A simple and accurate method for approximate conditional inference in linear exponential family models

By RICHARD P. WATERMAN

Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

and

BRUCE G. LINDSAY

Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

Summary

A simple method for approximate conditional inference is described. The methodology is applied to canonically parameterized linear exponential family models where it is shown to provide remarkably accurate approximations to fully conditional estimates. The approximation technique can be applied much more generally than in this particular class of models. The technique depends only on the construction of certain "projected scores" derived from higher-order likelihood derivatives and their covariances, and so can be used in many problems where it is relevant to control for the estimation of nuisance parameters.

Some key words: Conditional inference, Conditional score; Generalized linear models; Linear exponential family, Nuisance parameters; Projected score.

1. Introduction

This paper presents two applications of a conceptually simple method for approximate conditional inference. One is the problem of adjusting for ties in Cox's partial likelihood for the proportional hazards model, and the other involves inference for item parameters in a Rasch model. The methodology used to provide these approximations allows for inference about a multivariate parameter of interest in the presence of a multivariate nuisance parameter and permits any set of conditioning events that correspond to linear combinations of the response variable. A theoretical treatment of the projective approach to nuisance parameter problems with an emphasis on approximating conditional score functions can be found in Waterman and Lindsay (1994). Here we restrict ourselves to a less formal presentation. A variety of methods are available for approximate or, in some cases, exact conditional inference. These include the exact methods of Hirji et al. (1987), the saddlepoint approximations described by Davison (1988) and the orthogonal parameterization technique of Cox and Reid (1987). We believe that our approach provides a useful alternative to these techniques because it is simple to interpret, can deal simultaneously with a multivariate parameter of interest and very large numbers of nuisance parameters, and can be extended easily to a much wider range of problems than just the linear exponential family that we consider here.

2. Construction of the projected scores

Consider n independently and identically distributed random variables from a distribution with density $f(y;\psi,\lambda)$, where ψ denotes an r-dimensional parameter of interest and λ denotes an s-dimensional nuisance parameter. Let $T(\psi)$ be a complete sufficient statistic for λ at each fixed ψ. Then previous results (Godambe, 1976; Lindsay, 1982, 1983) have established the conditional score function, $U_c(\psi,\lambda;y)$, as an optimum estimating equation within the class of estimating equations satisfying the unbiasedness condition $E_{\psi\lambda_0}\{g(\psi,\lambda;y)\} = 0$ for all ψ, λ and λ_0, when the optimality criterion is that of maximal information (Godambe, 1976).

The conditional score function $U_c(\psi,\lambda;y)$ can be expressed as

$$U_c(\psi,\lambda;y) = U_0 - E\{U_0 \mid T(\psi)\},$$

where U_0 is the score for the parameter of interest and $E\{U_0 \mid T(\psi)\}$ is the projection of U_0 onto the space of T-measurable functions (Lindsay, 1982). This space will be denoted by Λ_∞. Under appropriate regularity conditions the functions

$$V_{i_1,\ldots i_t} = \frac{\partial^t f(y;\psi,\lambda)}{\partial\lambda_{i_1}\cdots\partial\lambda_{i_t}} \Big/ f(y;\psi,\lambda), \qquad 1 \le i_j \le s, \quad j = 1,\ldots,t \qquad t = 1,\ldots$$

provide a basis for Λ_∞. These functions will be called the Bhattacharyya basis since they were used extensively by A. Bhattacharyya (1946).

The first approximation to U_c, denoted by U_1, is obtained by removing from U_0 its projection onto the closed linear space generated by $\{V_{i_1}\}$ rather than its projection onto the space Λ_∞. The second approximation U_2 is formed by removing from U_0 its projection onto the closed linear space generated by $\{V_{i_1}, V_{i_1 i_2}\}$. The obvious extension to higher order projected scores is easily seen, but we will not make use of them here. The projected score U_1 is in fact the score for the least favorable family distribution for ψ, and U_2 has been used by McLeish and Small (1988) and Small and McLeish (1989) in their work on desensitizing inference functions to nuisance parameters. Here we focus on the projected scores' ability to approximate a conditional score function. The projected scores are invariant under transformations of the nuisance parameter, which is seen in the simplest case, that is the construction of U_1, by noting that a reparameterization of λ modifies the V_1 score function only by the introduction of a multiplicative term, the Jacobian of the transformation. When the Jacobian is of full rank the space spanned by the original score functions V_{i_1} is identical to that spanned by the transformed score functions, and consequently the projection of U_0 onto these spaces is the same.

