Advances in Pure Mathematics, 2016, 6, 331-341
Published Online April 2016 in SciRes. http://www.scirp.org/journal/apm
http://dx.doi.org/10.4236/apm.2016.65024

Inference in the Presence of Likelihood Monotonicity for Polytomous and Logistic Regression

John E. Kolassa

Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ, USA
Email: kolassa@stat.rutgers.edu

Received 21 November 2015; accepted 19 April 2016; published 22 April 2016

Abstract

This paper addresses the problem of inference for a multinomial regression model in the presence of likelihood monotonicity. This paper proposes translating the multinomial regression problem into a conditional logistic regression problem, using existing techniques to reduce this conditional logistic regression problem to one with fewer observations and fewer covariates, such that probabilities for the canonical sufficient statistic of interest, conditional on remaining sufficient statistics, are identical, and translating this conditional logistic regression problem back to the multinomial regression setting. This reduced multinomial regression problem does not exhibit monotonicity of its likelihood, and so conventional asymptotic techniques can be used.

Keywords

Polytomous Regression, Likelihood Monotonicity, Saddlepoint Approximation

1. Introduction

We consider the problem of inference for a multinomial regression model. The sampling distribution of responses for this model, and, in turn, its likelihood, may be represented exactly by a certain conditional binary regression model.

Some binary regression models and response variable patterns give rise to likelihood functions that do not have a finite maximizer; instead, there exist one or more contrasts of the parameters such that as this contrast is increased to infinity, the likelihood continues to increase. For these models and response patterns, maximum likelihood estimators for regression parameters do not exist in the conventional sense, and so monotonicity in the likelihood complicates estimation and testing of binary regression parameters. Because of the association be-

How to cite this paper: Kolassa, J.E. (2016) Inference in the Presence of Likelihood Monotonicity for Polytomous and Logistic Regression. *Advances in Pure Mathematics*, **6**, 331-341. http://dx.doi.org/10.4236/apm.2016.65024

tween binary regression and multinomial regression, multinomial regression methods inherit this difficulty. In particular, methods like those suggested by [1], using higher-order asymptotic probability approximations like those of [2], are unavailable in these cases, since the methods of [2] use values of the maximized likelihood, both with the parameter of interested fixed and allowed to vary, and use the second derivatives of the likelihood at these two points.

[3] provides a method for diagnosing and adjusting for likelihood monotonicity for conditional testing in binary regression models. This manuscript extends this method to facilitate estimation in multinomial regression, for approximate inference, and in particular makes practical the use of the approximation of [2].

Section 2.1 reviews binary and multinomial regression models, and relations between these models that let one swap back and forth between them. Section 2.2 reviews conditional inference for canonical exponential families. Section 2.3 reviews techniques of [3] for performing conditional inference in the presence of likelihood monotonicity for binary regression, makes a suggestion for improving the efficiency of this earlier technique, and expands on its implication for estimation. Section 2.4 reviews some existing techniques for addressing likelihood monotonicity. Section 2.5 develops new techniques for detection of likelihood monotonicity in multinomial regression models, and explores discusses non-uniqueness of maximum likelihood estimates in this case. Section 3 applies the techniques of Section 2.5 to some examples. Section 4 presents some conclusions.

2. Methods and Materials

This section describes existing methods used in cases of likelihood monotonicity in multinomial models, and presents new methods for addressing these challenges.

2.1. Multinomial and Logistic Regression Models

Methods will be developed in this manuscript to address both multinomial and binary regression models. In this section, relationships between these models are made explicit.

Consider first the multinomial distribution. Suppose that M multinomial trials are observed; for trial $m \in \{1,\cdots,M\}$, one of J_m alternatives is observed, with alternative $j \in \{1,\cdots,J_m\}$ having probability

$$p_{mj} = n_{mj}\exp\left(\boldsymbol{x}_{mj}\boldsymbol{\beta}\right)\Big/\sum_{k=1}^{J_m} n_{mk}\exp\left(\boldsymbol{x}_{mk}\boldsymbol{\beta}\right). \tag{1}$$

Here $\boldsymbol{x}_{mj} \in \Re^K$ (for $\Re$ representing the real numbers and K a positive integer) are covariate vectors associated with each of the alternatives, and n_{mj} are the number of replicates with this covariate pattern. These probabilities depend only on on the differences between $\boldsymbol{x}_{mj}$ and $\boldsymbol{x}_{mj'}$ for $j \neq j'$; without loss of generality we will take $\boldsymbol{x}_{mJ_j} = (0,\cdots,0)^{\mathrm{T}}$, treating the last category as a baseline. Let $\{W_m, m=1,\cdots,M\}$ be independent random variables such that $P[W_m = j] = p_{mj}$, and let $Y_{mj} = 1$ if $W_m = j$ and $Y_{mj} = 0$ if $W_m \neq j$. Then the variables W_m are the indices of the selected multinomial outcomes, and Y_{mj} are related indicator variables. The likelihood is given by

$$\underset{\boldsymbol{\beta}}{P}\left[Y_{mj} = y_{mj}\forall m, j\right] = L(\boldsymbol{\beta}) = \prod_{m=1}^{M}\prod_{j=1}^{J_m} p_{mj}^{y_{mj}}. \tag{2}$$

