Statistica Sinica **8**(1998), 941-963

BIAS CORRECTION FOR CENSORED DATA WITH EXPONENTIAL LIFETIMES

A. N. Pettitt, J. M. Kelly and J. T. Gao

Queensland University of Technology

Abstract: The analysis of censored data presents several problems including infinite maximum likelihood estimates and biased estimates. In this paper we consider modifying the score equation for the maximum likelihood estimate so that the bias is reduced, following Firth (1993). This method is considered for the case of right censored failure time data having an exponential distribution and the means of the observations are given by a log-linear model. For some situations the modified score equations can be integrated and the method is equivalent to a penalised maximum likelihood approach. We additionally show that the estimates are finite under weak conditions. A small sample study indicates that the modified estimates have good properties and have mean square error behaving like $1/n$.

Key words and phrases: Bias reduction, censored data, exponential model, failure data, penalised likelihood, regression model, Type I censoring.

1. Introduction

Several problems arise with the analysis of censored data including infinite maximum likelihood estimates and biased estimates. An additional problem concerns the degrees of freedom associated with censored observations: in one extreme case observations can be practically completely missing while in the other they can be almost uncensored. Here we follow an approach which is to modify the maximum likelihood estimating equations so that the bias is reduced and in some cases finite estimates are guaranteed. For some cases the modified estimating equations can be used to obtain the estimates as maximisers of a penalised likelihood. We extend the ideas of Firth (1993) to investigate the adjustment to the likelihood score which reduces the bias of resulting estimates in a sampling theory framework. When it exists the penalty function can be interpreted as a Bayesian prior or, in a sampling framework, as available prior data.

A motivation for penalising the likelihood might be to ensure finite estimates. There is considerable literature on this topic. For example, it is well known that infinite likelihood estimates can occur with discrete, censored and generally incomplete data (see, for example, Clarkson and Jennrich (1991), Geyer and

Thompson (1992), Haberman (1974), Hamada and Tse (1988), Silvapulle and Burridge (1986), Verbeek (1989)). (By an "infinite maximum likelihood estimate", it is meant in this paper that the likelihood takes a supremum by letting a parameter tend to plus or minus infinity. Some authors refer to this situation by saying that the maximum likelihood estimate does not exist but Verbeek (1989) considers the compact parameter space $[-\infty, \infty]^K$ allowing for formal use of the expression "infinite maximum likelihood estimate".) One approach to the problem involves the idea of "extended maximum likelihood estimation" due to Haberman (1974) and more recently Clarkson and Jennrich (1991). The latter authors' idea is to provide a method by which finite maximum likelihood estimates can be found from a subset of the cases in the sample resulting then, inevitably, in aliasing of effects or non-identifiability problems. Some parameters have infinite maximum likelihood estimates and for these parameters no attempt is made to make any inferences. Pettitt (1996) illustrates some practical problems of interpretation and inference which arise if the ideas of extended maximum likelihood estimation are applied to the analysis of fractional factorial experiments with incomplete data. If one is fitting a sequence of nested models it is then disconcerting that parameter estimates which are finite for smaller models become infinite for larger models.

Specific results are given in this paper for right censored data modelled by an exponential distribution. For estimation of the mean, it is found that the bias reducing penalty function for censored identically distributed variables is equivalent to knowing that one observation is less than a common censoring point (whereas the information from a right censored observation is knowing that it is greater than the censoring point). For the general log-linear model, an estimating equation is found and it does not appear possible to define a corresponding penalty function.

The results of a small simulation study are reported for the log-linear model. Two estimates are considered: the maximum likelihood estimator and the new bias corrected estimate. Generally it was found that the modified estimate had substantially smaller bias and variance than the maximum likelihood estimate when the latter existed. For those situations where there was a substantial chance of obtaining an infinite maximum likelihood estimate, the modified estimate had good bias properties irrespective of the existence or otherwise of the maximum likelihood estimate.

We give an example of how the modified estimate can be used and suggest the bootstrap can be used to find confidence intervals for the parameters of interest.

In Section 2 we derive the appropriate penalty function for right censored data and Section 3 the results are extended to the log-linear regression model.

In Section 4 simulation results are given and Section 5 gives an example. Finally appendices give mathematical results on the derivation of the estimates and prove finiteness.