A simple but extremely useful recurrence relation between the Bhattacharyya basis elements is given by

$$V_{i_1\ldots i_{t+1}} = \frac{\partial V_{i_1\ldots i_t}}{\partial\lambda_{i_{t+1}}} + V_{i_1\ldots i_t} V_{i_{t+1}}. \tag{1}$$

Since we will not be making use of the higher order projected scores the general indexing notation of $i_1,\ldots,i_t$ can be simplified so that the first order Bhattacharyya basis elements are indexed by i and the second order elements by (i,j). In this indexing notation a special case of (1) which will be used in §3 is

$$V_{ij} = \frac{\partial V_i}{\partial\lambda_j} + V_i V_j. \tag{2}$$

3. Canonically parameterized Linear Exponential Family models

3·1. *Derivation of the U_2 projected score equation*

Following McCullagh and Nelder (1989, p.23) the kernel of the log-likelihood for the generalized linear model with n independent observations y_k, canonical link and known dispersion parameter can be written as

$$l(\theta; y) = \sum_{k=1}^{n} y_k \eta_k - \sum_{k=1}^{n} b_k(\eta_k) \quad \text{where } \theta = (\psi, \lambda)^T \text{ and } \eta_k = x_k \theta.$$

The coefficient η_k is the canonical parameter and the dispersion parameter is included in the y_k and b_k. Corresponding to the partition of θ into ψ and λ there is a partition of the covariate vector for the k^{th} observation, x_k, into the two components $x_k = (z_k, \gamma_k)$, so that the matrix of covariates can be written as $X = (Z, \Gamma)$, where Z is the covariate matrix for the parameters of interest and Γ the covariate matrix for the nuisance parameters.

In this model a complete sufficient statistic for the nuisance parameter λ is $\Gamma^T y$ which does not depend on the parameter of interest. This means that the conditional score function U_c is in fact a function of ψ alone and provides a globally optimal estimating equation for ψ. Ideally one would carry out inference for ψ after having conditioned on $\Gamma^T y$ (Cox and Hinkley, 1979, Ch.5.2). In most problems the resulting conditional distribution will be intractable, hence the need to approximate it. The approximations will depend on the nuisance parameter λ and in order to estimate it we use the obvious estimate $\hat{\lambda}_\psi$, the maximum likelihood estimator of λ for ψ fixed.

To construct the approximation to U_c the Bhattacharyya basis elements must be regressed out of each component of the score function for the parameter of interest. While in theory it is possible to derive an arbitrary order approximation U_t, it would of course be desirable if a lower order approximation, say U_2, was sufficiently close to U_c to make higher order approximations unnecessary. To determine if this is indeed the case we will compare the U_2 projected score with the U_c score function. Straightforward calculations show that $U_0 = Z^T\{y - b'(\eta)\}$, where $b'(\eta)$ is a vector whose k^{th} element is $b'(\eta_k)$, and $'$ denotes differentiation with respect to η. Similarly the Bhattacharyya basis function V_i, is the i^{th} component of the vector $\Gamma^T\{y - b'(\eta)\}$. By making use of

(2) it is easy to show that V_{ij} is the $(i,j)^{th}$ element of the matrix

$$\Gamma^T[\{y - b'(\eta)\}\{y - b'(\eta)\}^T - b''(\eta)]\,\Gamma,$$

where $b''(\eta)$ is a diagonal matrix with elements $b''(\eta_k)$. Since we are dealing with canonically parameterized models, the vector $b'(\eta)$ is simply the mean vector of the y_k, which will be denoted by m, and the diagonal matrix $b''(\eta)$ is the variance-covariance matrix of the y_k, which we write as M_2. In this notation the U_0 score function and the first order Bhattacharyya basis elements are components of the vectors $Z^T(y - m)$ and $\Gamma^T(y - m)$ respectively. The second order elements are components of the matrix

$$\Gamma^T\{(y - m)(y - m)^T - M_2\}\,\Gamma.$$

Performing the regression of U_0 against the first and second order Bhattacharyya basis functions is initially a daunting task since there are $s(s+1)/2$ linearly independent functions V_{ij} and s might well be of the same order as the sample size. However, it is possible to achieve a large degree of simplification due to the following facts. Firstly, it is not necessary to calculate the full U_2 score function in order to find its solution provided that $\hat{\lambda}_\psi$ is used to estimate λ. Secondly, the highly structured form of the Bhattacharyya basis elements induces structural properties in the projected scores themselves, which can be exploited for computational purposes.