Let $\boldsymbol{X}_m$ be the matrix with J_m rows and K columns, with row j given by $\boldsymbol{x}_{mj}$, let $\boldsymbol{Y}_m = \left(Y_{m1},\cdots,Y_{mJ_m}\right)^{\mathrm{T}}$, and let $\boldsymbol{X} = \left(\boldsymbol{X}_1^{\mathrm{T}},\cdots,\boldsymbol{X}_M^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^{\mathrm{T}},\cdots,\boldsymbol{Y}_M^{\mathrm{T}}\right)^{\mathrm{T}}$. Then sufficient statistics for $\boldsymbol{\beta}$ are

$$\boldsymbol{S} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{3}$$

The binary regression model is similar; let $\{V_m, m=1,\cdots,M\}$ be independent binomial random variables with mass function

$$\underset{\boldsymbol{\beta}}{P}[\boldsymbol{V} = \boldsymbol{v}] = \mathcal{L}(\boldsymbol{\beta}) = \prod_{m=1}^{M}\binom{n_m}{v_m} q_m^{v_m}\left(1-q_m\right)^{n_m-v_m} \tag{4}$$

for

$$q_m = \frac{\exp\left(\boldsymbol{z}_m \boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{z}_m \boldsymbol{\beta}\right)}, \tag{5}$$

where $\boldsymbol{z}_m \in \Re^{K'}$, and K' a positive integer. Sufficient statistics for $\boldsymbol{\beta}$ are given by

$$\boldsymbol{T} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V} \tag{6}$$

for $\boldsymbol{Z}$ the $M \times K'$ matrix with row *m* equal to $\boldsymbol{z}_m$.

The binary regression model can be recast as a multinomial regression model. Furthermore, the multinomial regression model may be expressed as a conditional binary regression model. Suppose that (1) and (2) hold. Let $\boldsymbol{E}_m$ be a matrix with J_m rows and *M* columns, such that the entries in column *m* are all 1, and all other entries are 0. Let $\boldsymbol{Y}_m = \left(Y_{m1},\cdots,Y_{mJ_m}\right)^{\mathrm{T}}$

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{E}_1 \\ \vdots & \vdots \\ \boldsymbol{X}_m & \boldsymbol{E}_m \end{pmatrix}, \text{ and } \boldsymbol{V} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_M \end{pmatrix}, \tag{7}$$

with $\boldsymbol{v}$ defined analogously, and let $\boldsymbol{T} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}$ and $\boldsymbol{t} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{v}$.

2.2. Conditional Inference

The model for $\boldsymbol{T}$ given by (4)-(5) represents a canonical exponential family, and so inference on some components of $\boldsymbol{\beta}$ (without loss of generality, $\beta_1,\cdots,\beta_I$) may be performed by considering the sampling distribution of $T_1,\cdots,T_I$ conditional on $T_{I+1},\cdots,T_{K'}$. Let $Q_{\boldsymbol{\beta},I}\left(\boldsymbol{t}\right)$ represent the probability mass function for this conditional distribution. The probability mass function for $\boldsymbol{S}$ of (3), evaluated at $\boldsymbol{s}$, is exactly the same as

$$Q_{\boldsymbol{\beta}}\left(\boldsymbol{t}\right) = \underset{\boldsymbol{\beta}}{P}\left[T_1 = t_1,\cdots,T_I = t_I \mid T_{I+1} = t_{I+1},\cdots,T_{K'} = t_{K'}\right],$$

for $\boldsymbol{t} = \left(\boldsymbol{s}, n_1,\cdots,n_M\right)^{\mathrm{T}}$. Note here the conditioning event for the larger model is expressed in terms of sample sizes in the smaller model. Furthermore, one can relate probabilities calculated with the regression parameters set to zero, to the probabilities for a general parameter vector $\boldsymbol{\beta}$, by

$$Q_{\boldsymbol{\beta},I}\left(\boldsymbol{t}\right) = \frac{\exp\left(\sum_{j=1}^{I} t_j\beta_j\right) Q_{\boldsymbol{0},I}\left(\boldsymbol{t}\right)}{\sum_{(t_1,\cdots,t_I)\in\mathcal{T}} \exp\left(\sum_{j=1}^{I} t_j\beta_j\right) Q_{\boldsymbol{0},I}\left(\boldsymbol{t}\right)}. \tag{8}$$

When $I = 1$, define the confidence interval for one of the regression parameters (without loss of generality β_1) with nominal coverage $1-\alpha$ for $\alpha \in (0,1)$, by

$$\mathcal{I}\left(t_1,\alpha\right) = \left\{\beta_1 \mid \sum_{s\le t_1} Q_{\beta_1}\left(s,t_2,\cdots,t_K\right) \ge \frac{\alpha}{2}, \sum_{s\ge t_1} Q_{\beta_1}\left(s,t_2,\cdots,t_K\right) \ge \frac{\alpha}{2}\right\}. \tag{9}$$

Then $\mathcal{I}\left(T_1,\alpha\right)$ has coverage probability at least $1-\alpha$, and can be used as a $1-\alpha$ confidence interval. In fact, the coverage $P_{\beta_1}\left[\beta_1 \in \mathcal{I}\left(T_1,\alpha\right)\right]$ may be strictly greater than $1-\alpha$, and more precise intervals with at least $1-\alpha$ coverage may be constructed as $\mathcal{I}\left(T_1,\alpha^*\right)$, for

$$\alpha^* = \min\left\{\alpha^\dagger \mid \underset{\beta_1}{P}\left[\beta_1 \in \mathcal{I}\left(T_1,\alpha^*\right)\right] \ge 1-\alpha\ \forall \beta_1\right\}. \tag{10}$$

