Penalized Maximum Likelihood Estimation for Multinomial Logistic Regression using the Jeffreys Prior

by

Shelley B. Bull
Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Prosserman Centre for Health Research, & University of Toronto, Departments of Public Health Sciences & Statistics

and

Juan Pablo Lewinger
University of Toronto, Department of Statistics, and Samuel Lunenfeld Research Institute of Mount Sinai Hospital

and

Sophia S.F. Lee
University of Toronto, Department of Public Health Sciences, and Samuel Lunenfeld Research Institute of Mount Sinai Hospital

Technical Report No. 0505 July 29, 2005

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics
Penalized Maximum Likelihood Estimation for Multinomial Logistic Regression using the Jeffreys Prior

Shelley B. Bull*
Samuel Lunenfeld Research Institute of Mount Sinai Hospital,
Prosserman Centre for Health Research, and
University of Toronto, Departments of Public Health Sciences and Statistics

Juan Pablo Lewinger
University of Toronto, Department of Statistics, and
Samuel Lunenfeld Research Institute of Mount Sinai Hospital

Sophia S.F. Lee
University of Toronto, Department of Public Health Sciences, and
Samuel Lunenfeld Research Institute of Mount Sinai Hospital

28 July 2005

*Corresponding author. Samuel Lunenfeld Research Institute, Prosserman Centre for Health Research, Lebovic Building 5th floor, Mount Sinai Hospital, 600 University Avenue, Toronto, Ont. Canada M5G 1X5. Tel.: +416-586-8245; fax: +416-586-8404. E-mail address: bull@mshri.on.ca (SB Bull).
Summary:

Logistic regression is one of the most widely used regression models in practice, but alternatives to conventional maximum likelihood estimation methods may be more appropriate for small or sparse samples. Modification of the logistic regression score function to remove first order bias is equivalent to penalizing the likelihood by the Jeffreys prior, and yields penalized likelihood estimates (PLEs) that always exist, even in samples in which MLEs are infinite. PLEs are an attractive alternative in small to moderate-sized samples, and are preferred to exact conditional MLEs when there are continuous covariates. We present methods to construct confidence intervals (CI) in the penalized multinomial logistic regression model, and compare CI coverage and length for the PLE-based methods to that of conventional MLE-based methods in trinomial logistic regressions with both binary and continuous covariates. Based on simulation studies in sparse data sets, we recommend profile CIs in preference to asymptotic Wald-type intervals for the PLEs. Furthermore, when finite sample bias and data separation are likely to occur, we prefer PLE profile CIs over MLE methods.

Keywords: Asymptotic bias; Bayesian estimates; Bias reduction; Continuous covariate; Data separation; Infinite estimates; Jeffreys prior; Odds ratio; Polychotomous logistic regression; Polytomous logistic regression; Small samples
1. Introduction

In finite samples, the usual maximum likelihood estimates (MLEs) of the log odds ratio parameters in logistic regression are biased, and there is a non-zero probability that an MLE is infinite, which corresponds to the problem of separation (Albert and Anderson 1984, Lesaffre and Albert 1989). In exponential family models with canonical parameterization, which includes the binomial logistic model, Firth (1993) showed that introducing bias into the score function to remove the order $n^{-1}$ bias of the MLEs is equivalent to penalizing the likelihood by the Jeffreys prior. Ibrahim and Laud (1991) provided theoretical support for the use of the Jeffreys priors in Bayesian analysis of generalized linear models, including the existence of posterior moments, with an application to the binomial logistic model. The penalized likelihood has the advantage that estimates can be obtained in samples in which the MLEs are infinite, and is attractive for routine applications in logistic regression of mixed binary and continuous covariates, in which exact methods may be difficult to apply due to over-conditioning. For example, when one or more of the covariates is continuous and the number of unique covariate combinations is large, support for the exact conditional distribution of the sufficient statistic for the parameter of interest can become extremely discrete or even degenerate, resulting in little or no information left on which to base inferences.

Bull et al. (2002) extended Firth’s modification to the multinomial logistic model, specified a general form for the penalized likelihood, and developed an algorithm for estimation of the regression parameters. In small sample studies of binomial and trinomial logistic regression in a cohort study design with binary and normally distributed covariates, they found that the penalized estimates (PLEs) were effectively unbiased, had smaller mean squared error (MSE)
than the MLEs (with relative MSE as low as 30%), and were more effective in reducing finite-sample bias than alternative methods (Bull et al. 1997). Heinze and Schemper (2002) reported similar results from a simulation study of multiple binary covariates in binomial logistic regression analysis under a case-control design. They also found PLEs to be less biased than alternatives, including exact logistic regression estimates.

Asymptotically, the MLEs are normally distributed around the true parameter value with variance given by the inverse of the Fisher information matrix, but in finite samples the quadratic approximation to the log likelihood may not apply (Jennings 1984). Wald test statistics and confidence intervals (CIs) based on large sample standard errors can have poor properties when the parameter is far from zero (Hauck and Donner 1977); CIs based on the profile likelihood are generally recommended in small samples (Alho 1992). The PLEs are likewise asymptotically distributed, and the first order asymptotic covariance matrix of the PLEs is the same as that of the MLEs (Firth 1993). In finite samples, the PLEs always exist, so both Wald and profile-likelihood CIs can always be constructed. However, use of symmetric Wald-type CIs based on the PLEs may be ill-advised because the small samples in which PLEs are most useful will also be those in which the log-likelihood is not quadratic. Profile-likelihood CIs for infinite MLEs can be constructed with one finite and one infinite endpoint. In contrast, those for the PLEs, while also nonsymmetric, have two finite endpoints.

Based on simulation studies of the penalized binomial logistic regression model with multiple binary covariates reported by Heinze (1999), Heinze and Schemper (2002) recommended profile CIs based on the penalized likelihood in situations with a high probability of separation, but found the Wald CIs to be satisfactory only for balanced covariate distributions and modest log odds ratios less than 1.4. In simulations comparing PLE and MLE profile 95%
CIs, also limited to the binomial model, Bull and Lewinger (2001) reported close to nominal coverage for the PLE profile CIs for log odds ratios as large as 2, even in very small samples sizes of 25 with one binary and one continuous covariate, but in some cases found less than nominal coverage for the MLE profile CIs. They also found greater than nominal coverage for both the MLE and PLE Wald CIs.

In this report, we extend methods for CI construction to the general multinomial logistic regression model, including a derivation of the information matrix for the penalized likelihood that is useful for computation. We compare the performance of the PLE-based methods to that of conventional MLE-based methods with respect to CI coverage and length. We begin in section 2 with a review of notation and methods for point and interval estimation in the usual and penalized multinomial logistic likelihoods. In section 3, we revisit sparse data from a large disease prevention trial in which comparisons of alternative CIs reveal interesting differences. These differences are further investigated in section 4, using finite sample simulations of trinomial logistic models with both binary and continuous covariates, focussing on scenarios in which the ratio of the sample size to the number of parameters is low. A closing discussion compares the simulation findings to those reported previously for the case of a two by two table, and recommends the use of profile CIs for the PLEs in sparse data.

2. Inference for the penalized multinomial logistic likelihood

2.1 Estimation of regression parameters

We consider a multicategory outcome \( y \) that is a multinomial variable with \( J+1 \) categories. For each category \( j = 1, \ldots, J \) there is a regression function in which the log odds
of response in category $j$, relative to category 0, is a linear function of regression parameters and a vector $x$ of $P$ covariates (including a constant): 
$$
\log \left\{ \frac{\text{prob}(y = j \mid x)}{\text{prob}(y = 0 \mid x)} \right\} = \beta_j^T x.
$$
Let $y_i$ be a $J$ by 1 vector of indicators for the observed response category for observation $i$, with the corresponding $J$ by 1 vector of probabilities $\Theta_i = (\Theta_{i1}, ..., \Theta_{iJ})^T$.

The vector of MLEs, $\hat{\beta} = \text{vec}[(\hat{\beta}_1, ..., \hat{\beta}_J)^T]$, is estimated from observations $(y_i, x_i)$, $i=1,...,n$, by solving the score equations of the log-likelihood $l(B)$. The score function is $U(B) = X^T \text{vec}(Y - \Theta)$, with $X^T = X_D^T \otimes I_J$, $X_D^T = (x_1 \mid x_2 \mid ... \mid x_n)$, $Y = (y_1 \mid y_2 \mid ... \mid y_n)$, and $\Theta = (\Theta_1 \mid \Theta_2 \mid ... \mid \Theta_n)$; $\otimes$ is the Kronecker product matrix operator. The $PJ$ by $PJ$ Fisher information matrix is $A = (X^T M X)$, and $M$ is an $nJ$ by $nJ$ block diagonal matrix with $n$ $J$ by $J$ blocks $M_i = \{m_{ij}\}$, $m_{ij} = \Theta_{ij}(1 - \Theta_{ij})$ for $j=k$ and $m_{jk} = -\Theta_{ij} \Theta_{ik}$ otherwise.