To illustrate the first of these points consider for notational simplicity a model with univariate ψ and λ. Then the conventional maximum likelihood estimators are the roots of the simultaneous system of equations $U_0 = 0$ and $V_1 = 0$. Since U_1 is defined as U_0 with its projection onto the space spanned by V_1 removed, we can write U_1 as $U_0 - \rho V_1$ where ρ is the regression coefficient. The coefficient ρ is found as the solution to the orthogonality constraint $E\{(U_0 - \rho V_1)V_1\} = 0$, so that $\rho = E(U_0V_1)/E(V_1^2)$, (note that these expectations are taken at the same values that the score functions are evaluated at). The estimate for ψ obtained as the root of the simultaneous equations $U_1 = 0$ and $V_1 = 0$ is again the maximum likelihood estimator because at the solution $V_1 = 0$ so that $U_1 = U_0 - \rho\, 0 = U_0$. The U_2 score can be written as $U_0 - \rho_1 V_1 - \rho_2 V_{11}$ where ρ_1 and ρ_2 are the regression coefficients. Using th e representation (2) of V_{11} as $\partial V_1/\partial\lambda + V_1^2$ it is clear that the only part of U_2 that needs to be calculated in order to solve $U_2 = 0$ and $V_1 = 0$ is $U_0 - \rho_2 \partial V_1/\partial\lambda$. Consequently the adjusted estimates of ψ can be calculated as the root of the simultaneous equations $U_0 - \rho_2\partial V_1/\partial\lambda = 0$ and $V_1 = 0$. The adjusted estimate is therefore the root of a modified maximum likelihood estimating equation, where the modification to U_0 is the simple correction term $\rho_2 \partial V_1/\partial\lambda$.

As noted earlier the multivariate ψ and λ case is technically more complicated but conceptually identical to the univariate ψ and λ situation. After some lengthy but routine calculations that involve finding the correction term $\rho_2 \partial V_1/\partial\lambda$ for the matrix case, the relevant part of the U_2 score equation, denoted by $\tilde{U}_2$ is

$$\tilde{U}_2 = Z^T(y-m) + Z^T \Xi M_3 \Delta^T (\Gamma \otimes \Gamma) \left[(\Gamma^T \otimes \Gamma^T) \Omega (\Gamma \otimes \Gamma) \right]^+ (\Gamma^T \otimes \Gamma^T) \Delta M_2 1. \quad (3)$$

Before proceeding to identify the components of (3) note that $\tilde{U}_2$ can indeed be written as $U_0 + Correction$. The correction term may appear at first sight complicated, but it is no more than a product of matrices and the appendix shows its calculation to much less computationally demanding than it would first appear.

The components of (3) are defined as follows: m is the vector of means and the diagonal matrices M_2, M_3 and M_4 contain the second, third and fourth cumulants of the random variables Y_k.

- $\Xi = I - M_2 \Gamma (\Gamma^T M_2 \Gamma)^{-1} \Gamma^T$
- $A = M_2 \odot M_2 + \Delta (M_4 - M_3 M_2^{-1} M_3) \Delta^T$
- Δ is an $n^2 \times n$ dimensional matrix with components δ^r_{ij} where $\delta^r_{ij} = 1$ if $i = j = r$ and 0 otherwise.
- $\Omega = A + \Delta (M_3 M_2^{-1} \Xi M_3) \Delta^T$
- The vector 1 denotes an n-dimensional vector of ones.

Note the dependence on the first four cumulants and the appearance of the projection matrix Ξ with an inverse variance metric. The generalized inverse, denoted by $^+$ is of an $s^2 \times s^2$ dimensional matrix, simplifications of which exist principally because the $n^2 \times n^2$ dimensional matrix $M_2 \odot M_2$, a permutation product of the matrix M_2, and defined elementwise as $(M_2 \odot M_2)_{ij,rs} = M_{2\,ir} M_{2\,js} + M_{2\,is} M_{2\,jr}$, has a simple to calculate generalized inverse (McCullagh, 1987, p.20, ex. 1.12). The appendix contains a more detailed look at how the simplifications are achieved. Some comments on the form of (3) are in order:

Remark 1. For models in which the random component is normal the entire correction term is 0 because the third cumulant, M_3, that is the skewness, is 0.