The cumulative probabilities implicit in (9) may be approximated as

$$\sum_{s\ge t_1} Q_{\beta_1,1}\left(s,t_2,\cdots,t_K\right) \approx 1-\Phi\left(\omega\right)+\phi\left(\omega\right)\left(1/\zeta - 1/\omega\right), \tag{11}$$

for Φ and ϕ the standard normal distribution function respectively, $\omega = \sqrt{2\left(\ell\left(\hat{\boldsymbol{\beta}}\right)-\ell\left(\tilde{\boldsymbol{\beta}}\right)\right)}$, for $\ell\left(\boldsymbol{\beta}\right)$ the logarithm of the likelihood (in the multinomial regression case, given by (2)), $\hat{\boldsymbol{\beta}}$ maximizing $\ell(\boldsymbol{\beta})$, with data adjusted for continuity so that t_1 is reduced by half, $\tilde{\boldsymbol{\beta}}$ maximizing $\ell(\boldsymbol{\beta})$, with β_1 fixed at its null value, and $\zeta = \left(\hat{\beta}_1 - \beta_1\right)\sqrt{\det\left(\ell''\left(\hat{\boldsymbol{\beta}}\right)\right) \Big/ \det\left(\ell''_{-1}\left(\tilde{\boldsymbol{\beta}}\right)\right)}$, with ℓ''_{-1} the matrix of second derivatives of ℓ with respect to all but the first component of the argument [2].

Similar techniques may be applied to the multinomial regression model, with $\boldsymbol{S}$ of (3) replacing $\mathbf{T}$ of (6). Denote the conditional sample space by

$$\mathcal{S} = \left\{\sum_{m=1}^{M} z_{mj} v_m \ \forall j \le I \mid \sum_{m=1}^{M} z_{mj} v_m = s_j \ \forall j > I\right\}. \tag{12}$$

2.3. Infinite Estimates

This section reviews and clarifies techniques for inference in the presence of monotonicity in the logistic regression likelihood (4) given by [3], who built on results of [4]-[6], and [7]. Choose $\boldsymbol{\lambda} \in \Re^K$. Let

$$\mathcal{U} = \left\{\boldsymbol{u} \in \Re^K \mid \lim_{a \to \infty} \mathcal{L}\left(\boldsymbol{\lambda} + \boldsymbol{u}a\right) > 0\right\},$$

for $\mathcal{L}$ the logistic regression likelihood of (4) [3], building on the results of [4], determines which observations correspond to extreme fitted probabilities by maximizing the total number of positive entries in both $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ subject to constraints outlined in the next theorem.

Theorem 1. *Suppose that random vectors* $\boldsymbol{V}$ *arise from the model* (4) *and* (5), *and matrix* $\boldsymbol{Z}$ *is of full rank. Then* $\mathcal{U} - \mathbf{0}$ *is exactly the set of vectors*

$$\boldsymbol{u} = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{\sigma} - \boldsymbol{\rho}\right), \tag{13}$$

with $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ column vectors such that $\boldsymbol{\sigma} \ne \boldsymbol{\rho}$ and such that

$$\sigma_j \ge 0, \rho_j \ge 0 \ \forall j, \tag{14}$$

$$\boldsymbol{t}^{\mathrm{T}}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\left(\boldsymbol{\sigma} - \boldsymbol{\rho}\right) - \boldsymbol{n}^{\mathrm{T}}\boldsymbol{\sigma} = 0, \text{ and } \boldsymbol{Z}^{\perp}\left(\boldsymbol{\sigma} - \boldsymbol{\rho}\right) = \mathbf{0}, \tag{15}$$

where $\boldsymbol{T} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}$ are the sufficient statistics associated with $\boldsymbol{\beta}$, $\boldsymbol{t}$ is the observed value of this vector $\boldsymbol{T}$, and $\boldsymbol{Z}^{\perp}$ is a matrix with M rows and rank $M - K'$, such that $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}^{\perp} = \mathbf{0}$.

Furthermore, the conditional probabilities are the same as those arising if observations with positive entries in $\boldsymbol{\sigma}$ or $\boldsymbol{\rho}$ are omitted, and collinear covariates among columns $I+1,\cdots,K$ removed.

The matrix $\boldsymbol{Z}^{\perp}$ may be constructed from the QR decomposition of $\boldsymbol{Z}$. Inference on components of $\boldsymbol{\beta}$ for (4) may be performed conditionally, even when maximum likelihood estimates fail to exist, and hence (11) cannot be used [3].

2.4. Other Approaches

Suppose that there exists a vector $\boldsymbol{u}$ such that

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_{mj} \ge 0 \text{ if } Y_{mj} = 1, \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}_{mj} \le 0 \text{ if } Y_{mj} = 0, \tag{16}$$

with strict inequality holding in place of at least one of the inequalities. Then the likelihood $L\left(\boldsymbol{\lambda} + a\boldsymbol{u}\right)$, defined in (2), is a strictly increasing function of a for any $\boldsymbol{\lambda} \in \Re^K$, and so no finite maximizer of L exists. This lack of a finite maximizer leads to difficulties with maximum likelihood estimation and inference. This section reviews some existing approaches.

Bias-correction is possible for maximum likelihood estimators [8], and may be employed in this type of cituation [9]. Estimates with this correction applied are the same as those maximizing the posterior density of the parameters under the invariant prior of [10], in which the likelihood function is multiplied by the square root of the determinant of the information matrix; this is equivalent to maximizing a penalized likelihood. That is, if $\ell(\boldsymbol{\beta}) = \log\left(L(\boldsymbol{\beta})\right)$, for $L(\boldsymbol{\beta})$ as in (2), then the approach of [8] suggests maximizing $\ell^{\dagger}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - 12\det\ell''(\boldsymbol{\beta})$. These estimates are always finite. A similar approach is possible in the case of proportional hazards regression, which also gives advantages for testing [11].