In finite samples, the quadratic approximation to the log likelihood may not apply, leading to bias in the MLEs. Furthermore, there is a non-zero probability that an MLE is infinite, i.e. does not exist. Existence problems can occur when the data are sparse or when there are large covariate effects. This situation is characterized by separation in the sample space among the groups that correspond to the categories of the outcome variable, and failure of one or more of the elements of $B$ to converge. It is not unusual for separation to occur in small and moderate-sized datasets, especially in multinomial logistic regression.

In binary regression models, Cordeiro and McCullagh (1991) showed that the finite sample bias in the MLEs is proportional to the true log-odds-ratio parameter and depends on the ratio of the number of parameters to the sample size, provided the magnitude of the linear predictor is small. This relationship also held approximately in trinomial logistic regression simulations (Bull et al 1997).

The order $n^{-1}$ bias of estimates based on the usual likelihood $L(B)$ is removed by applying the penalty $|A|^{1/2}$, which is the square root of the determinant of the Fisher information matrix for the likelihood $L(B)$. Estimation is then based on the penalized
likelihood function $L^*(B) = L(B) \mid A \mid^{1/2}$. The vector $\hat{B}^*$ of penalized estimates (PLEs) is the solution to the score equations of the penalized log-likelihood $l^*(B)$. The introduction of bias into the score function through the penalty removes the leading term in the asymptotic bias of the MLEs. The modified score function proposed by Firth for the binomial logistic model extends directly to the multinomial model as

$$U^*(B) = U(B) - A b(B),$$

where $A$ is the Fisher information for the MLEs, and $b(B)$ is the leading term in the asymptotic bias of the MLEs, obtained from the Taylor series expansion of the log-likelihood of $B$ (Cox and Snell 1968). The bias term $b(B)$ is a function of the matrix of third derivatives of the $l^*(B)$ with respect to $B$, defined explicitly by Bull et al (2002).

As described by Firth (1993), the solution of $U^* = 0$ locates a stationary point of $l^*(B) = l(B) + 1/2 \log |A|$ which is equivalent to the penalized likelihood function $L^*(B) = L(B) \mid A \mid^{1/2}$ with the Jeffreys invariant prior as the penalty function. Use of the Jeffreys prior shrinks estimates toward the point $\Theta_j = 1 / (J + 1)$ which maximizes the determinant and corresponds to $\beta_j = 0$. Arguments for the existence and uniqueness of estimates in the binomial model given by Firth extend to the multinomial model in a straightforward manner (Bull et al 2002). In the simple case of a multinomial response and a binary covariate (considered in Appendix A), which can be summarized in a $(J + 1)$ by 2 contingency table, the PLEs correspond to the usual log odds ratios calculated from cross products from a table in which $1/2$ has been added to each of the table cells.

In general, the PLEs can be obtained by a modified scoring algorithm; using $t$ to denote the iteration number, the modified iterative updating equations are:

$$B^*_{(t+1)} = B^*_{(t)} + A_{(t)}^{-1} U^*(B^*_{(t)})$$
$$= B^*_{(t)} - b(B^*_{(t)}) + A_{(t)}^{-1} U(B^*_{(t)}).$$ (2.1)

with $U(B^*_{(t)}) = X^T vec(Y - \Theta_{(t)})$, and $\Theta_{(t)}$, $A_{(t)}$ and $Q_{(t)}$ evaluated at $B^*_{(t)}$. Thus, in comparison to the usual updating equation used to obtain the MLEs, the score function
modification operates by applying the asymptotic bias corrections at each step in the iterative process; this prevents the estimates from going off to infinity and failing to converge when there is separation in the data. As described previously (Bull et al 2002), we implemented a modified Fisher scoring algorithm based on (2.1) using the matrix programming language GAUSS (Aptech Systems 1990). This algorithm updates with the inverse of $A$, the Fisher information for the MLE's, rather than the information for the PLEs, $A^*$, which includes an additional term corresponding to the second derivatives of the penalty: $1/2 \log |A|$, (see Appendix B).

With increasing sample size, the effect of the penalty diminishes and the PLEs approach the MLEs. Other penalty functions of the form $|A|^c$ could also be applied; Greenland (2000) demonstrates these for the conditional logistic model. Values of $c > 1/2$ correspond to priors stronger than Jeffreys's, further reducing MSE at the cost of introducing negative bias on the log-odds-ratio scale. For this reason, we do not consider them further here.

2.2 Confidence interval construction for regression parameters

Under asymptotic normality of the MLEs, the estimated large sample variance-covariance matrix, $\text{Var}(\hat{B})$, is obtained by evaluating the inverse of the Fisher information, $A^{-1}$, at the MLEs, with standard errors of single parameters corresponding to the diagonal elements of this matrix. Construction of a two-sided $100(1 - \alpha)\%$ confidence interval for a single parameter $\beta_{jp}$ in the parameter vector $\beta_j$ corresponds to inversion of a 1 $df$ family of Wald tests of $H_0: \beta_{jp} = s$, with test statistic $W_M = (\hat{\beta}_{jp} - s)^2 / \text{Var}(\hat{\beta}_{jp})$, which yields the symmetric two-sided confidence interval $\hat{\beta}_{jp} \pm z_{\alpha/2} \sqrt{\text{Var}(\hat{\beta}_{jp})}$. When there is separation in a dataset, and the MLE lies on the boundary of the parameter
space, the Wald CI for the corresponding infinite component of the MLE can be defined as the entire real line \((-\infty, +\infty)\).

The standard error estimates of the PLEs obtained from the information matrix \(A\) evaluated at the PLEs, which can be referred to as plug-in estimates, are smaller than those for the MLEs because the PLEs are generally smaller in magnitude than the MLEs and the standard errors tend to be proportional to the magnitude of the regression estimates. This yields shorter Wald-type confidence intervals. As shown in Appendix B, the information matrix \(A^*\) for the penalized likelihood is considerably more complex than the matrix \(A\), due to the additional higher order terms arising from the log-likelihood penalty. A two-sided 100 \((1 - \alpha)\)% confidence interval for a single parameter \(\beta_{jp}\) is given by the symmetric confidence interval: 
\[
\hat{\beta}_{jp}^* \pm z_{\alpha/2} \text{Var}^*(\hat{\beta}_{jp}^*)
\]
where \(\text{Var}^*\) is a diagonal element of the inverse of the information matrix \(A^*\) evaluated at the PLEs. The higher order terms in \(A^*\) disappear as the sample size increases, and the first order asymptotic covariance of the PLEs is therefore the same as that of the MLEs. In small samples, \(\text{Var}^*(\hat{\beta}_{jp}^*)\) is generally smaller than \(\text{Var}(\hat{\beta}_{jp}^*)\) obtained from \(A\); in Appendix A, this is shown to be a strict inequality in the case of a single binary covariate. However, the latter provides an adequate approximation in moderately sized samples, or when sparse data occur within a larger overall sample.

Likelihood ratio tests, which are asymptotically equivalent to Wald tests, have better properties when the normality of the MLE is in doubt, and confidence intervals based on inversion of the likelihood ratio statistic are not necessarily symmetric, reflecting any departures from a quadratic log-likelihood (Alho 1992 and references therein). A likelihood ratio test statistic has the general form: 
\[
LR_M = 2 \left[ l(\hat{B}) - l(\hat{B}_0) \right],
\]
where \(\hat{B}_0\) are the MLEs estimated under the constraints of the null hypothesis. Profile-likelihood confidence intervals for a single parameter, \(\beta_{jp}\), are obtained by inversion of
the corresponding family of likelihood ratio tests of $H_0: \beta_{jp} = s$. The endpoints for the two-sided 100 $(1 - \alpha) \%$ CI are given by the solution to $LR_M(s) = 2 \left[ l(\hat{B}) - l_0(B(s)) \right] = q$, where $B(s)$ is the argument that maximizes $l$ when $\beta_{jp} = s$, and $q$ is the $(1 - \alpha)$ percentile of a $\chi^2$ distribution with $1 \, df$. When there is separation in a dataset, and the MLE lies on the boundary of the parameter space, the profile likelihood CI for the corresponding infinite component of the MLE has the form $(-\infty, u)$ or $(l, +\infty)$.