Remark 2. For Poisson models all cumulants are identical and equal the mean. Consequently the matrix $M_4 - M_3 M_2^{-1} M_3$, which is a component of the matrix A, is 0. Naturally this leads to simplifications in the calculations.

3·2. Computational details

To find the approximate conditional estimates the two sets of equations

$$\tilde{U}_2 = 0 \text{ and } \Gamma^T(y - m) = 0 \tag{4}$$

are solved simultaneously for ψ. A straightforward approach is to use a Newton-Raphson algorithm with the obvious candidates for starting values being the un-conditional maximum likelihood estimators. The Newton-Raphson procedure requires the derivative of (3). Since the derivative appears to be formidable a numerical derivative was used as an approximation to the true derivative. An alternative to the numerical differentiation is to use the derivative of the U_0 score function which gives a Fisher scoring type algorithm.

At present we make no claims to have found the most computationally efficient way of solving the equations in (4) and this remains a topic for future research. In particular it would be interesting to determine if an iteratively reweighted least squares algorithm is feasible.

The most complex task in the calculation of $\tilde{U}_2$ is to find the inverse of an $n \times n$ dimensional matrix. Greater simplifications can be achieved in problems with greater structure. For example, with a Poisson random component and $Z \oplus \Gamma = R^n$, as would be the case for inference concerning higher order interactions conditional on complete sufficient statistics for all lower order interactions and main effects in a multiway table, then the most computationally intensive part of the simplification involves finding the square root of an $r \times r$ full rank symmetric matrix.

4. Examples

In this section we present two examples that demonstrate the accuracy of the U_2 projected score equation for canonically parameterized generalized linear models, and a simple example that outlines the construction of the $\tilde{U}_2$ projected score equation. The section also includes a simulation study that addresses the conditional properties of the projected score estimates.

4·1. *Proportional hazards with ties*

In the standard Cox's proportional hazards model (Cox, 1972) and (McCullagh and Nelder, 1989, p.426) the j^{th} failure time makes a contribution to the partial likelihood of

$$\frac{\exp(\beta^T x_j)}{\sum \exp(\beta^T x)},$$

with summation over the risk set $\mathcal{R}(T_j) = \{j : T_k \geq T_j\}$, where T_j denotes the j^{th} failure time.

In the presence of ties arising from k discrete failure times this contribution becomes

$$\frac{\exp(\beta^T s_j)}{\sum \exp(\beta^T s)} \tag{5}$$

where the sum in the denominator is over all distinct groups of k individuals in $\mathcal{R}(T_j)$ and s denotes the sum of the covariate vectors for the k individuals. In the presence of heavy ties this contribution to the partial likelihood is inordinately cumbersome to calculate. There are a variety of ways to deal with ties. Two commonly used methods are those of Peto (1972) and Efron (1977). Peto's method is really an approximation to a grouped continuous time likelihood, whereas Efron's was suggested as an approximation to Cox's partial likelihood, which is more appropriate if the time scale is genuinely discrete. For a discussion of adjustment for ties see Kalbfleisch and Prentice (1973), Whitehead (1980) and Cox and Oakes (1984, p.101).

The partial likelihood (5) may be maximized by using its formal equivalence to a conditional logistic regression model in which the conditioning statistic for each failure time is the number of failures occurring at time j, in other words the complete sufficient statistic for the intercept parameter in the unconditional logistic regression model. It is this conditional interpretation that we will exploit with the use of the approximate conditional score function U_2.

The data used for this example was taken from McCullagh and Nelder (1989, Table 13.1). The parameter of interest is a log odds ratio and included in Table 1 are the parameter estimates and their estimated standard errors. The estimated standard error for the approximate conditional estimate is derived from a symmetrized estimated version of the "observed information" type matrix $-\nabla(\tilde{U}_2, V_1)^T$, where in a slight abuse of notation V_1 represents the vector containing all the first order Bhattacharyya basis elements. The information matrix is not symmetric because it does not correspond to the derivative of a score function derived from a genuine loglikelihood.