2. Adjustments for Right Censored Data

The basic idea in Firth (1993) is to modify the score equation $U(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = 0$, where $\ell(\theta)$ is the log likelihood to give

$$U^*(\theta) = U(\theta) + A(\theta), \tag{1}$$

where $A(\theta)$ is chosen to remove the bias of $O(1/n)$ of the maximum likelihood estimate as determined by expanions in $1/n$. Firth gives an expression for $A(\theta)$ involving expected values of derivatives of the log likelihood (his equation (4.1)). Additionally, $A(\theta)$ can be calculated using expected or observed information. From equation (1), it may be possible to integrate with respect to θ to obtain a modified log likelihood and therefore modified likelihood or penalised likelihood function $L^*(\theta) = L(\theta)\,p(\theta)$, where $p(\theta)$ is the penalty function. For θ being the canonical parameter of an exponential family distribution, the method leads to the penalty function being equal to $\|i(\theta)\|^{1/2}$, where $i(\theta)$ is the expected information matrix, which is identical to Jeffreys' invariant prior. Because the method is based upon bias reduction, the penalty function is not invariant to non-linear reparameterisation as a Bayesian prior would be.

Let us suppose we make observations upon independent $Y_1, \ldots, Y_n$, each of which is subject to right censoring at $y = c$, say, that is Type I censoring. We assume also that each Y has an exponential distribution with mean μ. Conditional on knowing which observations are censored and which are uncensored, the likelihood is given by

$$L_c(\mu) = \Big\{ \prod_{j:uncen} \frac{e^{-y_j/\mu}}{\mu} \Big\} \Big\{ \prod_{j:cen} e^{-c/\mu} \Big\}, \tag{2}$$

and this would normally be the likelihood that would be considered (see, for example, Cox and Oakes (1984)). However this formulation does not give a sampling distribution for each Y suitable for bias calculations. Only the uncensored observations have a distribution. We therefore consider a distribution for each Y which is unconditional on knowing whether censoring has taken place, and this is given by

$$dF(y) = \begin{cases} \dfrac{e^{-y/\mu}}{\mu}\, dy & 0 < y < c \\ e^{-y/\mu} & y = c, \end{cases} \tag{3}$$

giving $L(\mu) = \prod_{j=1}^{n} dF(y_j)$, which simplifies to (2) when the censoring information is known.

Under this distribution the mean of Y is $\mu\{1-e^{-c/\mu}\}$ rather than μ when no censoring is considered. Such distributions are considered in estimating equation approaches to censored data (see, for example, Buckley and James (1979)).

We note that the maximum likelihood estimate derived from equation (2) is $\hat{\mu} = \sum_{i=1}^{n} y_i/n_u$, where n_u is the number of uncensored observations and when $n_u \geq 1$. When $n_u = 0$ then $\hat{\mu} = \infty$. The event $n_u = 0$ occurs with probability $e^{-nc/\mu}$ under model (3). A consistent estimate, $\tilde{\mu}$, of μ is given by solving the estimating equation (which only implicitly uses the censoring information)

$$\sum_{i=1}^{n} \Big[y_i - \mu\Big\{1 - e^{-c/\mu}\Big\}\Big] = 0$$

or $\bar{y} = \mu\{1 - e^{-c/\mu}\}$.

Provided $\bar{y} < c$ the solution $\tilde{\mu}$ is finite but for $\bar{y} \to c$, $\tilde{\mu} \to \infty$. Although the probability of obtaining infinite estimates $\hat{\mu}$ and $\tilde{\mu}$ is $e^{-nc/\mu}$, which becomes negligible for reasonable sample sizes and obviously goes to zero for $n \to \infty$, there is a considerable chance of obtaining infinite estimates if one were considering several small samples.

The bias reduction idea can be first applied to the parameter μ. Routine calculations then give the modified score function

$$U^*(\mu) = U(\mu) - \frac{c}{\mu^2\{e^{c/\mu} - 1\}} \tag{4}$$

using the expected information. Integration of (4) then gives

$$L^*(\mu) = L(\mu)\Big\{1 - e^{-c/\mu}\Big\}. \tag{5}$$

The penalty function $1 - e^{-c/\mu}$ corresponds to the information that one observation is known to be in the interval $(0, c)$, whereas knowing an observation is right censored is equivalent to knowing that it lies in (c, ∞).