Similarly, we can construct asymmetric confidence intervals based on the profile for the penalized likelihood. The profile log-likelihood for the true parameter $\beta_{jp}$ derived from the penalized likelihood $L^*$ is the function $l_0^*(B(s))$, where $B(s)$ is the argument that maximizes $l^*$ under the constraint $H_0: \beta_{jp} = s$. Under this null hypothesis, the likelihood ratio statistic $LR_P(s) = 2 \left[ l^*(\hat{B}^*) - l_0^*(B(s)) \right]$ also follows an asymptotic $\chi^2$ distribution with $1 \, df$. Then the 100 $(1 - \alpha) \%$ CI for $\beta_{jp}$ is given by all parameter estimate values that are compatible with the data, i.e. all $s$ such that $LR_P(s) \leq q$, which is equivalent to $l_0^*(B(s)) \geq \{l^*(\hat{B}^*) - \frac{1}{2} q\}$. The endpoints of the interval are then found by numerically solving the equality for values of $s$. Our preferred algorithm for finding these roots does not require computing $l_0^*(B(s))$, which in itself would involve maximizing $l^*(B)$, but proceeds directly by solving the following constrained optimization problem: maximizing $l^*(B)$ such that $l^*(B) = \{l^*(\hat{B}^*) - \frac{1}{2} q\}$ and $\beta_{jp} = s$. The two points we seek are the intersection of the contour curve $l^*(B) = \{l^*(\hat{B}^*) - \frac{1}{2} q\}$ and the curve defined by $B(s)$.

We modified the algorithm employed in SAS Proc Logistic for MLEs which does not require computing the profile likelihood (SAS Institute 1999). In comparison, this new algorithm is slower, but simple and more robust. SAS uses an iterative scheme, with starting point $\hat{B}$, in which $l(B)$ is replaced by its second order approximation. Our
modification consists of finding a better starting point to improve convergence (see Appendix C for details).

3. Application to a sparse sample

We return to a data set we examined previously (Bull et al. 2002), and apply the confidence interval methods to a trinomial outcome from a multicentre trial of a population intervention designed to prevent post-transfusion hepatitis (Blajchman et al. 1995). The preventive intervention (Treatment factor) under evaluation was the screening of donor units for two surrogate markers of non-A non-B hepatitis infection. Blood transfusion recipients were randomized to receive units from one of two sources: from the general blood supply or from a supply that had excluded units that were positive for the surrogate markers. In addition, while the trial was on-going there was a change in national blood screening policy whereby a new test was introduced to screen all units for hepatitis C antibodies (Time factor). This had the effect of decreasing the incidence of hepatitis C. To evaluate whether the intervention was equally effective before and after the change in screening policy, a test for interaction between the Treatment and Time factors was of interest.

Although the total sample size is large, the disease outcomes are rare, producing empty cells in some subgroups (Table 1). As a result, in the model with an interaction between Time and Treatment, the usual logistic regression MLEs are infinite for two of the parameters (Table 2). The corresponding Wald CIs, which are undefined, are set to be the entire real line (Agresti 1999). The profile likelihood confidence intervals for the infinite MLEs have one infinite endpoint, indicating that we cannot rule out a coefficient of $\beta = +\infty$ for the interaction, and $\beta = -\infty$ for the treatment effect. This corresponds to the possibility that a person with a given covariate value can be affected or unaffected with certainty, which is implausible in most cases, and in this study more likely reflects the low frequency of the hepatitis C outcome. The PLEs, however, can be obtained for all parameters, and in
contrast to the MLEs, the Wald and the profile likelihood confidence intervals for the PLEs have two finite endpoints. In these data, the Wald* intervals, based on A*, are equivalent to the Wald intervals based on A, differing only in the third decimal place. As demonstrated graphically in Figure 1, the PLE profile CI for the simple treatment effect for hepatitis C in the first time period excludes $\beta = 0$, while the corresponding Wald CI based on a quadratic approximation does not, consistent with our expectation that these CIs may have different coverage properties in sparse data. Simulation results reported in the following section suggest that the PLE profile interval is more accurate than the PLE Wald intervals, and that the MLE profile and Wald intervals can be unsatisfactory in sparse data.

4. A Monte Carlo simulation study

4.1 Design

The purpose of the simulation study was to evaluate the properties of the profile and Wald CIs for the PLEs, and to compare their coverage and length to the profile and Wald CIs for the usual MLEs in multinomial logistic regressions that included correlated binary and continuous covariates. Wald CIs for the PLEs were calculated using both A and A* as defined in section 3.2. We programmed the simulations in the matrix language GAUSS. To detect separations, we applied a general algorithm adapted from others and used in previous studies (Lesaffre and Albert 1989; Bull et al. 1997, 2002).

A series of simulations was conducted over a range of sample sizes and slope parameter effect sizes for regressions with one binary and one continuous covariate (details concerning the parameter values are provided in Table 3). We generated the response category by comparing the probabilities calculated from the linear predictor(s), $\beta_j^T x_j$, to a uniform random number. To generate a covariate vector $x_i$ for each observation in a dataset, we first generated variates from a bivariate normal distribution with zero means, unit
variances, and a correlation of 0.8, followed by dichotomization at zero to produce a binary covariate. The covariate correlation induced by dichotomization was estimated to be 0.6 from the simulated covariate vectors. Datasets generated from a model with positively correlated covariates are more likely to show separation than those generated from the same model with uncorrelated covariates.

We tabulated coverage for two-sided 95% CIs, and for upper and lower one-sided 97.5% CIs, as well as the median length of the two-sided CIs. These quantities were also tabulated in the subset of datasets in which all the MLEs were finite, in order to observe differences in the treatment of datasets with separation. To further explore the relationship between coverage and length, as well as the distributions of estimates, their standard errors, CI endpoints, and CI length, we conducted a second set of larger simulations, each with 10,000 replications, that focussed on selected regression models. In this case, the results for each parameter were stratified into deciles of 1000 replicates sorted by the value of the PLE, and summary statistics were tabulated within each decile.

For the studies with 10,000 replications, the precision of the coverage estimates is such that values greater than 95.44% (less than 94.56%) are significantly different from the nominal 95% (exact p<0.025), leading to a conclusion of over (under) coverage on average. For 8,700 replications, the corresponding values are 95.46% (94.53%). We used exact tests for marginal homogeneity (StatXact-5 for Windows) to compare the coverage properties of the PLE profile interval to the PLE Wald* and the MLE profile intervals, assuming ordered categories corresponding to the true parameter being less than the lower CI endpoint, within the CI, or greater than the upper CI endpoint.

Because it was usually one of the binary covariate parameter estimates that produced separation in a dataset, the summaries presented mainly focus on results for the binary covariate estimated in the presence of a correlated normally distributed covariate.
4.2 Results

To illustrate the general patterns observed in the simulations, we present detailed results for a sample size of 50 in models 8 and 9 (Table 4(a) and (b) respectively, with coverage comparisons in Table 5), with density plots of distributions across replications for model 9 (Figure 2). Under model 8, the true slope parameter values are all zero, while under model 9, they are all non-zero. Inspection of these and companion tables reveals that:

(1) As expected, the MLEs and PLEs are both unbiased when the slope parameter is zero, but only the PLEs are unbiased when the parameter is large (Table 4, Figure 2). In the latter case, there is a proportion of the datasets with MLEs that, although not meeting criteria for separation, have log odds ratio estimates greater than 15 (and correspondingly extreme standard errors). Comparison of the distribution of the PLEs in all datasets compared to those in datasets with finite MLEs indicates that datasets with separations tend to yield PLEs with large values (Figure 2).

(2) The median lengths of the PLE intervals are shorter than those of the MLEs, for both Wald and profile CIs (Table 4), although the median interval lengths for all methods increase when datasets with separations are included, due to the CIs being longer (or infinite) for large (or infinite) estimates. The secondary mode in the distribution of the PLE CI lengths corresponds to these separated datasets (Figure 2), and reflects the fact that there is an abrupt transition from lack of separation to complete or partial separation associated with the discreteness of the response frequencies.

(3) The MLE Wald interval is shorter than the corresponding MLE profile interval in the same dataset in over 90% of replicates, while the PLE Wald* interval is always shorter than the PLE profile interval. Furthermore, we find the PLE profile interval to be shorter than the MLE profile interval in all replicates.