TABLE 1 ABOUT HERE

Clearly the approximation is very successful for both the point estimate and its estimated standard error. In this problem the number of nuisance parameters is equal to the number of failure times, which for this particular data set was 17 and clearly illustrates the ability of the methodology to cope with multivariate nuisance parameters.

4·2. Rasch model

As a second example we use the data presented in Lindsay, Clogg, and Grego (1991, p.100) concerning test scores of 216 people on a four item test. A Rasch model for the probability that the i^{th} person gets the j^{th} item correct is

$$f(y_{ij};\psi_j,\lambda_i) = \frac{\exp\{(\lambda_i - \psi_j)y_{ij}\}}{1+\exp(\lambda_i - \psi_j)},$$

where ψ_j is an item difficulty parameter and λ_i is a subject ability parameter. Note that in this case both the parameter of interest and the nuisance parameter are multivariate, being of dimensions 3 and 14 respectively. To complete the Rasch model two assumptions are needed; firstly no interaction between subject and item parameters and secondly independence of responses across items.

In order to make the comparison with the exact estimate we have exploited links between the logistic model and a multiplicative Poisson model (Tjur, 1982). The parameters estimates are given in Table 2 and the estimate derived from U_2 is referred to as the "approximate conditional" estimate. To make the parameters identifiable ψ_4 has been set equal to 0.

TABLE 2 ABOUT HERE

An efficient coding of the data makes use of weights, which explains why there are only essentially 14 nuisance parameters and not 216, the number of people who took the test. We note that (3) can be easily adapted to deal with weights, in fact the 1 vector simply becomes a column of weights and the vector $Z^T(y-m)$ is appropriately weighted exactly as in the case of generalized linear models.

The unconditional estimates are the unadjusted maximum likelihood estimators for the parameters of interest. In practice of course one would not consider using them, but they are included here because they demonstrate the distance that the approximate conditional estimates have to jump to reach the exact conditional estimates. It is clear that the gap between unconditional and exact conditional is large, but that the U_2 score function again does an extremely good job of approximating U_c both in terms of the point estimate and its estimated standard error. We have not included estimated covariances in Table 2, but the relative error between the exact and projected estimates is of the order of 3%, whereas it is about 2% for the variances.

4·3. *Illustration of the calculations for a paired binomial experiment*

This subsection provides a simple example to illustrate the calculation of the projected score equation (3) and demonstrates approximate conditional inference for a 2 by 2 table. Two independent binomial random variables are observed with sample sizes and probabilities of success n_1 and n_2, and π_1 and π_2 respectively. The model is

$$Y_1 \sim Bin\left(n_1, \frac{\exp(\lambda+\psi)}{1+\exp(\lambda+\psi)}\right) \qquad Y_2 \sim Bin\left(n_2, \frac{\exp(\lambda-\psi)}{1+\exp(\lambda-\lambda)}\right).$$

The parameter of interest is 2ψ, the log odds ratio.

The kernel of the loglikelihood is given by

$$y_1 \log\left(\frac{\pi_1}{1-\pi_1}\right) + y_2 \log\left(\frac{\pi_2}{1-\pi_2}\right) + n_1 \log(1-\pi_1) + n_2 \log(1-\pi_2). \qquad (6)$$

Using the definitions of π_1 and π_2, (6) can be written as

$$y_1\,\eta_1 + y_2\,\eta_2 - b_1(\eta_1) - b_2(\eta_2)$$

where $\eta_1 = \lambda + \psi$, $\eta_2 = \lambda - \psi$ and $b_1(\eta_1) = n_1 \log\{1+\exp(\eta_1)\}$. The function $b_2(\eta_2)$ is defined analogously. In terms of covariates η can be expressed as

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \psi \\ \lambda \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \psi + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \lambda,$$

which identifies the matrices Z and Γ as $(1,-1)^T$ and $(1,1)^T$ respectively. Once the first four cumulants of the binomial random variables have been found then (3) can be calculated.

We state without proof the interesting fact that when $n_1 = n_2 = 1$, so that the experiment consists of observing a Bernoulli pair then the projected score function U_2 is identically equal to the conditional score function U_c obtained by conditioning the data on the complete and sufficient statistic for λ, $y_1 + y_2$.