If we were to interpret the function $1 - e^{-c/\mu}$ as a prior density for μ then this prior is improper since as $\mu \to \infty$ the function behaves like $1/\mu$. But the function does have interpretation as providing prior sample information, in the sense, as stated above, that $1-e^{-c/\mu}$ is the likelihood of obtaining an observation in $(0, c)$. So an alternative way of interpreting Firth's idea within the sampling context is that of thinking of the penalty function (when it exists) introducing prior imaginary sample information with a corresponding likelihood.

The modified likelihood $L^*(\mu)$, equation (5), can be seen, when all observations are censored (the situation that gives rise to $\hat{\mu} = \infty$) to be equivalent to the likelihood for $n + 1$ binary observation Z with $\text{pr}(Z = 0) = e^{-c/\mu}$,

$\text{pr}(Z = 1) = 1 - e^{-c/\mu}$, and there is one 0 observation and n observations equal to 1.

Following earlier work on extended maximum likelihood, for example Clarkson and Jennrich (1991), it is straightforward to show $L^*(\mu)$ takes its maximum for $\mu < \infty$.

To demonstrate the dependence on the parameterisation consider now $\theta = -\log\mu$, a natural parameterisation for extension to incorporate covariates.

Straightforward application of Firth's method to this model now gives

$$U^*(\theta) = U(\theta) - \frac{1}{2} + \frac{ce^\theta}{e^{ce^\theta} - 1}.$$

Integrating this and exponentiating then gives the modified likelihood as $L^*(\theta) = L(\theta)e^{-\theta/2}(1 - e^{-ce^\theta})$ with penalty function

$$p(\theta) = e^{-\theta/2}(1 - e^{-ce^\theta}). \tag{6}$$

The penalty function $p(\theta)$, equation (6), is equivalent to the additional information that one observation is known to be in the interval $(0, c)$ and the measure $e^{-\theta/2}$ for $-\infty < \theta < \infty$ which has no direct sample interpretation.

The likelihood $L(\theta)$ achieves a supremum as $\theta \to -\infty$ when all cases are right censored so that the behaviour of $p(\theta)$ as $\theta \to -\infty$ determines whether $L^*(\theta)$ can take a supremum as $\theta \to -\infty$. Now, as $\theta \to -\infty$, $p(\theta) \sim e^{\theta/2}$ so that, as $L(\theta)$ is bounded above, $L^*(\theta) \to 0$ as $\theta \to -\infty$ and achieves a finite maximum for all possible data. The penalty function $p(\theta)$ is bounded above and tends to 0 for $\|\theta\| \to \infty$.

When $p(\theta)$ is interpreted as an unnormalised prior density for θ we find the normalising constant is given by

$$\int e^{-\theta/2}\left(1 - e^{-ce^\theta}\right) d\theta = 2\Big\{e^{ce^x - x/2} - e^{-x/2} + \frac{1}{2}\pi c\phi(\sqrt{2ce^x})\Big\},$$

where ϕ is the standard normal density. This gives the normalised density as $p(\theta)/(2\pi c^{1/2})$.

3. Estimation and Extension for Log Linear Models

3.1. Regression model

Here we consider the more practically important and more interesting log-linear model for μ_i with $\log\mu_i = -x_i^T\beta$. Here β has p components, $\beta_1, \ldots, \beta_p$.

Now the score for β_r $(r = 1, \ldots, p)$ is given by

$$U_r = \sum_i \Big(1 - \delta_i - \frac{y_i}{\mu_i}\Big)x_{ir} \tag{7}$$

with δ_i defined to be 0 if $0 < y < c$ and 1 if $y = c$; also $\log \mu_i = -x_i^T \beta$. Using expected information, the modified score U_r^* is given by $U_r^* = U_r + A_r^E$ where

$$A_r^E = \frac{1}{n} \sum_i \sum_{u,v} \kappa^{u,v} x_{ir} x_{iu} x_{iv} \Big\{ \frac{c}{\mu_i e^{c/\mu_i}} - \frac{1}{2}(1 - e^{-c/\mu_i}) \Big\}, \tag{8}$$

and $\kappa^{u,v}$ is the inverse of the expected information matrix $\kappa_{u,v}$ which is given by $1/n(X^T W X)$ in matrix notation. Here W is the diagonal matrix with diagonal term $(1 - e^{-c/\mu_i})$ and X is the n by p model matrix $\{x_{ir}\}$. Using ideas similar to Firth (1993), Section 3, let h_i be the ith diagonal element of the (expected) 'hat' matrix $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$; then A_r^E can be written as (see Appendix for more details)

$$A_r^E = \sum_i h_i \Big\{ \frac{c}{\mu_i (e^{c/\mu_i} - 1)} - \frac{1}{2} \Big\} x_{ir}, \tag{9}$$

so that U_r^* is given by

$$U_r^* = \sum_i \Big[(1 - \delta_i - \frac{h_i}{2}) - \frac{1}{\mu_i} \{ y_i - \frac{c\, h_i}{e^{c/\mu_i} - 1} \} \Big] x_{ir}.$$