(4) Coverage is greater than nominal for the MLE Wald, and for the PLE Wald and Wald* intervals (Table 4). When the parameter is zero, the proportion of datasets in which the
95% CI excludes the parameter value is roughly equal at the lower and upper CI endpoints, but less than 2.5%; in this case, the overcoverage in the tails corresponds to the Wald test having less than a nominal size of 5%. When the parameter is large and positive, it is excluded close to or greater than 2.5% of the time at the upper endpoint, but at the lower endpoint the proportion can be less than 0.5%. Essentially the lower endpoint is biased away from the true parameter value. These intervals therefore may fail to exclude small parameter values as being consistent with the observed data.

(5) The PLE profile intervals yield nominal or slightly higher than nominal coverage, but they also tend to underestimate the parameter at the lower endpoint when it is large and positive.

(6) Comparison of the PLE Wald* interval coverage to the PLE profile interval (Table 5) indicates that the former tends to be generally too conservative in comparison to the latter when the true parameter is zero, although the test for marginal homogeneity is not significant (p=0.25). When the true value of the parameter is large, however, the systematic underestimation of the parameter at the lower endpoint of the Wald* interval is more severe (test for marginal homogeneity p=0.001). Most of the disagreements between the Wald* and the profile intervals occur in either the first or the last decile of the PLE estimates, i.e. when the estimate observed in a particular dataset is farthest from the true parameter value.

(7) Coverage is generally close to nominal for the MLE profile interval, but can fall below nominal in some cases, particularly for the correlated normal covariate, when the true parameters for both covariates are large (Table 4(b)).

(8) When the true parameter for the binomial covariate is zero, the coverage of the PLE profile interval does not differ significantly from that of the corresponding MLE profile interval (test for marginal homogeneity p=0.44). However, when the true parameter is large, the coverage distributions differ (p=0.0001), with the MLE profile interval
overestimating the parameter at the upper endpoint, which appears to be at least partly due to use of an opened-ended interval when the MLE is infinite. Again, disagreements occur in the extreme deciles, with disagreements at the upper endpoint when the observed PLE is less than the true parameter, and disagreements at the lower endpoint when the observed PLE is larger than the true parameter.

Differences among the methods and imbalances in coverage diminish as the sample size increases from 25 through 50 and 100 to 200 (see Figure 3), albeit somewhat less quickly when the true parameter is large due to the bias in the MLEs. The performance of the Wald-type intervals is particularly sensitive to the sample size, whereas the profile intervals are somewhat less so. Attempting to estimate six parameters with a sample size of only 25 yields conservative but relatively uninformative CIs, and more than 50% of the datasets for the model with non-zero parameters are separated, producing distortion in the CI coverage among the datasets without separations. Even for moderate sample sizes of 100, the MLE profile intervals tend to exhibit less than nominal coverage, particularly when the underlying parameter is non-zero. This lack of coverage is at the lower endpoint, in contrast to the PLE intervals in which there is modest overcoverage. For sample sizes of 200, in which no separations were observed, the MLE Wald CI coverage is remarkably similar to that of the PLE profile CI, and both yield close to or greater than nominal 95% coverage, but the latter has shorter median length (data not shown).

5. Discussion

In this report, we present methods to construct confidence intervals for multinomial logistic regression parameters that perform better than standard methods in sparse data sets. Our investigations demonstrate several advantages for the penalized maximum likelihood estimator (PLE). This method always yields finite estimates and confidence interval endpoints for logistic regression parameters. In addition to having
smaller bias and MSE than conventional MLEs (Bull et al 2002), the PLEs generally yield more accurate profile CIs with shorter length and close to or greater than nominal coverage. Comparisons between the Wald and profile CIs show that despite having closer to nominal coverage than Wald CIs, profile CIs have greater length. This can be explained by the symmetric construction of Wald CIs that produces marked overcoverage at the lower endpoint.

Although profile CIs (with one finite endpoint) can be defined for conventional MLEs when an estimate is non-finite, these intervals have less than nominal coverage in some cases, and the absence of a finite point estimate of a parameter will usually be unsatisfactory. Even in datasets without separation, in which all parameter estimates are finite, penalized likelihood profile CIs have comparable coverage and shorter length than CIs based on conventional ML estimation.

Clogg et al (1991), Firth (1993), and others have noted the equivalence between a Bayesian estimator based on the Jeffreys prior and the correction that adds $\frac{1}{2}$ to each cell in a 2 by 2 table (Haldane 1956). The penalized likelihood estimator examined here is a generalization of Haldane's estimator to a general multinomial logistic regression setting. The classical Woolf and Gart logit intervals studied in the 2 by 2 table by Agresti (1999) are thus equivalent to the MLE Wald and PLE Wald* intervals, respectively, in a binomial logistic model with one binary covariate. Although mean coverage levels for these intervals exceeded the nominal, the Woolf interval doing so more greatly than the Gart, Agresti found both intervals on average underestimated the magnitude of the log odds ratio, as might be expected for a non-quadratic likelihood. Minimum coverage, however, could be less than nominal, with minimum coverage of the Gart interval below a nominal 95% level when the log odds ratio was large (in the range of log odds ratios of 1.6 to 3.0), while that of the Woolf interval retained minimum coverage.
close to nominal. In contrast, the profile likelihood interval for the MLE had mean coverage closer to, but slightly less than the nominal level, and minimum coverage generally less than nominal.

In simulations for a binary covariate log odds ratio estimated in a multinomial logistic regression model with a correlated normally distributed covariate, our findings are generally consistent with those of Agresti for a 2 by 2 table. We also found less than nominal coverage for the MLE profile CI in some cases, and greater than nominal coverage for the Woolf and Gart Wald intervals, but did not observe any cases in which the Gart Wald interval had less than nominal coverage. The latter may be a consequence of our simulation design which did not consider a low frequency binary covariate or log odds ratios greater than 2.0. Our results, however, also suggest that a PLE profile interval corresponding to Haldane's estimate would perform better than the Gart Wald interval, and similarly or better than the MLE profile interval, at least for moderate effect sizes.

Rather than smoothing cell probabilities toward equiprobability as in the Gart interval, Agresti (1999) recommended an alternative logit interval based on smoothing toward independence. For binomial logistic regression with multiple categorical covariates, Clogg et al (1991) proposed a general class of shrinkage estimators, with Wald-type confidence intervals for the regression parameters, including both equiprobability and independence smoothing as special cases. The former involves shrinking all regression parameters toward zero, as in the method investigated here, while the latter shrinks the slope parameters toward zero and the intercept toward the marginal response distribution. While Clogg et al (1991) found that independence smoothing performed better for prediction when the response distribution was skewed, and gave sensible inferences in sparse data regressions where maximum likelihood
failed, adjustment in this way does not fully remove the bias in the slope parameter estimates (see for example the simulations reported in Heinze and Schemper 2002) which will usually be unsatisfactory when estimates of association via the log odds ratio are of interest.

As the sample size increases, the penalized estimates become equivalent to the usual MLEs, so routine application of penalized estimation appears to bear only the cost of implementation and additional computation (Heinze and Ploner 2003). We concur with the conclusion of Heinze and Schemper (2002) that the penalized likelihood estimation procedure originally developed by Firth (1993) to reduce finite sample bias in logistic regression parameter estimates is an attractive solution to the problem of separation in logistic regression, and conclude in addition that it may be especially valuable in multinomial regression models where the number of parameters can be large relative to the sample size. We also recommend that the profile CI be used for the PLEs, and emphasize that profile CIs based on the standard likelihood MLEs are not appropriate in sparse datasets when finite sample bias or data separation is likely to occur.
Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Network for Centres of Excellence in Mathematics (Canada). SBB holds a Senior Investigator Award from the Canadian Institutes of Health Research.

References


Heinze, G. (1999, updated 2001) Technical Report 10: The application of Firth's procedure to Cox and logistic regression, Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna.