4·4. *Conditional Simulation*

The results presented for the Rasch model certainly indicate that the projected score has the potential to be close to the conditional score but questions still remain regarding the conditional properties of the projected score estimate. The two questions we will answer here are whether or not the degree of accuracy of the approximation that we observed in Table 2 is consistently repeatable and if there is any systematic bias between the estimates provided by the U_2 projected score and the exact conditional estimates. To investigate these questions we have carried out a conditional simulation based on the Rasch model example data. 1000 data sets were constructed that mirrored the conditioning restrictions of the original data, that is that the number of people that got three questions correct was 33, the number that got two correct was 63 and the number who got just one question correct was 55. (Note that those students who get perfect scores or those that get zeroes do not enter into a conditional analysis.) The components of ψ were set equal to 2.9431, 1.2440, 1.3204 and 0.0 respectively and λ_i was set equal to -1 for all λ_i simply for convenience. The values of the nuisance parameters, λ_i, are not relevant in the exact conditional analysis or for the conditional simulation but their values would be relevant for assessing properties of the estimates un-conditionally.

Table 3 contains comparisons of the un-conditional maximum likelihood estimates and the projected score estimates with the exact conditional estimates. For each method two quantities have been calculated. The first is the average relative error of the estimate compared the exact conditional estimate and the second is the average absolute error relative to the estimated standard error of the conditional estimate. To be more explicit, if the estimates of ψ_1 are denoted by $\hat{\psi}_1$, $\tilde{\psi}_1$ and $\hat{\psi}_{1c}$ standing for the un-conditional maximum likelihood estimator, the U_2 projected score estimator and the finally the exact conditional maximum likelihood estimator respectively, then the four columns in the table correspond to the means over the 1000 replicates of the quantities

$$\frac{\text{abs}(\hat{\psi}_1 - \hat{\psi}_{1c})}{\hat{\psi}_{1c}} \qquad \frac{\text{abs}(\tilde{\psi}_1 - \hat{\psi}_{1c})}{\hat{\psi}_{1c}} \qquad \frac{\text{abs}(\hat{\psi}_1 - \hat{\psi}_{1c})}{\widehat{s.e.}(\hat{\psi}_{1c})} \qquad \frac{\text{abs}(\tilde{\psi}_1 - \hat{\psi}_{1c})}{\widehat{s.e.}(\hat{\psi}_{1c})}.$$

The same summary measures have been calculated for the other parameters of interest, ψ_2 and ψ_3.

TABLE 3 ABOUT HERE

The two most striking features of the table are just how disastrous the un-conditional maximum likelihood estimates are, and the relative success of the U_2 projected score estimates. The relative absolute error of the U_2 projected score estimates to the exact conditional estimate is clearly small, and the absolute error, based on the standard error scale is also acceptable when one takes into account the fact that a 95% confidence interval has length of about four standard errors. The simulation showed that a systematic bias between $\tilde{\psi}_i$ and $\hat{\psi}_{ic}$ was indeed present and more noticeable for ψ_2 and ψ_3 than for ψ_1. The worst situation was for ψ_2 where the mean of $\tilde{\psi}_2 - \hat{\psi}_{2c}$ over the 1000 replicates was -0.0196, with a standard deviation of 0.00725. At issue here from a practical perspective is whether or not this systematic bias seriously detracts from the usefulness of the estimates. If one considers that the standard error of $\hat{\psi}_{2c}$ is roughly 0.26 so that a 95% confidence interval would have a width of about 1, then the systematic bias in this instance, is not a major component of the inherent uncertainty associated with the entire estimation procedure.

Given the systematic bias one could consider correcting for it. If it were deemed to pose a real threat to the quality of the analysis then the natural step in our formulation of the problem would be to construct the U_3 projected score and use that for inference. For the Rasch model example, the additional computational complexity involved would outweigh its usefulness.

5. Conclusion

In this paper we have presented a simple method for approximate conditional inference, and demonstrated its remarkable accuracy in two commonly applied models. We believe that the approximation will complement existing exact inference methods because it can cope with models with extremely large numbers of nuisance parameters. Furthermore, its ability to deal with a multivariate parameter of interest plus an essentially arbitrary linear (in y) conditioning event make it an extremely flexible tool. In effect the analyst simply has to specify two matrices, the usual design matrix for the parameters of interest and a matrix Γ containing contrasts for the nuisance parameters.