Comparing U_r with U_r^*, y_i is adjusted by the amount

$$\frac{c\, h_i}{e^{c/\mu_i} - 1}$$

and δ_i by the amount $h_i/2$.

As is well known, the score equations $U_r = 0$, $r = 1, \ldots, p$ can be solved by iteratively reweighted least squares by taking adjusted dependent variable

$$\frac{(1 - \delta_i - y_i/\mu_i)}{(1 - e^{-c/\mu_i})}$$

and weight matrix W with diagonal element $1 - e^{-c/\mu_i}$.

The equation $U_r^* = 0$, $r = 1, \ldots, p$, can be solved for β by applying Newton-Raphson but h_i is assumed to be constant, independent of β. Details are given later in Section 4.1.

3.2. Finite estimates

Two specific cases of the log-linear model can be investigated simply. These are (i) the situation where all observations have the same mean and dealt with earlier, and (ii) the situation where all observations have different means or, equivalently, $p = n$ and the model matrix X is of full rank, equal to n. For the latter case where all observations have different means $-\log \mu_j = \theta_j$ the penalty function $p(\theta)$ is the product of functions given by equation (6), namely

$$p(\theta) = \prod_{j=1}^{n} e^{-\theta_j/2}(1 - e^{-ce_j^{\theta}}).$$

Since the deviation of the adjustment is based on bias and bias is invariant under linear transformation, β^* is finite if and only if the θ_j^* are finite.

There remains the task of demonstrating that the estimate β^* is finite always for $1 < p < n$. This is considered in detail in Appendix 2. However some preliminary remarks are given here. Consider the penalised likelihood $L(\beta)\,p(\beta)$. Then if $\log p(\beta)$ is strictly concave and unbounded below as $|\beta| \to \infty$, and $\ell(\beta)$ is strictly concave and bounded above then it follows that the maximum penalised likelihood estimate β^* exists and is unique. Now $p(\beta)$ is strictly concave if

$$\frac{\partial^2}{\partial\beta'\partial\beta} \log p(\beta) < 0$$

for all possible β. If $p(\beta)$ exists then

$$\frac{\partial^2}{\partial\beta_s\partial\beta_r} \log p(\beta) = \frac{\partial}{\partial\beta_S} A_r^E$$

and this matrix is symmetric.

Upon differentiation of A_r^E with respect to β_S we obtain a quantity such that, in general,

$$\frac{\partial A_r^E}{\partial\beta_s} \neq \frac{\partial A_s^E}{\partial\beta_r}. \tag{10}$$

The inequality (10) then implies that there is no function $p(\beta)$ which satisfies

$$A_r^E = \frac{\partial}{\partial\beta_r} \log p(\beta),\ r = 1, \ldots, p.$$

The situation is analogous to the non-existence of the quasi-likelihood function (see McCullagh and Nelder (1989), Section 9.3.2, for example). Further details are considered in the Appendix 2 where finiteness is proved.

4. Computational Details and Small Sample Study

4.1. Introduction

We carried out a small simulation study to investigate the small sample properties of the modified estimate. The theoretical development is based on asymptotic series expansions in $1/n$ and needs to be validated for small samples. Additionally our theoretical developments provide no information on the variance of the estimates. Two simple log-linear models were investigated. These were

(i) regression model, $\log \mu_i = \beta_1 + \beta_2 x_i$ with $x_i = \frac{2i-2n-1}{2n-1}$, $i = 1, \ldots, n$

(ii) factorial model, $\log \mu_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i}$, with $x_2 = (0,0,1,1)$, $x_3 = (0,1,0,1)$, $i = 1, \ldots, n$.

Observations y_i were generated to have an exponential distribution with mean μ_i and censored at a fixed point e^{-c}. For the regression model n was taken to be 5, 10, 20. For the factorial model n was taken to be 4 and the design replicated twice to produce the $n = 8$ design. The censoring point e^{-c} was chosen to take the values 0.01, 0.1, 0.2 and 0.5 for all choices of n and model. The true value of (β_1, β_2) was (0, 1) for the regression model and that of $(\beta_1, \beta_2, \beta_3)$ for the factorial model was (0, 0, 0).