Table 1. Hepatitis prevention trial data

<table>
<thead>
<tr>
<th></th>
<th>Hepatitis Outcome:</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>Non-ABC</td>
<td>no disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 1: Treated</td>
<td></td>
<td>0</td>
<td>2</td>
<td>400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td></td>
<td>5</td>
<td>3</td>
<td>389</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time 2: Treated</td>
<td></td>
<td>3</td>
<td>10</td>
<td>1896</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untreated</td>
<td></td>
<td>5</td>
<td>11</td>
<td>1864</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Conventional and penalized maximum likelihood estimates, (95% CIs)

(a) hepatitis C outcome

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Time Period</th>
<th>Treatment by Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLE</strong></td>
<td>$-\infty$</td>
<td>$-1.567$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>Wald CI</td>
<td>$(-\infty, +\infty)$</td>
<td>$(-2.811, -0.322)$</td>
<td>$(-\infty, +\infty)$</td>
</tr>
<tr>
<td>Profile CI</td>
<td>$(-\infty, -0.782)$</td>
<td>$(-2.850, -0.283)$</td>
<td>$(-0.152, +\infty)$</td>
</tr>
<tr>
<td><strong>PLE</strong></td>
<td>$-2.426$</td>
<td>$-1.566$</td>
<td>1.957</td>
</tr>
<tr>
<td>Wald CI</td>
<td>$(-5.330, 0.478)$</td>
<td>$(-2.754, -0.378)$</td>
<td>$(-1.242, 5.156)$</td>
</tr>
<tr>
<td>Wald* CI</td>
<td>$(-5.326, 0.475)$</td>
<td>$(-2.753, -0.378)$</td>
<td>$(-1.239, 5.152)$</td>
</tr>
<tr>
<td>Profile CI</td>
<td>$(-7.304, -0.244)$</td>
<td>$(-2.787, -0.345)$</td>
<td>$(-0.702, 6.950)$</td>
</tr>
</tbody>
</table>

(b) hepatitis non-ABC outcome

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Time Period</th>
<th>Treatment by Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLE</strong></td>
<td>$-0.433$</td>
<td>$-0.268$</td>
<td>0.321</td>
</tr>
<tr>
<td>Wald CI</td>
<td>$(-2.228, 1.361)$</td>
<td>$(-1.549, 1.014)$</td>
<td>$(-1.669, 2.311)$</td>
</tr>
<tr>
<td>Profile CI</td>
<td>$(-2.464, 1.369)$</td>
<td>$(-1.437, 1.222)$</td>
<td>$(-1.668, 2.501)$</td>
</tr>
<tr>
<td><strong>PLE</strong></td>
<td>$-0.364$</td>
<td>$-0.376$</td>
<td>0.256</td>
</tr>
<tr>
<td>Wald CI</td>
<td>$(-1.996, 1.268)$</td>
<td>$(-1.580, 0.827)$</td>
<td>$(-1.579, 2.092)$</td>
</tr>
<tr>
<td>Wald* CI</td>
<td>$(-1.994, 1.266)$</td>
<td>$(-1.578, 0.826)$</td>
<td>$(-1.577, 2.090)$</td>
</tr>
<tr>
<td>Profile CI</td>
<td>$(-2.161, 1.278)$</td>
<td>$(-1.486, 0.995)$</td>
<td>$(-1.581, 2.219)$</td>
</tr>
</tbody>
</table>
Table 3. Specifications for the Monte Carlo simulation models: Parameters evaluated for trinomial logistic regression ($J = 2$) with one binary ($x_1$) and one normal covariate ($x_2$) and correlation $\rho(x_1, x_2) = 0.6$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Regression Parameters</th>
<th>$\beta_1^T$</th>
<th>$\beta_2^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>(-1.4, 0, 1)</td>
<td>(-1.4, 2, 0)</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>(-1.4, 1, 1)</td>
<td>(-1.4, 1, -1)</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>(-1.4, 2, 1)</td>
<td>(0.0, 1, -1)</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>(-1.4, 2, 1)</td>
<td>(0.0, 2, -1)</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>(0.0, 0, 1)</td>
<td>(0.0, 2, 0)</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>(0.0, 1, 0)</td>
<td>(0.0, 2, 0)</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>(0.0, 2, 1)</td>
<td>(0.0, 2, 1)</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>(-1.4, 0, 0)</td>
<td>(-1.4, 0, 0)</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>(-1.4, 2, 1)</td>
<td>(-1.4, 2, 1)</td>
</tr>
</tbody>
</table>

Models 1-7 run with a variable number of replications for $n = 50, 75, 100, 200$. Minimum number of replications, respectively, was 3000, 1500, 1200, 600. The number of datasets simulated for each combination of parameters was determined such that the standard error of the mean of the uncorrected MLE was less than 1% of the largest slope parameter value.

Models 4, 6, 7 run with 10,000 replications for $n = 50$.
Models 8, 9 run with 10,000 replications for $n = 25, 50, 100, 200$. 
Table 4. Simulation results (10,000 replicates) for binary and normal covariate slope parameters in $\beta_1^T (n=50)$

(a) Model 8 with parameters $\beta_1^T = (-1.4, 0, 0), \beta_2^T = (-1.4, 0, 0)$

<table>
<thead>
<tr>
<th>Method</th>
<th>Binary Covariate ($\beta=0.0$)</th>
<th>Normal Covariate ($\beta=0.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean $\hat{\beta}$ (Median)</td>
<td>Percent CI Coverage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.006 (0.002)</td>
<td>96.86</td>
</tr>
<tr>
<td>Profile</td>
<td></td>
<td>95.06</td>
</tr>
<tr>
<td>PLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>-0.004 (-0.011)</td>
<td>97.97</td>
</tr>
<tr>
<td>Wald*</td>
<td></td>
<td>97.63</td>
</tr>
<tr>
<td>Profile</td>
<td></td>
<td>96.30</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td></td>
<td>96.90</td>
</tr>
<tr>
<td>Profile</td>
<td></td>
<td>95.19</td>
</tr>
<tr>
<td>All Datasets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=10,000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>0.006 (-0.001)</td>
<td>97.65</td>
</tr>
<tr>
<td>Wald*</td>
<td></td>
<td>97.29</td>
</tr>
<tr>
<td>Profile</td>
<td></td>
<td>95.61</td>
</tr>
</tbody>
</table>
(b) Model 9 with parameters $\beta_1^T = (-1.4, 2, 1)$, $\beta_2^T = (-1.4, 2, 1)$

<table>
<thead>
<tr>
<th>Method</th>
<th>Binary Covariate ($\beta=2.0$)</th>
<th>Normal Covariate ($\beta=1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean $\hat{\beta}$ (Median)</td>
<td>Percent CI Coverage</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>2.10 (2.07)</td>
<td>97.26</td>
</tr>
<tr>
<td>Profile</td>
<td></td>
<td>95.78</td>
</tr>
<tr>
<td>PLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>97.68 (1.85)</td>
<td>97.20</td>
</tr>
<tr>
<td>Wald*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>96.64 (1.85)</td>
<td></td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>---</td>
<td>97.45</td>
</tr>
<tr>
<td>Profile</td>
<td>(2.18)</td>
<td>96.22</td>
</tr>
<tr>
<td>PLE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>97.65 (2.00)</td>
<td>97.17</td>
</tr>
<tr>
<td>Wald*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile</td>
<td>95.78 (1.94)</td>
<td></td>
</tr>
</tbody>
</table>
Table 5. Comparison of CI coverage for simulation Models 8 and 9 for the binary covariate parameter in $\beta_1$ (n=50), including all 10,000 datasets.

<table>
<thead>
<tr>
<th>Model 8</th>
<th>PLE Wald*</th>
<th>MLE Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta=0.0$</td>
<td>$\beta &lt; l$</td>
<td>$l \leq \beta \leq u$</td>
</tr>
<tr>
<td>PLE Profile</td>
<td>$\beta &lt; l$</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>$l \leq \beta \leq u$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$u &lt; \beta$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1.47%</td>
<td>97.29%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 9</th>
<th>PLE Wald*</th>
<th>MLE Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta=2.0$</td>
<td>$\beta &lt; l$</td>
<td>$l \leq \beta \leq u$</td>
</tr>
<tr>
<td>PLE Profile</td>
<td>$\beta &lt; l$</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>$l \leq \beta \leq u$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$u &lt; \beta$</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.26%</td>
<td>97.17%</td>
</tr>
</tbody>
</table>

Note: $l$ and $u$ denote the lower and upper CI endpoints, respectively.
Figure 1.

