldots, \tilde{\phi}_n^{(j)})$ to calculate new values $\tilde{H}_r^{(j)}$ $(r=1,\ldots,n)$,
3. fit model (2) by maximum likelihood but using the adjusted responses $y_{rs} + \tilde{h}_{rss}^{(j)}/2$ in place of y_{rs} to get new estimates $\phi^{(j+1)}$ and $\beta^{(j+1)}$ $(r=1,\ldots,n; s=1,\ldots,k)$.

The β-block of the inverse of the expected information matrix evaluated at the reduced-bias estimates can be used to produce valid standard errors for the estimators.

Note that H_r depends on the model parameters only through the Poisson expectations $\mu_{r1},\ldots,\mu_{rk}$ $(r=1,\ldots,n)$ and that the first step implies the rescaling of the current values of

those expectations so that they sum to the corresponding multinomial totals. It is straightforward to implement this iteration using standard software for univariate-response generalized linear models; a documented program for the R statistical computing environment (R Development Core Team, 2010) is available in the Appendix.

Acknowledgement

The comments of the Editor, Associate Editor and a referee have resulted in a clearer account of the work and are much appreciated. Part of this work was carried out when the first author was a member of the Centre for Research in Statistical Methodology, University of Warwick. Financial support from the Engineering and Physical Sciences Research Council, for the work of both authors, is gratefully acknowledged.

Appendix

Preamble

The proof and algebraic derivations that follow are taken from Appendices B.5 and B.6 of the 2007 University of Warwick PhD thesis by I. Kosmidis (Kosmidis, 2007), available online at `http://www.ucl.ac.uk/~ucakiko/ikosmidis_thesis.html`.

This appendix corrects minor typographical errors in Appendices B.5 and B.6 of the PhD thesis, and makes the notational and textual adjustments needed to match material in the paper.

Proof of Theorem 1

Write $Z_r^\dagger = \left[Q_{1,r} \vdots Q_{2,r} \right] \quad (r = 1, \ldots, n)$, where

$$Q_{1,r} = \begin{bmatrix} G_r \\ \hdashline 0_{pq}^T \end{bmatrix} - 1_k \otimes (\pi_r^T G_r) \quad \text{and} \quad Q_{2,r} = 1_k \otimes (\tau_r^{-1} e_r^T)\,. \tag{7}$$

In Palmgren (1981) it is shown that the expected information on $\theta^\dagger$ is the block diagonal matrix

$$F^\dagger = \begin{bmatrix} F_\beta^\dagger & \\ \hdashline & F_\tau^\dagger \end{bmatrix},$$

where $F_\beta^\dagger$ is the expected information on β and $F_\tau^\dagger = \mathrm{diag}(1/\tau_1, \ldots, 1/\tau_n)$ is the expected information on τ. Palmgren (1981) also showed that if the parameter space is restricted by $\tau_1 = m_1, \ldots, \tau_n = m_n$ then $F_\beta^\dagger = E$ where E is the information on β corresponding to the likelihood function of the multinomial logistic regression model. Noting that the $k \times k$ matrix H_r is parameterization invariant,

$$\begin{aligned} H_r &= Z_r^\dagger \left(F^\dagger\right)^{-1} \left(Z_r^\dagger\right)^T W_r \\ &= Q_{1,r} \left(F_\beta^\dagger\right)^{-1} Q_{1,r}^T W_r + Q_{2,r} \mathrm{diag}(\tau_1, \ldots, \tau_n) Q_{2,r}^T W_r \quad (r = 1, \ldots, n)\,, \end{aligned} \tag{8}$$

where $W_r = m_r \text{diag}(\pi_{r1}, \ldots, \pi_{rk})$. Substituting (7) in (8) and restricting the parameter space by $\tau_1 = m_1, \ldots, \tau_n = m_n$ gives

$$H_r = \left[\begin{array}{c:c} V_r\Sigma_r^{-1} & 0_q \\ \hdashline 0_q^T & 0 \end{array}\right] W_r - \left[\begin{array}{c} 1_k^T \otimes (V_r\Sigma_r^{-1}\pi_r) \\ \hdashline 0_k^T \end{array}\right] W_r - \left[\, 1_k \otimes (\pi_r^T\Sigma_r^{-1}V_r^T) \,\vdots\, 0_k \,\right] W_r$$
$$+ \pi_r^T V_r \Sigma_r^{-1} \pi_r 1_{k\times k} W_r + \left[\, (1_k \otimes \pi_r) \,\vdots\, 1_k \pi_{rk} \,\right] \qquad (9)$$

where $\pi_{rk} = 1 - \sum_{s=1}^{q} \pi_{rs}$ $(r = 1, \ldots, n)$ and $1_{k\times k}$ is a $k \times k$ matrix of ones. The matrix $V_r = G_r E^{-1} G_r^T \Sigma_r$ is the $q \times q$, rth diagonal block of the asymmetric 'hat matrix' for the multinomial logistic regression model, and Σ_r is the $q \times q$ variance-covariance matrix of the incomplete multinomial vector $(Y_{r1}, \ldots, Y_{rq})^T$, with (s, u)th component

$$\sigma_{rsu} = \begin{cases} m_r \pi_{rs}(1 - \pi_{rs}) \,, & s = u \\ -m_r \pi_{rs} \pi_{ru} \,, & s \neq u \end{cases} \qquad (s, u = 1, \ldots, q) \,.$$

The inverse Σ_r^{-1} of Σ_r has (s, u)th component

$$\rho_{rsu} = \begin{cases} m_r^{-1}(\pi_{rs}^{-1} + \pi_{rk}^{-1}) \,, & s = u \\ m_r^{-1}\pi_{rk}^{-1} \,, & s \neq u \end{cases} \qquad (s, u = 1, \ldots, q) \,.$$