For each generated sample it was possible that the maximum likelihood estimate could be infinite and this possibility was tested for by using the technique of Hamada and Tse (1989) implemented in their FORTRAN code. In the case of an infinite maximum likelihood estimate, the modified estimate was found and the cases carefully investigated to check convergence of the modified estimate. The two estimates can be compared for bias and variance when the maximum likelihood estimate is finite, otherwise we just consider the bias and variance of the modified estimate.

To find β^* we investigated two schemes. First, we used an iterative scheme based on applying Newton-Raphson to $U_r^* = 0$, as in the standard approach to solving $U_r = 0$ to give the maximum likelihood estimates, with the exception that h_i, the 'hat' matrix element, is assumed to vary slowly with β, that is $\frac{\partial}{\partial \beta} h_i = 0$.

The second approach is to solve $U_r^* = 0$ using a standard root finding algorithm such as the NAG Fortran Library Routine C05NBF (The Numerical Algorithms Group Limited (1993)) which uses an adjustment to the Powell (1970) hybrid method for solving a system of non-linear equations $f_i(x_1, \ldots, x_n) = 0$, $i = 1, \ldots, n$.

4.2. Results

The results of the limited simulation study are given in Tables 1 and 2. Where infinite maximum likelihood estimates have been obtained, the bias, variance and mean square error of the maximum likelihood estimates, $\hat{\beta}$ and modified estimates, β^* are given conditional on the maximum likelihood estimates being finite. However, for β^* we can give the simulation summary statistics for all cases, whether or not the maximum likelihood estimate is finite. (That is, unconditionally on what property the maximum likelihood estimate has.) These estimates are denoted by $\hat{\beta}$(cond'l), β^*(cond'l), and β^*(uncond'l).

	$n = 5$					n	
$\exp(-c)$	pr(∞)		Bias	Var	MSE	pr(∞)	
0.01	0	$\hat{\beta}_1$	0.224	1.285	1.335	0	$\hat{\beta}_1$
		β_1^*	0.002	0.938	0.938		β_1^*
		$\hat{\beta}_2$	0.155	3.769	3.793		$\hat{\beta}_2$
		β_2^*	−0.056	2.492	2.495		β_2^*
0.1	0.009	$\hat{\beta}_{1\,(\text{cond'l})}$	0.198	1.950	1.989	0.0001	$\hat{\beta}_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{cond'l})}$	0.047	1.172	1.174		$\beta^*_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{uncond'l})}$	0.044	1.303	1.305		$\beta^*_{1\,(\text{uncond'l})}$
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.364	6.346	6.478		$\hat{\beta}_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{cond'l})}$	0.020	3.408	3.408		$\beta^*_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{uncond'l})}$	0.059	4.078	4.081		$\beta^*_{2\,(\text{uncond'l})}$
0.2	0.0414	$\hat{\beta}_{1\,(\text{cond'l})}$	0.142	2.648	2.668	0.0001	$\hat{\beta}_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{cond'l})}$	0.066	1.432	1.436		$\beta^*_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{uncond'l})}$	0.036	1.815	1.816		$\beta^*_{1\,(\text{uncond'l})}$
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.403	8.603	8.765		$\hat{\beta}_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{cond'l})}$	0.003	4.220	4.219		$\beta^*_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{uncond'l})}$	0.106	6.028	6.038		$\beta^*_{2\,(\text{uncond'l})}$
0.5	0.2664	$\hat{\beta}_{1\,(\text{cond'l})}$	0.144	3.335	3.356	0.0249	$\hat{\beta}_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{cond'l})}$	0.277	1.538	1.614		$\beta^*_{1\,(\text{cond'l})}$
		$\beta^*_{1\,(\text{uncond'l})}$	0.030	3.094	3.094		$\beta^*_{1\,(\text{uncond'l})}$
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.215	10.434	10.479		$\hat{\beta}_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{cond'l})}$	−0.194	4.567	4.604		$\beta^*_{2\,(\text{cond'l})}$
		$\beta^*_{2\,(\text{uncond'l})}$	0.016	9.952	9.952		$\beta^*_{2\,(\text{uncond'l})}$

Table 1. Bias, variance and MSE for parameter estimates, co

Table 2. Bias, variance and MSE for parameter estimates, conditional and unconditional, for the factorial model.