Standard errors may be calculated from the second derivative of the unpenalized log likelihood [8]; one could also calculate standard errors from the second derivative of the penalized log likelihood [11]. This second approach is used in this manuscript for comparison purposes, and so the asymptotic confidence intervals considered below for β_1 is

$$\tilde{\beta}_1 \pm z_{\alpha/2}\sqrt{\ell^{\dagger 11}\left(\tilde{\boldsymbol{\beta}}\right)} \text{ for } \tilde{\boldsymbol{\beta}} = \operatorname{argmax}\left(\ell^{\dagger}\left(\boldsymbol{\beta}\right)\right). \tag{17}$$

Here the superscript 11 on $\ell^{\dagger}$ represents component 11 of the inverse second derivative matrix of ℓ with respect to $\boldsymbol{\beta}$.

The approach using Jeffreys' prior to penalize the likelihood has some disadvantages. The union of all possible confidence intervals resulting as the penalized estimator plus or minus a multiple of the standard error has a finite range, and so the confidence region procedure as described above has vanishing coverage probability for large values of the regression parameter.

2.5. Estimation

We investigate the behavior of maximum likelihood estimates in the multinomial regression model (2). Maximizers for both the original likelihood and for the likelihood of the distribution of sufficient statistics of interest (6) or (3) conditional on the remaining canonical sufficient statistics are considered. Conditional probabilities arising from the logistic regression model (4) are of form (2), and so may also be handled as below.

Consider the occurrence of infinite estimates for model (2). Denote the sample space for the $\boldsymbol{S}$ of (3) by $\mathcal{T}$. For a canonical exponential family with finite support, the maximizer of the likelihood associated with a data point $\boldsymbol{s}$ exists if and only if $\boldsymbol{s} \in \operatorname{hull}\left(\mathcal{T}\right)^{\circ}$, where $\operatorname{hull}\left(\cdot\right)$ represents the convex hull of its argument, and the superscript $\circ$ represents the interior of the set it modifies ([12], Theorem 9.4). The following two corollaries may be used to determine if the finite maximizers of (2) exist, in terms of the observed sufficient statistic $\boldsymbol{s}$ of (3). The first is a corollary to ([12], Theorem 9.4).

Corollary 1. *Unique finite maximizers of the likelihood given in* (2) *exist if and only if* $\boldsymbol{X}$ *is of rank at least K, and the maximizer of* c *subject to*

$$\boldsymbol{s} = \sum_{m=1}^{M}\sum_{j=1}^{J_m}\alpha_{mj}\boldsymbol{x}_{mj}, \sum_{j=1}^{J_m}\alpha_{mj} = 1\ \forall m \in \left\{1,\cdots,M\right\}. \tag{18}$$

$$\alpha_{mj} \geq c\ \forall m \in \left\{1,\cdots,M\right\},\ j \in \left\{1,\cdots,J_m\right\} \tag{19}$$

is greater than zero.

Proof. If $\boldsymbol{X}$ is not of full rank, then multiple parameter values give the same value of $\boldsymbol{X\beta}$, and maximizers cannot be unique.

Suppose that $\boldsymbol{X}$ is of full rank. The set $\mathcal{T}$ is defined from (3) as the set of multiples of rows of $\boldsymbol{X}$, with the multipliers taken from $\left\{0,1\right\}$, with the sum of indicators associated with a fixed *m* equal to 1. Let $\mathcal{V} = \left\{\boldsymbol{s} \mid (18) \text{ and } (19) \text{ hold}, c \geq 0\right\}$, and let $\mathcal{V}^{+} = \left\{\boldsymbol{s} \mid (18) \text{ and } (19) \text{ hold}, c > 0\right\}$. The convex hull of $\mathcal{T}$ is given by $\mathcal{U}$. By the theorem of [12], a finite estimator at the data point $\boldsymbol{s}$ exists if and only if $\boldsymbol{s} \in \operatorname{hull}\left(\mathcal{T}\right)^{\circ}$. In this case, the interior is defined in terms of the topology of the subset of $\Re^{K}$ containing $\boldsymbol{s}$ subject to (18). The proof is complete when one shows that $\mathcal{V}^{+} = \operatorname{hull}\left(\mathcal{T}\right)^{\circ}$. Suppose that $\boldsymbol{s} \in \operatorname{hull}\left(\mathcal{T}\right)^{\circ}$. Let $\boldsymbol{\alpha}$ achieve the maximum as described in the statement of the corrolary. Choose *m* and *j* such that α_{mj} is as small as any component of $\boldsymbol{\alpha}$. There exists $j' \neq j$, $j' \leq J_m$, such that $\alpha_{mj'} \geq 1/J_m$. Let $\boldsymbol{\gamma}$ be the vector configured like $\boldsymbol{\alpha}$, consisting of all zeros except for 1 in the m, j place and -1 in the m, j' place. There exists $\epsilon > 0$ such that $\boldsymbol{s} - \epsilon\boldsymbol{X\gamma} \in \mathcal{V}$. Hence $\boldsymbol{\alpha}$ may be chosen so that all components of $\alpha - \epsilon\boldsymbol{\gamma}$ are non-negative, and, in particular, $\alpha_{mj} \geq \epsilon > 0$. Hence $\boldsymbol{s} \in \mathcal{V}^{+}$.