This methodology undoubtably has parallels with techniques for second order correct inference. Part of our ongoing research is to contrast the various techniques available. A method for the adjustment of profile likelihoods proposed by McCullagh and Tibshirani (1990) is an obvious candidate for comparison. At this stage we just note that the projected score approach is certainly different as demonstrated by McCullagh and Tibshirani's Example 6, where they consider n binary matched pairs. The projected score technique yields the exact conditional score function in this problem, whereas McCullagh and Tibshirani's adjusted profile likelihood and also Barndorff-Nielsen's modified profile log-likelihood (Barndorff-Nielsen, 1986) do not.

Two interesting questions for future research are to determine the effect on the estimates of ψ of using a simpler estimator for the nuisance parameter λ in place of $\hat{\lambda}_\psi$. Also, since the score equations are solved in an iterative fashion it would be pertinent to examine the efficiency of a "one-step" approximation to the root of U_2.

As a final point, a referee has pointed out that using standard errors to asses the accuracy of the estimates is not necessarily useful if the conditional likelihood is asymmetric about its maximum, see Davison (1988). The focus of this paper has been to show that the conditional score function can be well approximated about its root by the projected score function, and the whole issue of confidence interval construction and hypothesis testing based on the projected scores is a part of our ongoing research and a subject that we hope to have more to say on later.

A computer program to calculate the roots of the $\widetilde{U}_2$ score function and their estimated standard errors is available from the first author.

Acknowledgments

We would like to thank Professor Clifford. C. Clogg for many illuminating discussions and comments on a draft version of this paper. The paper was improved by a referee's helpful review, who also suggested the conditional simulation described in §4.4. Bruce Lindsay's research was partially supported by NSF grant DMS-9106895.

Appendix

Simplification of the $\widetilde{U}_2$ score function

The simplifications exploit a standard result for the Moore Penrose inverse of a sum of non-negative definite matrices (Boullion and Odell, 1971, Theorem 6,p.16). We use a special case in which the matrix Q lies in the column space of a matrix P when the result becomes (I denotes an identity matrix of appropriate dimension):

$$(PP^T + QQ^T)^+ = P^{+T}\{I - P^+Q(I + Q^T P^{+T} P^+ Q)^{-1} Q^T P^{+T}\} P^+.$$

Two steps are taken to simplify the $n \times n$ dimensional matrix

$$\Delta^T(\Gamma \otimes \Gamma)\{(\Gamma^T \otimes \Gamma^T)\Omega(\Gamma \otimes \Gamma)\}^+(\Gamma^T \otimes \Gamma^T)\Delta, \tag{A·1}$$

which is the computationally fearsome component of $\widetilde{U}_2$.

For the first stage the matrix $M_3 M_2^{-1} \Xi M_3$ is factored into CC^T where C is an $n \times (n-s)$ dimensional matrix. This allows the matrix Ω to be written as $\Omega = A + \Delta C C^T \Delta^T$, and (A·1) simplifies to

$$R - RC(I + C^T RC)^{-1} C^T R$$

where

$$R = \Delta^T(\Gamma \otimes \Gamma)\{(\Gamma^T \otimes \Gamma^T) A (\Gamma \otimes \Gamma)\}^+(\Gamma^T \otimes \Gamma^T)\Delta.$$

It can then be shown that

$$R = H \mp HD(I \pm D^T HD)^{-1} D^T H \tag{A·2}$$

where

$$H = \tfrac{1}{2}\{\Gamma(\Gamma^T M_2 \Gamma)^{-1}\Gamma^T\} * \{\Gamma(\Gamma^T M_2 \Gamma)^{-1}\Gamma^T\}, \tag{A·3}$$

and D is the square root of the diagonal matrix $|M_4 - M_3 M_2^{-1} M_3|$, with the $*$ operator being the Hadamard product. The $\mp$ and $\pm$ depends on the sign of the elements of $(M_4 - M_3 M_2^{-1} M_3)$. The signs in (A·2) are $-$ then $+$ if the elements of $(M_4 - M_3 M_2^{-1} M_3)$ are positive and $+$ then $-$ if the elements are negative.

The key point here is that (A·1) which involves the generalized inverse of an $s^2 \times s^2$ dimensional matrix has been expressed in terms of the matrices C, D and H, all of which can be readily calculated.

References

Barndorff-Nielsen, O. E. (1986) Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, **73**, 307–22.