	$n=4$					$n=8$				
$\exp(-c)$	$\mathrm{pr}(\infty)$		Bias	Var	MSE	$\mathrm{pr}(\infty)$		Bias	Var	MSE
0.01	0.0005	$\hat{\beta}_{1\,(\text{cond'l})}$	0.373	1.226	1.365	0	$\hat{\beta}_1$	0.192	0.496	0.533
		$\beta^*_{1\,(\text{cond'l})}$	0.029	0.821	0.821		β^*_1	0.027	0.421	0.421
		$\beta^*_{1\,(\text{uncond'l})}$	0.028	0.824	0.825					
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.010	1.698	1.698		$\hat{\beta}_2$	−0.002	0.683	0.683
		$\beta^*_{2\,(\text{cond'l})}$	0.009	1.038	1.038		β^*_2	−0.003	0.566	0.566
		$\beta^*_{2\,(\text{uncond'l})}$	0.009	1.041	1.041					
		$\hat{\beta}_{3\,(\text{cond'l})}$	0.006	1.720	1.720		$\hat{\beta}_3$	0.007	0.668	0.668
		$\beta^*_{3\,(\text{cond'l})}$	0.004	1.049	1.049		β^*_3	0.006	0.553	0.553
		$\beta^*_{3\,(\text{uncond'l})}$	0.005	1.052	1.052					
0.1	0.036	$\hat{\beta}_{1\,(\text{cond'l})}$	0.326	1.410	1.516	0.0005	$\hat{\beta}_{1\,(\text{cond'l})}$	0.129	0.612	0.629
		$\beta^*_{1\,(\text{cond'l})}$	0.116	0.829	0.843		$\beta^*_{1\,(\text{cond'l})}$	0.036	0.456	0.458
		$\beta^*_{1\,(\text{uncond'l})}$	0.066	0.985	0.989		$\beta^*_{1\,(\text{uncond'l})}$	0.035	0.460	0.461
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.020	2.033	2.033		$\hat{\beta}_{2\,(\text{cond'l})}$	0.001	0.849	0.849
		$\beta^*_{2\,(\text{cond'l})}$	0.012	1.124	1.124		$\beta^*_{2\,(\text{cond'l})}$	0.000	0.622	0.622
		$\beta^*_{2\,(\text{uncond'l})}$	0.020	1.309	1.310		$\beta^*_{2\,(\text{uncond'l})}$	0.001	0.625	0.625
		$\hat{\beta}_{3\,(\text{cond'l})}$	0.007	2.025	2.025		$\hat{\beta}_{3\,(\text{cond'l})}$	−0.010	0.835	0.835
		$\beta^*_{3\,(\text{cond'l})}$	0.005	1.126	1.126		$\beta^*_{3\,(\text{cond'l})}$	−0.008	0.611	0.611
		$\beta^*_{3\,(\text{uncond'l})}$	0.007	1.306	1.306		$\beta^*_{3\,(\text{uncond'l})}$	−0.008	0.614	0.615
0.2	0.1282	$\hat{\beta}_{1\,(\text{cond'l})}$	0.339	1.485	1.599	0.0068	$\hat{\beta}_{1\,(\text{cond'l})}$	0.072	0.748	0.753
		$\beta^*_{1\,(\text{cond'l})}$	0.206	0.843	0.885		$\beta^*_{1\,(\text{cond'l})}$	0.039	0.523	0.525
		$\beta^*_{1\,(\text{uncond'l})}$	0.062	1.226	1.230		$\beta^*_{1\,(\text{uncond'l})}$	0.029	0.553	0.553
		$\hat{\beta}_{2\,(\text{cond'l})}$	0.016	2.070	2.070		$\hat{\beta}_{2\,(\text{cond'l})}$	0.001	1.002	1.002
		$\beta^*_{2\,(\text{cond'l})}$	0.012	1.133	1.133		$\beta^*_{2\,(\text{cond'l})}$	0.000	0.690	0.690
		$\beta^*_{2\,(\text{uncond'l})}$	0.001	1.651	1.651		$\beta^*_{2\,(\text{uncond'l})}$	0.005	0.718	0.717
		$\hat{\beta}_{3\,(\text{cond'l})}$	0.011	2.096	2.096		$\hat{\beta}_{3\,(\text{cond'l})}$	0.005	1.026	1.026
		$\beta^*_{3\,(\text{cond'l})}$	0.005	1.147	1.147		$\beta^*_{3\,(\text{cond'l})}$	0.004	0.706	0.706
		$\beta^*_{3\,(\text{uncond'l})}$	0.012	1.626	1.626		$\beta^*_{3\,(\text{uncond'l})}$	0.002	0.737	0.736
0.5	0.5586	$\hat{\beta}_{1\,(\text{cond'l})}$	0.659	1.274	1.708	0.1907	$\hat{\beta}_{1\,(\text{cond'l})}$	0.022	0.803	0.804
		$\beta^*_{1\,(\text{cond'l})}$	0.652	0.679	1.104		$\beta^*_{1\,(\text{cond'l})}$	0.160	0.488	0.514
		$\beta^*_{1\,(\text{uncond'l})}$	0.055	1.715	1.718		$\beta^*_{1\,(\text{uncond'l})}$	−0.013	0.824	0.824
		$\hat{\beta}_{2\,(\text{cond'l})}$	−0.015	1.867	1.867		$\hat{\beta}_{2\,(\text{cond'l})}$	0.028	1.223	1.223
		$\beta^*_{2\,(\text{cond'l})}$	−0.008	0.959	0.959		$\beta^*_{2\,(\text{cond'l})}$	0.023	0.752	0.752
		$\beta^*_{2\,(\text{uncond'l})}$	0.002	2.625	2.625		$\beta^*_{2\,(\text{uncond'l})}$	0.005	1.182	1.182
		$\hat{\beta}_{3\,(\text{cond'l})}$	−0.024	1.783	1.784		$\hat{\beta}_{3\,(\text{cond'l})}$	0.004	1.233	1.233
		$\beta^*_{3\,(\text{cond'l})}$	−0.018	0.914	0.914		$\beta^*_{3\,(\text{cond'l})}$	0.003	0.758	0.758
		$\beta^*_{3\,(\text{uncond'l})}$	0.018	2.593	2.593		$\beta^*_{3\,(\text{uncond'l})}$	−0.006	1.208	1.208