Now take $\boldsymbol{s} \in \mathcal{V}^{+}$, let $\boldsymbol{\alpha}$ and *c* be as defined in (19), and choose any vector $\boldsymbol{v}$. Since $\boldsymbol{X}$ is of full rank, $\boldsymbol{v}$ is expressible as $\boldsymbol{v} = \boldsymbol{X\gamma}$, for $\boldsymbol{\beta} = \left(\gamma_{11},\cdots,\gamma_{1J_1},\cdots,\gamma_{M1},\cdots,\gamma_{MJ_M}\right)^{\mathrm{T}}$. Since $\boldsymbol{x}_{mJ_m} = \left(0,\cdots,0\right)^{\mathrm{T}}$ for all *m*, $\boldsymbol{\gamma}$ can be selected such that $\sum_{j=1}^{J_m}\gamma_{mj} = 1$ for all *m*. Let $\epsilon = c/\left(2\max\left|\gamma_{mj}\right|\right)$. Then $\boldsymbol{s} - \epsilon\boldsymbol{v} \in \operatorname{hull}\left(\mathcal{T}\right)$. □

One may determine whether such a *c* exists by maximizing *c* over non-negative *c* and $\alpha_{11},\cdots,\alpha_{MJ_M}$ satisfying (18) and (19), and checking to see if the maximum is greater than zero. The above maximization may be done via the simplex algorithm. If $\boldsymbol{s} \in \mathcal{T}$, then (18) and (19) are satisfied by the vector $\boldsymbol{\alpha}$ with zeros in each component except that corresponding to $\boldsymbol{s}$; this realization makes optimization of c more efficient. If the optimization indicates that such a positive c exists, the maximizer of (2) may be determined using Fisher scoring.

The second corollary follows directly from Theorem 1.

Corollary 2. *Suppose that the random vector* $\boldsymbol{Y}$ *arises from the multinomial regression model* (1) *and* (2). *Use* (7) *to construct the implied conditional logistic regression model, and Theorem* 1 *to reduce the model to one with finite maximum likelihood estimates. Then standard asymptotic methods for conditional inference on model parameters of interest, including normal theory techniques and those of* [1], *can be used for testing and estimation*.

If either Corollary 1 or Corollary 2 indicates that finite maximum likelihood estimators do not exist, one might look for estimators in the extended real numbers $\overline{\Re} = \{-\infty, \infty\} \cup \Re$. In certain highly--structured cases, for example, when the sample space for certain stratified rank--based tests is embedded in a canonical exponential family [13], unique maximizers in $\overline{\Re}^K$ can be show to exist. In general, unique estimators are known to exist in the extended real number only in the case when $K = 1$, since the profile likelihood obtained by setting one of the parameters to $\pm\infty$ will correspond to the likelihood of a similar model, with one parameter, and one or more observations, deleted. This reduced regression model need not be of full rank, and so the profile likelihood need not have a unique maximizer. The second example, in Section 3, involves a sample space containing points for which (conditional) maximum likelihood estimates cannot be extended unambiguously, even if allowing infinite values for some components.

[9] motivates the penalized likelihood approach for estimation in order to reduce biases of estimators. Since the standard approach to estimation in this case allows for infinite estimators, expectations of the conventional estimators do not exist, and so bias is an inappropriate criterion for our estimator. In what follows, the median bias (that is, the difference between the median of the sampling distribution of the estimator, minus the true value of the parameter) is used to assess quality of estimation.

3. Results

The following date reflects the results of a randomized clinical trial testing the effectiveness of a screening procedure designed to reduce hepatitis transmission in blood transfusions [9] [14]. The clinical trial was divided into two time periods. The data may be summarized as in **Table 1** [9]. Hepatitis outcome is modeled as a function of period ($S = 0$ for early, and $S = 1$ for late), screening treatment ($T = 0$ for standard, and $T = 1$ for the new method), and the interaction between period and treatment ($S \times T$). The model also contains an intercept term (I). We treat the response as unordered categories.

A log linear model is used, implying a comparison between each of the two hepatitis categories and the third no-disease baseline category. Each of these comparisons involves parameters for I, S, T, and $S \times T$, for a total of 8 parameters. This model is saturated, in that there are as many parameters as there are potential table probabilities. In our case, $M = 4588$, $J_m = 3$ for all m, and $q = 8$.

The lack of any Hepatitis C cases among the treated individuals in the early period gives rise to infinite estimate for the T and $S \times T$ parameters in the comparison between Hepatitis C and healthy subjects.

In this simple case, closed-form maximizers of (2) under the alternative hypothesis exists, since the alternative hypothesis may be viewed as a saturated model for a $2 \times 2 \times 3$ table. No closed-form maximizers of (2) exist under the null hypothesis, since this model is equivalent to the model with no three-way interactions.

The sampling distribution of the sufficient statistics is available in under the hypothesis that all six S, T, and

Table 1. Hepatitis data.

Group	Hepatitis Outcome: Response Variable		
	C	Non-ABC	No disease
Time 0 Treated	0	2	400
Time 0 Untreated	5	3	389
Time 1 Treated	3	10	1896
Time 1 Untreated	5	11	1864

$S\times T$ coefficients are zero. Exact enumeration techniques may be used to enumerate all possible tables 4×3 tables, and their probabilities [15], under the assumption that treatment and time combinations are independent of disease status. There are 2,046,240 such tables.

Inference on the two $S\times T$ parameters may be performed by conditioning the distribution of the associated sufficient statistics for the interaction terms on sufficient statistics associated with the T, S, and I parameters. This conditioning is equivalent to conditioning on Hepatitis C and non-ABC Hepatitis totals for each treatment group, and each time period, separately. After performing this conditioning, 24 tables remain, with 6 distinct values for the effect of the $S\times T$ interaction on Hepatitis C, and 4 distinct values for the effect of the $S\times T$ interaction on non-ABC hepatitis. This yields six tables for inference on the interaction effect on Hepatitis C, and four for inference on the interaction effect on non-ABC hepatitis. **Figure 1** shows the median bias for estimation of the effect on Hepatitis C, indicating a small advantage for the uncorrected estimates. **Figure 2** shows coverage of penalized likelihood asymptotic confidence intervals (17), and exact intervals (9), using (10). In this case, exact intervals are readily available, and have far better coverage properties than the asymptotic intervals; in particular, note that the asymptotic intervals have zero coverage for sufficiently large absolute values of the parameter of interest. Values presented in **Figure 1** are summarized in **Table 2**, and values presented in **Figure 1** are summarized in **Table 3**. The range presented in **Table 2** are heavily dependent on the range of parameter values examined in **Figure 1**, and are intended only for comparison among the two methods.