Bhattacharyya, A. (1946) On some analogues of the amount of information and their uses in statistical estimation. *Sankhyā*, **8**, 1–14: 201–218: 315–328.

Boullion, T. L. and Odell, P. L. (1971) *Generalized Inverse Matrices*. New York: Wiley.

Cox, D. R. and Hinkley, D. V. (1979) *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall.

Cox, D. R. (1972) Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **74**, 187–220.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B*, **49**, 1–39.

Davison, A. C. (1988) Approximate conditional inference in generalized linear models. *J. R. Statist. Soc. B*, **50**, 445–61.

Efron, B. (1977) The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Assoc.*, **72**, 557–565.

Godambe, V. P. (1976) Conditional likelihood and unconditional optimal estimating equations. *Biometrika*, **63**, 277–284.

Hirji, K. F., Mehta, C. R., and Patel, N. R. (1987) Computing distributions for exact logistic regression. *J. Am. Statist. Assoc.*, **82**, 1110–1117.

Kalbfleisch, J. D. and Prentice, R. L. (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267–278.

Lindsay, B. G. (1982) Conditional score functions, some optimality results. *Biometrika*, **69**, 503–512.

Lindsay, B. G. (1983) Efficiency of the conditional score in a mixture setting. *Ann. Statist.*, **11**, 486–497.

Lindsay, B. G., Clogg, C. C., and Grego, J. (1991) Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Am. Statist. Assoc.*, **86**, 96–107.

McCullagh, P. (1987) *Tensor Methods in Statistics.* London: Chapman and Hall.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd. ed. No. 37 in Monographs on Statistics and Applied Probability. London: Chapman and Hall.

McCullagh, P. and Tibshirani, R. (1990) A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc. B*, **52**, 325–344.

McLeish, D. L. and Small, C. G. (1988) *The Theory and Applications of Statistical Inference Functions* Vol. 44 of *Lecture Notes in Statistics.* New York: Springer-Verlag.

Peto, R. (1972) Contribution to the discussion of Cox. *J. R. Statist. Soc. B*, **34**, 205–207.

Small, C. G. and McLeish, D. L. (1989) Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, **76**, 693–703.

Tjur, T. (1982) A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Stat.*, **9**, 23–30.

Waterman, R. P. and Lindsay, B. G. (1994) The accuracy of projected score methods in approximating conditional scores. *Submitted*,.

Whitehead, J. (1980) Fitting Cox's regression model to survival data using GLIM. *Appl. Statist.*, **29**, 268–75.

Captions for tables

Table 1:

Comparison of estimates for a log-odds ratio in Cox's proportional hazards model.

Table 2:

Comparison of unconditional, approximate conditional and exact inference in a Rasch model.

Table 3:

A conditional comparison of approximate conditional and exact conditional estimates in a Rasch model.

Table 1:

METHOD	Estimate	Est.std.err.
EXACT (COX)	1.6282	0.4332
WATERMAN/LINDSAY	1.6270	0.4325
EFRON	1.5721	0.4124
PETO	1.5088	0.4096

Table 2:

Param.	Estimate			Est.std.err.		
	Un-conditional	Approx. conditional	Exact conditional	Un-conditional	Approx. conditional	Exact conditional
ψ_1	4.1605	2.9431	2.9445	0.3733	0.2973	0.2988
ψ_2	1.7167	1.2247	1.2440	0.3130	0.2604	0.2631
ψ_3	1.8219	1.3016	1.3204	0.3137	0.2607	0.2633

Table 3:

Parameter	Mean $\frac{\text{abs}(\hat{\psi}_i - \hat{\psi}_{ic})}{\hat{\psi}_i}$	Mean $\frac{\text{abs}(\tilde{\psi}_i - \hat{\psi}_{ic})}{\hat{\psi}_i}$	Mean $\frac{\text{abs}(\hat{\psi}_i - \hat{\psi}_{ic})}{\widehat{s.e.}(\hat{\psi}_{ic})}$	Mean $\frac{\text{abs}(\tilde{\psi}_i - \hat{\psi}_{ic})}{\widehat{s.e.}(\hat{\psi}_{ic})}$
ψ_1	0.415	0.00175	4.0623	0.0172
ψ_2	0.382	0.01549	1.7884	0.0732
ψ_3	0.382	0.01429	1.8931	0.0714