The results show for the regression model that for negligible censoring ($e^{-c} = 0.01$), the modified estimates have negligible bias while the maximum likelihood estimates have substantially more bias. Surprisingly, as Firth's (1993) theory predicts otherwise, the variance of the modified estimate is less than that of the maximum likelihood estimate, but this difference decreases, as expected, with increasing n. Generally for larger values of $e^{-c} = 0.1, 0.2$ and 0.5, the modified estimate is better than the maximum likelihood estimate in terms of both bias and variance when samples are restricted to those cases where the maximum likelihood estimate is finite. A curious exception is the case $n = 5$ and $e^{-c} = 0.5$ when there is about a 27 percent probability the maximum likelihood is infinite, the maximum likelihood estimate of the intercept, β_1 has smaller bias than the modified estimate, but unconditionally, the modified estimate has negligible bias, demonstrating the effectiveness of the bias correction. Of course, unconditionally the maximum likelihood estimate has infinite bias and variance. Unexpectedly, when the conditional modified estimate is compared with the unconditional modified estimate it is sometimes observed that the latter has smaller bias than the former, particularly when the probability of infinite estimates is high, whereas the variance of the former is always smaller. This is a finite sample property that the theory appears not to be able to explain, as asymptotically, infinite maximum likelihood estimates occur with zero probability.

Overall, the modified estimate has substantially smaller bias and variance than the maximum likelihood estimate, but this advantage, as predicted by the theory, reduces as the sample size increases. The conclusion is that the modification is worthwhile.

For the factorial model with $n = 4$ the modified estimate is better in terms of bias and variance than the maximum likelihood estimate. For the main effects parameters β_2 and β_3, bias is negligible for the two estimates but the modified estimate has somewhat smaller variance. For the intercept parameter β_1, bias is substantial for both estimates but the modified estimate has smaller bias and variance. For $n = 8$, the situation is similar to that for $n = 4$ but as expected biases and variances are relatively smaller. This is clearly shown in Figure 1 which shows the values of the beta estimates of β_1, β_2, and β_3 obtained under the original and modified score for $n = 8$, and $e^{-c} = 0.5$. Again the exception arises in the case $n = 8$ and $e^{-c} = 0.5$ when the maximum likelihood estimate of the intercept β_1, has smaller bias (but larger variance) than the modified estimate, but the unconditional modified estimate has negligible bias.