Depiction of the MLE and PLE profile likelihoods and the PLE quadratic approximation for the treatment effect parameter for the hepatitis C outcome. The scale on the vertical axis in this figure corresponds to \( \{l^*(B) - (l^*(\hat{B}^*)) - \frac{1}{2} q\} \) for the PLE profile likelihood (as defined in section 2.2), where \( q \) is the \((1 - \alpha)\) percentile of a \( \chi^2 \) \((1 \text{ df})\) distribution, with \( \alpha = 0.05 \). The intersection with the horizontal axis thus yields the 95\% interval endpoints. The endpoints for the 95\% PLE Wald interval are the solution to \( \{(\hat{\beta}^* - \beta)^2 / \text{Var}^*(\hat{\beta}^*)\} - q = 0 \), based on a quadratic approximation to the log-likelihood. The MLE profile likelihood is plotted similarly, but intersects the horizontal axis only at the upper CI endpoint.
Figure 2. Distributions of estimates, with density estimation by kernel smoothing, for the binary covariate parameter in $\beta_i^T$, and the corresponding Wald and Profile CI lengths, over 10,000 replicates of datasets of size $n = 50$ from Model 9 for

(a) MLEs in 9,467 datasets without separations (all MLEs are finite),
(b) PLEs in 9,467 datasets without separations (all MLEs are finite), and
(c) PLEs in all 10,000 datasets.
Figure 3.

Proportions of datasets (%) in which the 95% CI for the binary covariate parameter in $\beta_1^T$ excludes the true parameter value at the upper endpoint (above) and the lower endpoint (below), over 10,000 replicates of datasets of size $n = 25, 50, 100, \text{ and } 200$ generated from Model 8 or 9. The proportions at the upper endpoint are plotted as $(1 - \text{ the proportion})$ to show departures from the nominal 97.5%.
**Model 8** with parameters $\beta_1^T = (-1.4, 0, 0), \beta_2^T = (-1.4, 0, 0)$ for

*Upper left*: replicates without separations: all MLEs are finite in 6852, 9467, 9991, and 10,000 datasets of size 25, 50, 100, 200 respectively; *Upper right*: all 10,000 datasets.

**Model 9** with parameters $\beta_1^T = (-1.4, 2, 1), \beta_2^T = (-1.4, 2, 1)$ for

*Lower left*: replicates without separations: all MLEs are finite in 4850, 8699, 9909, and 9999 datasets of size 25, 50, 100, 200 respectively; *Lower right*: all 10,000 datasets.
Appendix A. General multinomial case \((J > 1)\) with a binary covariate \(x \ (P = 2)\)

The data can be summarized in a contingency table with \((J+1)\) rows and two columns.

<table>
<thead>
<tr>
<th>(Y =)</th>
<th>(x = 1)</th>
<th>(x = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(a_0)</td>
<td>(b_0)</td>
</tr>
<tr>
<td>1</td>
<td>(a_1)</td>
<td>(b_1)</td>
</tr>
<tr>
<td>2</td>
<td>(a_2)</td>
<td>(b_2)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(\cdot)</td>
<td>(\cdot)</td>
<td>(\cdot)</td>
</tr>
<tr>
<td>(J)</td>
<td>(a_J)</td>
<td>(b_J)</td>
</tr>
<tr>
<td>(Total)</td>
<td>(n_1)</td>
<td>(n_0)</td>
</tr>
</tbody>
</table>

The penalized likelihood estimates of the table column frequencies are:

\[
\theta_{(1)} = \frac{(a_j + \frac{1}{2})}{n_1^*} \quad \text{and} \quad \theta_{(0)} = \frac{(b_j + \frac{1}{2})}{n_0^*},
\]

where \(n_1^* = \{ n_1 + \frac{1}{2} \ (J+1) \}\) and \(n_0^* = \{ n_0 + \frac{1}{2} \ (J+1) \}\), and the corresponding log odds ratio estimates are:

\[
\beta_j^* = \log \left\{ \left( \frac{\theta_{(1)}}{\theta_{(0)}} \right) / \left( \frac{\theta_{(0)}}{\theta_{(0)}} \right) \right\}.
\]

The matrix \(A^*\) is composed of \(J\) by \(J\) block matrices \(M_0\) and \(M_1\):

\[
A^* = n_1^* \left( I_1 1^T \otimes M_1 \right) + n_0^* \left( i_1 i_1^T \otimes M_0 \right)
\]

with \(1^T = (1 \ 1)\) and \(i_1^T = (1 \ 0)\),

where \(M_1\) and \(M_0\) include the usual multinomial variance and covariance terms in the \(\theta_{(1)}\) and \(\theta_{(0)}\), respectively, as described in general in section 2.1.

Then, based on \(A^*\), the variance estimate for the log odds ratio estimate \(\beta_j^*\) is:

\[
\text{var}_{\beta_j^*} = \left\{ n_1^* \ \theta_{(1)} (1 - \theta_{(1)}) \right\}^{-1} + \left\{ n_0^* \ \theta_{(0)} (1 - \theta_{(0)}) \right\}^{-1},
\]

while in contrast, the variance estimate based on \(A\) is:

\[
\text{var}_{\beta_j} = \left\{ n_1 \ \theta_{(1)} (1 - \theta_{(1)}) \right\}^{-1} + \left\{ n_0 \ \theta_{(0)} (1 - \theta_{(0)}) \right\}^{-1}.
\]
Appendix B. Specification of the Fisher Information $A^*$ for the Penalized Likelihood

The penalized log-likelihood function is $l^*(B) = l(B) + \frac{1}{2} \log |A|$, where $|A|$ is the determinant of the Fisher information matrix: $A = \{- \frac{\partial l(B)}{\partial B} \frac{\partial B^{T}}{\partial B}\}$ The corresponding information $A^*$ for the penalized likelihood is $\{- \frac{\partial l^*(B)}{\partial B} \frac{\partial B^{T}}{\partial B}\}$ and the score function is $U^*(B) = \{- \frac{\partial l^*(B)}{\partial B}\} = U(B) + \frac{1}{2} R$, with $R = W \text{vec}(A^{-1})$ and $W = \{- \frac{\partial A}{\partial B^{T}}\}$. We index the observations by $i = 1, 2, \ldots, n$, and the regression equations (each with $P$ covariate parameters) by $j$ or $k$ or $l = 1, \ldots, J$. $I_{JP}$ is the $JP$ by $JP$ identity matrix, $I_p$ is the $P$ by $P$ identity matrix, and $i_j$ is a unit vector of length $J$, with 1 in position $j$ and 0's elsewhere. We also define $E_i = e_i \otimes I_p$, where $e_i$ is a unit vector of length $n$, with 1 in position $i$ and 0's elsewhere, and $I_J$ is the $J$ by $J$ identity matrix.

By matrix differentiation (details provided below), we have $A^* = A - \frac{1}{2} \{- \frac{\partial R}{\partial B^{T}}\}$. Furthermore, $\{- \frac{\partial R}{\partial B^{T}}\}$

$$= \{ \frac{\partial W}{\partial B^{T}} \} [\text{vec}(A^{-1}) \otimes I_p] + W \{ \frac{\partial \text{vec}(A^{-1})}{\partial B^{T}} \}$$

$$= V [\text{vec}(A^{-1}) \otimes I_p] - W [A^{-1} \{ \frac{\partial A}{\partial B^{T}} \} \text{vec}(A^{-1}) \otimes I_p]$$

$$= \{ V - W [(A^{-1} W) \otimes I_p] \} [\text{vec}(A^{-1}) \otimes I_p]$$

where $V = X^{T} S (X \otimes X \otimes X)$.

For purposes of computation in the general case, the general expressions needed for $A^*$ can be written as:

$$A = \Sigma_i (x_i \otimes I_J) M_i (x_i \otimes I_J)^{T}$$

$$= \Sigma_{jk} [\Sigma_i m_{ijk} x_i x_i^{T}] \otimes I_j I_k^{T}$$

$$W = \Sigma_i (x_i \otimes I_J) Q_i [ (x_i \otimes I_J)^{T} \otimes (x_i \otimes I_J)^{T} ]$$

$$= \Sigma_{jkl} \{ [\Sigma_i q_{jikl} x_i (x_i^{T} \otimes x_i^{T})] (I_p \otimes I_k^{T} \otimes I_p) \} \otimes I_j I_l^{T}$$

32
\[ V = \sum_i (x_i \otimes I_J) S_i \left[ (x_i \otimes I_J)^T \otimes (x_i \otimes I_J)^T \otimes (x_i \otimes I_J)^T \right] \]

\[ = \sum_{jklm} \left\{ \sum_i s_{ijklm} x_i (x_i^T \otimes x_i^T \otimes x_i^T) \right\} (I_p \otimes i_k^T \otimes I_p \otimes i_l^T \otimes I_p) \otimes i_m \]

Then
\[ A^* = A - \frac{1}{2} \left[ R_1 - R_2 \right] \]

with
\[ R_1 = V \left[ \text{vec}(A^{-1}) \otimes I_p \right] \]

\[ = \sum_i (x_i \otimes I_J) S_i \left[ \text{vec}\{ (x_i \otimes I_J)^T A^{-1} (x_i \otimes I_J) \} \otimes (x_i \otimes I_J)^T \right] \]

and
\[ R_2 = W \left[ A^{-1} W \text{vec}(A^{-1}) \otimes I_p \right] = W \left[ A^{-1} T \otimes I_p \right] \]

\[ = \sum_i (x_i \otimes I_J) Q_i \left[ (x_i \otimes I_J)^T A^{-1} T \otimes (x_i \otimes I_J)^T \right] \]

which follows from:
\[ T = W \text{vec}(A^{-1}) = \sum_{i'} (x_{i'} \otimes I_J) Q_{i'} \left[ \text{vec}\{ (x_{i'} \otimes I_J)^T A^{-1} (x_{i'} \otimes I_J) \} \right] \]

\[ W \left[ A^{-1} \otimes I_p \right] = \sum_i (x_i \otimes I_J) Q_i \left[ (x_i \otimes I_J)^T A^{-1} \otimes (x_i \otimes I_J)^T \right]. \]

All quantities (i.e. starting with the probabilities) are evaluated at \( B^* \), the penalized estimates.