Expression (9) is an expression of H_r in the restricted parameter space in terms of V_r and the multinomial probabilities $\pi_{r1}, \ldots, \pi_{rk}$. After performing the matrix multiplications and additions in (9) the diagonal elements of H_r $(r = 1, \ldots, n)$ are written as

$$h_{rss} = \pi_{rs} + v_{rss} - \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} v_{rsu} + \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} \sum_{w=1}^{q} \pi_{ru} v_{ruw} \quad (s = 1, \ldots, q) \,,$$
$$h_{rkk} = \pi_{rk} + \sum_{u=1}^{q} \sum_{w=1}^{q} \pi_{ru} v_{ruw} \,.$$

The proof is completed after showing that

$$\sum_{u=1}^{q} \pi_{ru} v_{rus} = \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} v_{rsu} - \frac{\pi_{rs}}{\pi_{rk}} \sum_{u=1}^{q} \sum_{w=1}^{q} \pi_{ru} v_{ruw} \quad (r = 1, \ldots, n; s = 1, \ldots, q) \,. \qquad (10)$$

If $G_r = 1_q \otimes x_r^T$ is substituted in $V_r = G_r E^{-1} G_r^T \Sigma_r$ and E_{su}^- denotes the (s, u)th, $p \times p$ block of E^{-1} (that is the block corresponding to β_s and β_u), then direct matrix multiplication gives

$$v_{rsu} = m_r \pi_{ru} \left(x_r^T E_{su}^- x_r - \sum_{w=1}^{q} \pi_{rw} x_r^T E_{sw}^- x_r \right) \quad (s, u = 1, \ldots, q) \,. \qquad (11)$$

Hence, because $E_{su}^- = E_{us}^-$, the left hand side of (10) is

$$\sum_{u=1}^{q} \pi_{ru} v_{rus} = m_r \pi_{rs} \sum_{u=1}^{q} \pi_{ru} x_r^T E_{su}^- x_r - m_r \pi_{rs} \sum_{u=1}^{q} \sum_{w=1}^{q} \pi_{ru} \pi_{rw} x_r^T E_{uw}^- x_r \quad (s = 1, \ldots, q) \,.$$

Substituting (11) in the right hand side of (10) gives the same result as above. □

Derivation of the adjusted scores via the multinomial likelihood

From Kosmidis & Firth (2009, p. 797) the bias-reducing adjusted score equations for the multinomial logistic regression model using adjustments based on the expected information matrix take the form

$$U_t^\dagger = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} - m_r\pi_{rs} + \frac{1}{2}\mathrm{tr}\left(V_r\Sigma_r^{-1}K_{rs}\right)\right]g_{rst} \quad (t = 1, \ldots, pq)\,, \tag{12}$$

where K_{rs} is the $q \times q$, sth block of rows of the $q^2 \times q$ matrix of third-order cumulants of the incomplete multinomial vector $(r = 1, \ldots, n; s = 1, \ldots, q)$ and has $(u, v)th$ component

$$\kappa_{rsuv} = \begin{cases} m_r\pi_{rs}(1-\pi_{rs})(1-2\pi_{rs}) & s = u = v \\ -m_r\pi_{rs}\pi_{rv}(1-\pi_{rs}) & s = u \neq v \\ 2m_r\pi_{rs}\pi_{ru}\pi_v & s, u, v \quad \text{distinct} \end{cases} \qquad (s, u, v = 1, \ldots, q)\,.$$

Direct matrix multiplication then gives,

$$\Sigma_r^{-1}K_{rs} = \begin{bmatrix} -\pi_{rs} & 0 & \ldots & 0 & \ldots & 0 \\ 0 & -\pi_{rs} & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ -\pi_{r1} & -\pi_{r2} & \ldots & 1-2\pi_{rs} & \ldots & -\pi_{rq} \\ \vdots & \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & -\pi_{rs} \end{bmatrix} \quad (s = 1, \ldots, q)\,,$$

and

$$\mathrm{tr}\left(V_r\Sigma_r^{-1}K_{rs}\right) = v_{rss} - \sum_{u=1}^{q}\pi_{ru}v_{rus} - \pi_{rs}\mathrm{tr}V_r \quad (r = 1, \ldots, n; s = 1, \ldots, q)\,.$$

Substituting the above expression into (12) gives

$$U_t^\dagger = \sum_{r=1}^{n}\sum_{s=1}^{q}\left[y_{rs} + \frac{1}{2}v_{rss} - \left\{m_r + \frac{1}{2}\mathrm{tr}(V_r)\right\}\pi_{rs} - \frac{1}{2}\sum_{u=1}^{q}\pi_{ru}v_{rus}\right]g_{rst} \quad (t = 1, \ldots, pq)\,.$$

□

References

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B: Methodological* **25**, 220–233.

Bull, S. B., Lewinger, J. B. & Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine* **26**, 903–918.

Bull, S. B., Mak, C. & Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57–74.

Firth, D. (1992). Generalized linear models and Jeffreys priors: An iterative generalized least-squares approach. In *Computational Statistics I*, Y. Dodge & J. Whittaker, eds. Heidelberg: Physica-Verlag.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Heinze, G. & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London* **186**, 453–461.

Kosmidis, I. (2007). *Bias Reduction in Exponential Family Nonlinear Models*. Ph.D. thesis, Department of Statistics, University of Warwick.

Kosmidis, I. & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793–804.

Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika* **68**, 563–566.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.