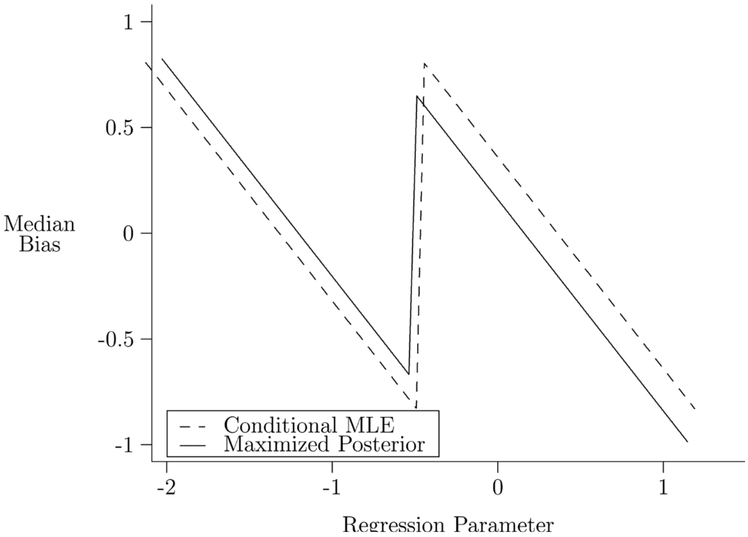

Figure 1. Median bias.

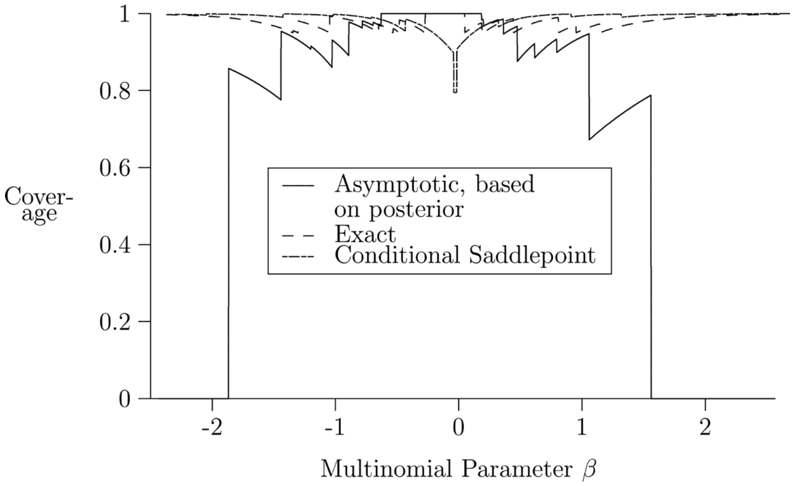

Figure 2. Coverage for various two-sided confidence interval procedures.

A second example concerns polling data related to British general elections [16]. The data set investigated here represents a subset of voters in one of eight geographic areas (London North, London South, Greater Manchester, Merseyside, South Yorkshire, Tyne and Wear, West Midlands, and West Yorkshire), and includes survey respondents who provided their ages and an informative response to a question measuring age at which education was finished (coded as 1 = 15 or younger, 2 = 16, 3 = 17, 4 = 18, 5 = 19 or older), and reported a party of preference, and party of intended vote in the 2005 election. Three parties (Conservative, Labor, and Liberal Democrat) were reported as answers to these last two questions. The resulting data set included 67 individuals. We model voting choice as a function of usual party preference and education, treated as an ordinal variable. In this case, $K = 8$, with covariates representing the effects of Labor and Liberal Democratic Party membership, and education, on the propensity to vote for Labor, and the same effects on the propensity to vote for the Liberal Democrats, plus intercept terms for Labor and Liberal Democrats. Membership in the Conservative Party, and choice of the Conservative Candidate, are taken as baseline. We explore the effect of education on the propensity to vote for the Labor candidates.

The null distribution of this data set cannot be trivially expressed as the independence distribution for a contingency table. One might enumerate the conditional sample space for the sufficient statistics vectors of (3), and the associated conditional probabilities [17]; this calculation, however, took over 16 hours to complete when coded in FORTRAN 90 and run on a 2.6 GHz processor with a 1 GB cache and 8 GB of memory, and is too intensive for routine use. This condition distribution is tabulated in **Table 4**. **Figure 3** shows the median bias for estimates [8] and the proposed method, explicitly recognizing infinite estimates associated with the extreme points in the conditional sample space. While the median bias for both estimators is poor, the corrected estimator generally performed better than the uncorrected estimator, as reported by other authors.

This manuscript is primarily concerned with producing confidence intervals. One might use the asymptotic intervals calculated using the penalized likelihood (17), or (9), with probabilities calculated exactly using **Table 4** in conjunction with (8) and the nominal level adjusted using (10), or (9), with cumulative probabilities approximated (in the present case using a double saddlepoint conditional distribution function approximation of [2]). **Table 5** shows the resulting confidence intervals, and **Figure 4** shows the coverage of these intervals. As

Table 2. Median bias in two estimation methods for hepatitis data.

	Maximized Posterior	Conditional MLE
Minimum	−0.986	$-\infty$
Maximum	0.967	∞

Table 3. Minimal coverage for two confidence interval methods for the hepatitis data.

	Asymptotic, based on posterior	Exact
Minimum	0	0.950

Table 4. Probabilities $Q_{0,1}(t)$ of sufficient statistic associating education with labor party voting, conditional on sufficient statistics for other variables in model.