**Details:**

In the \( pJ \text{ by } pJ \) Fisher information matrix based on the second derivatives of the log-likelihood with respect to \( B \):

\[ A = - \left\{ \frac{\partial l(B)}{\partial B} \right\} B^{-T} \right\} = (X^T M X), \]

the \( nJ \text{ by } nJ \) block diagonal matrix \( M \) has a \( n \text{ by } n \) block diagonal structure,

with \( i = 1, \ldots, n \) \( J \text{ by } J \) submatrices \( M_i \) such that \( M = \sum_i E_i M_i E_i^T \).

The elements of the submatrix \( M_i \) depend on the probabilities \( \Theta_{ij}, j = 1, 2, \ldots, J \)

with \( M_i = \sum_{jk} m_{ijk} i_j i_k^T \).
and \( m_{jk} = \Theta_y (1 - \Theta_y) \) for \( j = k \)
\[ - \Theta_y \Theta_k \] for \( j \neq k. \)

In the \( pJ \) by \( (pJ)^2 \) matrix of third derivatives:
\[ W = \partial (X^T M X) / \partial B^T = X^T Q (X \otimes X) \]
the \( nJ \) by \( (nJ)^2 \) matrix \( Q \) has a \( n \) by \( n^2 \) block structure,
with \( i = 1, \ldots, n \) \( J \) by \( J^2 \) submatrices \( Q_i \) such that \( Q = \sum_i E_i Q_i (E_i \otimes E_i)^T \).

The elements of the submatrix \( Q_i \) depend on the probabilities \( \Theta_y, j = 1, 2, \ldots, J \)
with \( Q_i = \sum_{jkl} q_{jkl} \ i_j (i_k \otimes i_l)^T, \)
and \( q_{jkl} \)
\[ \Theta_y (1 - \Theta_y) (1 - 2 \Theta_y) \] for \( j = k = l, \)
\[ - \Theta_y (1 - 2 \Theta_y) \Theta_l \] for \( j = k \neq l, \)
\[ - \Theta_y \Theta_k (1 - 2 \Theta_y) \] for \( j \neq k = l, \)
\[ - \Theta_y \Theta_k (1 - 2 \Theta_k) \] for \( j \neq k \) and \( k = l, \) and
\[ 2 \Theta_y \Theta_k \Theta_l \] for \( j \neq k \neq l. \)

Proof that the \( pJ \) by \( (pJ)^3 \) matrix of third derivatives is \( \partial (X^T M X) / \partial B^T = X^T Q (X \otimes X) \)
follows from:
\[ \partial m_{jk} / \partial \beta_i^T = q_{jkl} \ x_i^T \]
\[ \partial m_{jk} / \partial B^T = \sum_i q_{jkl} (x_i \otimes i_i)^T \]
\[ \partial M_i / \partial B^T = \sum_{jkl} q_{jkl} (i_j i_k^T) \otimes x_i^T \otimes i_l^T \]
\[ = \sum_{jkl} q_{jkl} [(i_j i_k^T) \otimes i_l^T] \ [I_j \otimes x_i^T \otimes I_l] \]
\[ \partial M_i / \partial B^T = \sum_i (e_i \otimes I_j) (\partial M_i / \partial B^T) (e_i^T \otimes I_j \otimes I_{lp}) \]
\[ = \sum_i (e_i \otimes I_j) \left\{ \sum_{jkl} q_{jkl} [(i_j i_k^T) \otimes i_l^T] \right\} [e_i^T \otimes I_j \otimes e_i^T X_D \otimes I_j] \]
\[ \frac{\partial (X^T M X)}{\partial B^T} = X^T (\frac{\partial M}{\partial B^T}) (X \otimes I_p). \]

Similarly, proof that the matrix of fourth derivatives is
\[ V = \frac{\partial (X^T Q (X \otimes X))}{\partial B^T} \]
\[ = X^T (\frac{\partial Q}{\partial B^T}) (X \otimes X \otimes I_p) = X^T S (X \otimes X \otimes X) \]
where \( S = \Sigma_i E_i S_i (E_i \otimes E_i \otimes E_i)^T \) and,
\[ S_i = \Sigma_{jklm} s_{ijklm} [ i_j (i_k^T \otimes i_l^T \otimes i_m^T) ] \]
follows from:
\[ \frac{\partial q_{ijkl}}{\partial \beta_m^T} = s_{ijklm} x_i^T \]
\[ \frac{\partial Q_i}{\partial B^T} = \Sigma_{jklm} s_{ijklm} i_j (i_k^T \otimes i_l^T \otimes x_i^T \otimes i_m^T) \]
\[ = \Sigma_{jklm} s_{ijklm} [ i_j (i_k^T \otimes i_l^T \otimes i_m^T) ] [ I_j \otimes I_j \otimes x_i^T \otimes I_j ] \]
where \( s_{ijklm} = \)
\[ \Theta_y (1 - \Theta_y) (1 - 6 \Theta_y (1 - \Theta_y)) \quad \text{for } j = k = l = m \]
\[ - \Theta_y \Theta_{lm} (1 - 6 \Theta_y (1 - \Theta_y)) \quad \text{for } j = k = l \neq m \]
\[ - \Theta_y \Theta_{il} (1 - 6 \Theta_y (1 - \Theta_y)) \quad \text{for } j = k \neq l, m = j \]
\[ - \Theta_{ij} \Theta_{il} (1 - 2 \Theta_y - 2 \Theta_{il} + 6 \Theta_y \Theta_{il}) \quad \text{for } j = k \neq l, m = l \]
\[ 2 \Theta_y \Theta_{il} \Theta_{lm} (1 - 3 \Theta_y) \quad \text{for } j = k \neq l \neq m \]
\[ - \Theta_y \Theta_{ik} (1 - 6 \Theta_y (1 - \Theta_y)) \quad \text{for } j \neq k, k = l = m \]
\[ - \Theta_y \Theta_{ik} (1 - 2 \Theta_y - 2 \Theta_{ik} + 6 \Theta_y \Theta_{ik}) \quad \text{for } j \neq k; j = l, k = m, \text{ or, } j = m, k = l \]
\[ 2 \Theta_{ij} \Theta_{ik} (1 - 3 \Theta_{ij}) \Theta_{lm} \quad \text{for } j \neq k \neq m, j = l \]
\[ 2 \Theta_{ij} \Theta_{ik} \Theta_{il} (1 - 3 \Theta_{lm}) \quad \text{for } j \neq k \neq l, m = j \text{ or } k \text{ or } l, \]
\[ - 6 \Theta_y \Theta_{ik} \Theta_{il} \Theta_{lm} \quad \text{for } j \neq k \neq l \neq m. \]
In the special case of the binomial model, when $J = 1$:

$$A = \sum_i \left[ \Theta_i (1 - \Theta_i) \right] \mathbf{x}_i \mathbf{x}_i^T$$

(which is $p$ by $p$)

$$W = \sum_i \left[ \Theta_i (1 - \Theta_i) (1 - 2 \Theta_i) \right] \mathbf{x}_i (\mathbf{x}_i^T \otimes \mathbf{x}_i^T)$$

(which is $p$ by $p^3$)

$$V = \sum_i \left[ \Theta_i (1 - \Theta_i) (1 - 6 \Theta_i (1 - \Theta_i)) \right] \mathbf{x}_i (\mathbf{x}_i^T \otimes \mathbf{x}_i^T \otimes \mathbf{x}_i^T)$$

