Biometrika (2011), **98**, 3, *pp.* 755–759
doi: 10.1093/biomet/asr026
Advance Access publication 13 July 2011

Multinomial logit bias reduction via the Poisson log-linear model

By IOANNIS KOSMIDIS

Department of Statistical Science, University College, London WC1E 6BT, U.K.
ioannis@stats.ucl.ac.uk

and DAVID FIRTH

Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K.
d.firth@warwick.ac.uk

Summary

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.

Some key words: 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; see Firth (1992) and a 2009 Centre for Research in Statistical Methodology working paper series of Kosmidis. 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

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

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 $\beta = (\beta_1^{\mathrm{T}}, \ldots, \beta_{k-1}^{\mathrm{T}})^{\mathrm{T}}$ 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 loglikelihood 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

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

where $G_r = I_q \otimes x_r^{\mathrm{T}}$ $(r = 1, \ldots, n)$, 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 $H_r = Z_r F^{-1} Z_r^{\mathrm{T}} W_r$, where F is the expected information for $\theta = (\beta^{\mathrm{T}}, \phi^{\mathrm{T}})^{\mathrm{T}}$ and $W_r = \mathrm{diag}\{\mu_{r1}, \ldots, \mu_{rk}\}$ $(r = 1, \ldots, n)$. The matrix H_r is the $k \times k$, rth diagonal block of the asymmetric so-called 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 $U_t^* = 0$ $(t = 1, \ldots, n + pq)$ 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 $\tau_r = m_r + \mathrm{tr}(H_r)/2$; 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 $\theta^\dagger = (\beta^{\mathrm{T}}, \tau^{\mathrm{T}})^{\mathrm{T}}$, 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,

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

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 functions using adjustments based on the expected information matrix take the form

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

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

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

with $\pi_r = (\pi_{r1}, \ldots, \pi_{rq})^{\mathrm{T}}$ 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

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

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 Supplementary Material.

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 Supplementary Material.

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:

(i) set $\tilde{\phi}_r^{(j)} = \log m_r - \log\left\{1 + \sum_{s=1}^{q}\exp\left(x_r^T\beta_s^{(j)}\right)\right\}$ $(r = 1, \ldots, n)$;
(ii) use $\tilde{\theta}^{(j)} = (\beta^{(j)}, \tilde{\phi}_1^{(j)}, \ldots, \tilde{\phi}_n^{(j)})$ to calculate new values $\tilde{H}_r^{(j)}$ $(r = 1, \ldots, n)$;
(iii) 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 up 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, 2011) is available in the Supplementary Material.

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 U.K. Engineering and Physical Sciences Research Council is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes a documented program for the R statistical computing environment, the proof of Theorem 1, and the algebraic manipulations required at the end of § 2.

References

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc.* B *Methodol.* **25**, 220–33.

Bull, S. B., Lewinger, J. B. & Lee, S. S. F. (2007). Confidence intervals for multinomial logistic regression in sparse data. *Statist. Med.* **26**, 903–18.

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

Firth, D. (1992). Generalized linear models and Jeffreys priors: an iterative generalized least-squares approach. In *Computational Statistics I*, Ed. Y. Dodge & J. Whittaker. 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. *Statist. Med.* **21**, 2409–19.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond.* **186**, 453–61.

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**, 563–6.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

[*Received June* 2010. *Revised March* 2011]