Sufficient Statistic	Number of Corresponding	Conditional
Value	Response Vectors	Probability
73	8673	0.03214
74	9009	0.03339
75	1335	0.00495
76	26,208	0.09712
77	62,412	0.23129
78	22,440	0.08316
79	15,120	0.05603
80	65,484	0.24268
81	59,160	0.21924

Table 5. Two-sided confidence intervals for the effect of education on voting for labor candidate.

Sufficient	Interval Type		
Statistic Value	(17)	Exact	Saddlepoint
73	(−1.934, 0.118)	(−∞, −0.086)	(−∞, −0.242)
74	(−1.508, 0.134)	(−2.996, 0.052)	(−8.668, 0.122)
75	(−1.264, 0.204)	(−1.478, 0.068)	(−3.420, 0.400)
76	(−1.092, 0.296)	(−1.436, 0.310)	(−2.246, 0.712)
77	(−0.958, 0.408)	(−1.180, 0.674)	(−1.624, 1.118)
78	(−0.848, 0.548)	(−0.784, 0.816)	(−1.216, 1.712)
79	(−0.760, 0.728)	(−0.666, 1.010)	(−0.910, 2.850)
80	(−0.694, 0.990)	(−0.616, 3.056)	(−0.632, 8.042)
81	(−0.696, 1.492)	(−0.406, ∞)	(−0.216, ∞)

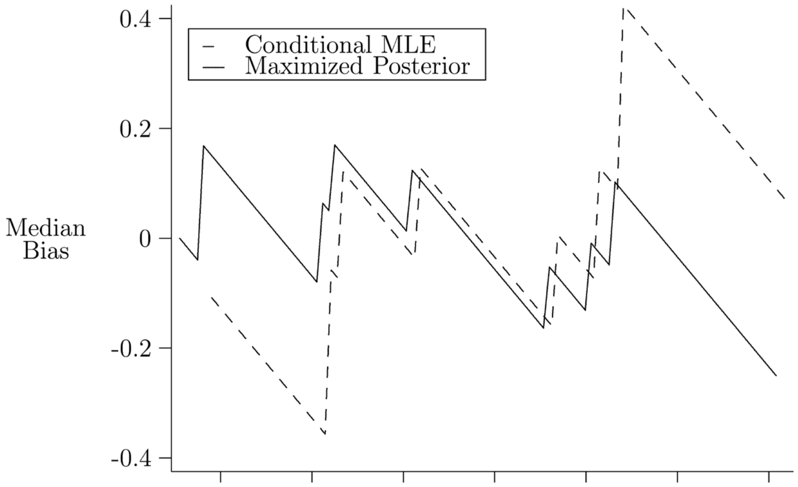

Figure 3. Median bias.

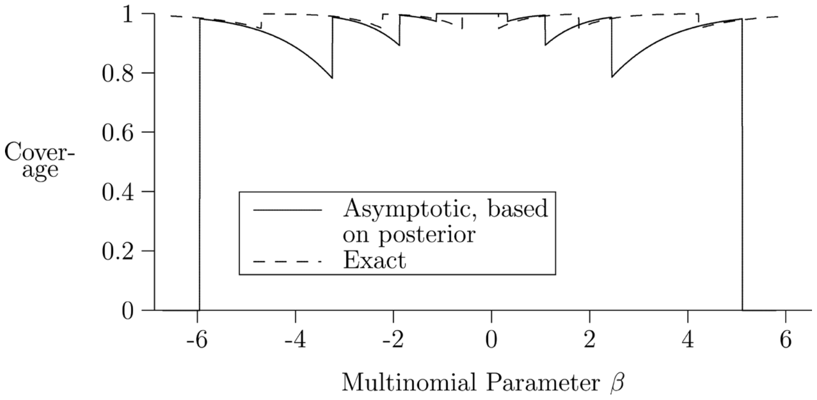

Figure 4. Coverage for various two-sided confidence interval procedures.

noted above, coverage for the asymptotic interval is zero for β_1 below the lowest possible confidence interval endpoint, and above the highest possible confidence interval endpoint; since each of these intervals is finite, and since there are only a finite number of these intervals, coverage is zero outside of a bounded interval for the penalized likelihood intervals (17). Coverage for the asymptotic saddlepoint interval is mostly quite good, except for one area in the middle. This poor performance near $\beta_1 = 0$ can be attributed to the fact that for $t_1 = 73$, the saddlepoint confidence interval is shifted to the left relative to the exact interval, and so the saddlepoint interval

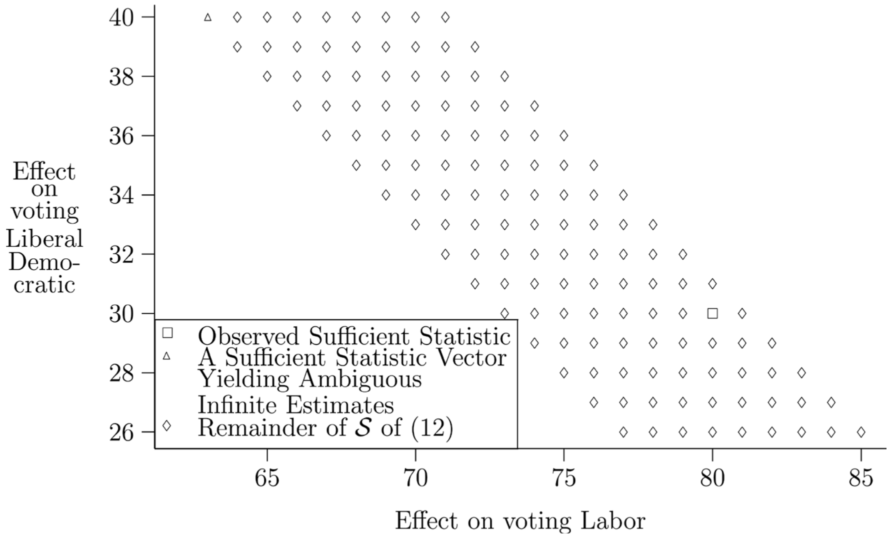

Figure 5. Conditional sample space for education effects.

fails to cover some values of β_1 that the exact interval covers. See the first row in **Table 5**. Furthermore, this most extreme value of t_1 has a considerable amount of conditional probability attached to it; see **Table 4**.