(which is $p$ by $p^3$)

and $A^* = A - \frac{1}{2} \left\{ V - W \left[ (A^{-1} W) \otimes I_p \right] \right\} \left[ \text{vec}(A^{-1}) \otimes I_p \right]$

$$= A - \frac{1}{2} \left\{ \sum_i \left[ \Theta_i (1 - \Theta_i) (1 - 6 \Theta_i (1 - \Theta_i)) \right] \mathbf{x}_i (\mathbf{x}_i^T A^{-1} \mathbf{x}_i) \mathbf{x}_i^T \right.$$  

$$- \sum_i \left[ \Theta_i (1 - \Theta_i) (1 - 2 \Theta_i) \right] \mathbf{x}_i \left[ \mathbf{x}_i^T \left( A^{-1} W \right) \text{vec}(A^{-1}) \right] \mathbf{x}_i^T \}$$

$$= A - \frac{1}{2} \sum_i \left[ \Theta_i (1 - \Theta_i) \right] \left\{ (1 - 6 \Theta_i (1 - \Theta_i)) \left( \mathbf{x}_i^T A^{-1} \mathbf{x}_i \right) \right.$$  

$$- (1 - 2 \Theta_i) \left( \mathbf{x}_i^T A^{-1} \mathbf{T} \right) \} \mathbf{x}_i \mathbf{x}_i^T$$

where $T = W \text{vec}(A^{-1})$

$$= \sum_i \mathbf{x}_i \left[ \Theta_i (1 - \Theta_i) (1 - 2 \Theta_i) \right] \left[ \mathbf{x}_i^T A^{-1} \mathbf{x}_i \right]$$

(which is $p$ by 1).

Furthermore, in the special case of a binary covariate, $x = \{0, 1\}$ with $p = 2$, $J = 1$

the matrix $A^*$ is 2 by 2 and simplifies further to:

$$A^* = a_0^* \mathbf{i}_1 \mathbf{i}_1^T + a_1^* \mathbf{1} \mathbf{1}^T$$

where $a_0^* = \{ (n_0 - 1) \Theta_0 (1 - \Theta_0) - \frac{1}{2} (1 - 1/n_0) (1 - 2 \Theta_0)^2 \}$,

$$a_1^* = \{ (n_1 - 1) \Theta_1 (1 - \Theta_1) - \frac{1}{2} (1 - 1/n_1) (1 - 2 \Theta_1)^2 \}$$

with estimates $\Theta_0 = (b + \frac{1}{2}) / (n_0 + 1)$ and $\Theta_1 = (a + \frac{1}{2}) / (n_1 + 1)$.

Therefore, based on $A^*$, the variance estimate for the log odds ratio estimate: $\beta^* = \log \{ (\Theta_1 / (1 - \Theta_1)) / (\Theta_0 / (1 - \Theta_0)) \}$ is $\text{var}^* = (a_0^*)^{-1} + (a_1^*)^{-1}$, while in contrast the variance estimate based on $A$ is $\text{var} = (a_0)^{-1} + (a_1)^{-1}$, where $a_0 = \{ n_0 \Theta_0 (1 - \Theta_0) \}$ and $a_1 = \{ n_1 \Theta_1 (1 - \Theta_1) \}$. 

36
Appendix C. Computation of Profile Likelihood Confidence Intervals

The endpoints for the two-sided 100 (1 - \( \alpha \)) % CI for the parameter \( \beta_{jp}, j=1,..J, \)
p=1,..P, are the solutions to \( \text{LR}_M(s) = 2 \left[ l(\hat{B}) - l(B(s)) \right] = q, \) where \( B(s) \) is the argument that maximizes \( l(B) \) when \( \beta_{jp} = s, \) and \( q \) is the \( (1 - \alpha) \) percentile of a \( \chi^2 \) distribution with 1 df. Here \( l(B) \) represents either the usual log-likelihood or the log of the penalized likelihood and \( \hat{B} \) is the MLE or the PLE, respectively. Thus, the endpoints of profile likelihood confidence intervals for \( \beta_{jp} \) are given by the values \( s \in \mathbb{R} \) defining the intersection of the ‘profile’ curve \( B(s) \) with the level curve \( l(B) = l_0, \) where \( l_0 = \{ l(\hat{B}) - \frac{1}{2} q \}. \) To obtain an iterative algorithm for computing the points of intersection of the profile and the level curves, the log-likelihood (or log-penalized likelihood) function is approximated in a neighbourhood of \( B \) by the quadratic function

\[
l_q(B + \delta) = l(B) + \delta^T U + \frac{1}{2} \delta^T A \delta
\]  

(A.1)

where \( U = U(B) \) is the gradient of \( l(B) \) with respect to \( B \) and \( A = A(B) \) is the Hessian. The increment \( \delta \) for the next iteration is obtained by solving the equations

\[
d/d\delta \{ l_q(B + \delta) + \lambda ( e_{jp}^T (B + \delta) - s ) \} = 0
\]  

(A.2)

where \( \lambda \) is a Lagrange multiplier, \( e_{jp} \) is the unit vector that extracts the element corresponding to \( \beta_{jp}, \) and \( s \) is an unknown parameter determined by the condition

\[l_q(B + \delta) = l_0.\]  
The solution is \( \delta = - A^{-1} (U - \lambda e_{jp}) \)

with \( \lambda = \pm \) square root of \( \left\{ 2 \left( l_0 - l(B) + U^T A^{-1} U \right) / (e_{jp}^T A^{-1} e_{jp}) \right\}. \)

To find the upper endpoint of the CI for \( \beta_{jp}, \) iteration starts at \( \hat{B} \) using positive \( \lambda \) values until convergence is achieved. The process is repeated with negative values of \( \lambda \) for the lower endpoint. This is the algorithm described in the documentation for SAS
PROC LOGISTIC (SAS Institute 1999). In our experience however, starting at the MLE (PLE) does not always guarantee convergence. A better starting point can be obtained by ‘bracketing’ the solution of

\[ \{ \mathbf{B}(s), \ s \in R^l \} \cap \{ l(\mathbf{B}) = l_0 \} \]  

(A.3)

using bisection. For this we find a point \( \beta^1 = \mathbf{B}(s_1) \) such that \( l(\beta^1) < l_0 \). Because \( l(\hat{\mathbf{B}}) = l(\mathbf{B}((\hat{s}_j + \beta_{jp})/2)) > l_0 \), there is a solution to (A.3) given by \( \mathbf{B}(t) \), with \( t \in (s_1, \beta_{jp}) \). The point \( \beta^2 = \mathbf{B}((s_1 + \beta_{jp})/2) \) is now closer to such solution. Bisection can continue in this fashion for a few iterations to find a better starting point for (A.2) or it can continue until convergence is attained. This provides a slower but more robust way to find the endpoints of profile CIs. To find \( \beta^1 \) we move in small steps \( t > 0 \) in the (positive or negative) direction \( e_{jp} \), i.e. parallel to the \( \beta_{jp} \) axis and starting at \( \hat{\mathbf{B}} \). Setting \( \gamma^0 = \hat{\mathbf{B}} \) and \( \gamma^n = \mathbf{B}(\gamma_{jp}^{n-1} + t) \), we iterate until \( l(\gamma^n) < l_0 \), in which case we set \( \beta^1 = \gamma^n \). This procedure requires computing \( \mathbf{B}(s) \) for \( s \in R \). Since the curve \( \mathbf{B} : R^l \rightarrow R_{jp} \) is defined for each \( s \in R^l \) as the solution to the maximization problem \( \mathbf{B}(s) = \text{argmax} \ l(\mathbf{B}) \) subject to \( \beta_{jp} = s \), this can be done using an iterative algorithm similar to (A.2). The log-likelihood function is approximated in a neighbourhood of \( \mathbf{B} \) by the quadratic function in (A.1) and the increment \( \delta \) for the next iteration is obtained by solving the score equations

\[ \frac{d}{d\delta} \{(l_0(\mathbf{B} + \delta) + \lambda(e_{jp}^T (\mathbf{B} + \delta) - s) = 0 \}

where \( \lambda \) is the the Lagrange multiplier. The solution is

\[ \delta = -A^{-1}(U - \lambda e_{jp}) \quad \text{with} \quad \lambda = -((s - \beta_{jp} + e_{jp}^T A^{-1} U) / (e_{jp}^T A^{-1} e_{jp})). \]

Iteration proceeds until convergence is achieved with step halving performed at each step to ensure the procedure is going uphill in \( l \).