In practice, **Table 4**, and hence the exact intervals in **Table 5** and **Figure 4**, are not computationally feasible. The saddlepoint confidence intervals are computationally feasible, but when the entire sufficient statistic vector lies on the boundary of the convex hull of the sufficient statistic sample space, the methods of Corollary 2 are required to apply these methods.

One can also consider simultaneous inference on both education parameters. **Figure 5** displays the conditional sample space for these the sufficient statistic vectors for the effects on education on preference for Labor and Liberal Democratic Candidates. Corollary 1 indicates that the observed sufficient statistic vector (indicated by □ in the figure) corresponds to finite estimates. Note that the point indicated by Δ is an example of one corresponding to ambiguous estimates; the parameter associated with the first component is clearly estimated as $-\infty$. The conditional likelihood associated with this sample space, with the first component of the parameter set to $-\infty$, corresponds to the sampling distribution consisting of only the point Δ; this space is degenerate, and any value for the second parameter is equally preferred. The point at the opposite corner of the parallelogram in **Figure 5** exhibits similar behavior. Values presented in **Figure 4** are summarized in **Table 2**, and values presented in **Figure 5** are summarized in **Table 3**. The range presented in **Table 2** are heavily dependent on the range of parameter values examined in **Figure 4**, and are intended only for comparison among the two methods.

4. Conclusion

This paper presents an algorithm for converting a multinomial regression problem that features nuisance parameters estimated at infinity to a similar problem in which all nuisance parameters have finite estimates; this conversion is such that the distribution of a sufficient statistic associated with the parameter of interest, conditional on all other sufficient statistics, remains unchanged. These conditional probabilities in the reduced model may be approximated using standard asymptotic techniques to yield confidence intervals with coverage behavior superior to those that arise from, for example, asymptotics derived from the likelihood after penalizing using Jeffreys' prior.

Acknowledgements

This research was supported in part by NSF grant DMS 0906569.

References

[1] Davison, A.C. (1988) Approximate Conditional Inference in Generalized Linear Models. *Journal of the Royal Statistical Society*, **50**, 445-461.

[2] Skovgaard, I. (1987) Saddlepoint Expansions for Conditional Distributions. *Journal of Applied Probability*, **24**, 875-887. http://dx.doi.org/10.2307/3214212

[3] Kolassa, J.E. (1997) Infinite Parameter Estimates in Logistic Regression, with Application to Approximate Conditional Inference. *Scandinavian Journal of Statistics*, **24**, 523-530. http://dx.doi.org/10.1111/1467-9469.00078

[4] Albert, A. and Anderson, J.A. (1984) On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**, 1-10. http://dx.doi.org/10.1093/biomet/71.1.1

[5] Jacobsen, M. (1989) Existence and Unicity of Miles in Discrete Exponential Family Distributions. *Scandinavian Journal of Statistics*, **16**, 335-349.

[6] Clarkson, D.B. and Jennrich, R.I. (1991) Computing Extended Maximum Likelihood Estimates for Linear Parameter Models. *Journal of the Royal Statistical Society*, **53**, 417-426.

[7] Santner, T.J. and Duffy, D.E. (1986) A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **73**, 755-758. http://dx.doi.org/10.1093/biomet/73.3.755

[8] Firth, D. (1993) Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, **80**, 27-38. http://dx.doi.org/10.1093/biomet/80.1.27

[9] Bull, S.B., Mak, C. and Greenwood, C.M. (2002) A Modified Score Function Estimator for Multinomial Logistic Regression in Small Samples. *Computational Statistics and Data Analysis*, **39**, 57-74. http://dx.doi.org/10.1016/S0167-9473(01)00048-2

[10] Jeffreys, H. (1961) Theory of Probability. 3rd Edition, Clarendon Press, Oxford.

[11] Heinze, G. and Schemper, M. (2001) A Solution to the Problem of Monotone Likelihood in Cox Regression. *Biometrics*, **57**, 114-119. http://dx.doi.org/10.1111/j.0006-341X.2001.00114.x

[12] Barndorff-Nielsen, O.E. (1978) Information and Exponential Families in Statistical Theory. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

[13] Robinson, J. and Samonenko, I. (2012) Personal Communication.

[14] Blajchman, M., Bull, S. and Feinman, S., C.P.-T.H.P.S. (1995) Group Post-Transfusion Hepatitis: Impact of Non-a, Non-b Hepatitis Surrogate Tests. *The Lancet*, **345**, 21-25. http://dx.doi.org/10.1016/S0140-6736(95)91153-7

[15] Pagano, M. and Halvorsen, K.T. (1981) An Algorithm for Finding the Exact Significance Levels of r × c Contingency Tables. *Journal of the American Statistical Association*, **76**, 931-934. http://dx.doi.org/10.2307/2287590

[16] Sanders, D.J., Whiteley, P.F., Clarke, H.D., Stewart, M. and Winters, K. (2007) The British Election Study. University of Essex.

[17] Hirji, K.F. (1992) Computing Exact Distributions for Polytomous Response Data. *Journal of the American Statistical Association*, **87**, 487-492. http://dx.doi.org/10.1080/01621459.1992.10475230