Implementation of the root finding procedure to solve $U_r^* = 0$ resulted in only two cases ($n = 10$, $e^{-c} = 0.2$ and 0.5), over all 200 000 simulations where a solution could not be found. These were extreme samples where the final observation was the only uncensored one, and this close to zero. Similar situations arose where initial efforts did not reveal a solution, and each was thoroughly examined to eventually reveal convergence. On incrementing the uncensored observation, thereby producing a new sample each time, and attempting to solve $U_r^* = 0$ for this sample, a value of the uncensored observation was eventually found for which the score could be resolved. By starting the problematic cases at this solution, the root finding procedure converged to a reasonable solution. For the two remaining unresolved cases, the weights became zero (according to machine precision) and iteration could not continue due to division by zero. This was considered to be a problem with machine precision, rather than a problem with the theory. Using weighted least square software, the number of cases unresolved was considerably higher, due to reduced precision. Both approaches otherwise produced comparable estimates.

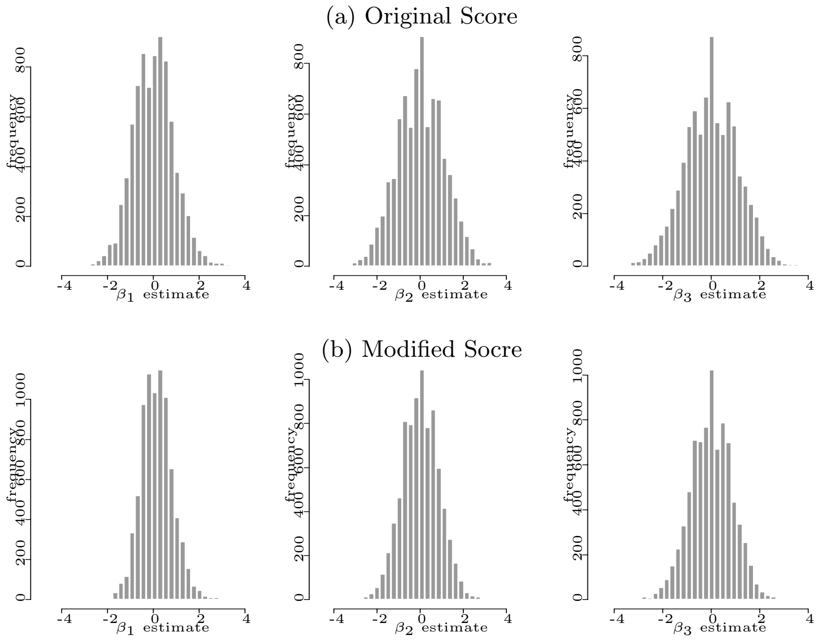

Figure 1. Count of beta estimates from the (a) original score and (b) modified score for $n = 8$ and $e^{-c} = 0.5$.

In general it appears that the modified estimate has good small sample properties and its mean square error behaves like $1/n$, whereas the maximum likelihood estimate depends upon its finiteness for reasonable properties.

5. Illustration

To illustrate the method, we consider the 2^{9-5} fractional design and censoring pattern presented by Hamada and Tse (1992) in their Table 2 (see also Table

3 of this paper). For the purpose of illustrating the techniques presented in this paper, we simulate a sample of size 16 from the unit exponential distribution. Although Hamada and Tse (1992) assume an exponential regression model, we require the exponential assumption to be true in order to compare the properties of our estimators. The censoring pattern is kept since the censoring and design configuration determine the finiteness, or otherwise, of the maximum likelihood estimates (Hamada and Tse (1992)). To obtain the censoring pattern, the largest four observations generated from the unit exponential distribution are associated with the censoring cases (in the order they were generated), and the remainder with the uncensored cases (again in the order they were generated). The censoring point is then the mid-point between the fourth and fifth largest observations, in this case $c = 1.547$. The censored data set can be found in Table 3.

Table 3. Design and data for the example of Hamada and Tse (1992)

	Design									
Data	Censored	A	B	C	D	E	F	G	H	I
1.547	1	1	1	1	1	1	1	1	1	1
0.229	0	1	1	1	−1	1	−1	−1	−1	−1
1.488	0	1	1	−1	1	−1	−1	−1	1	−1
1.547	1	1	1	−1	−1	−1	1	1	−1	1
0.991	0	1	−1	1	1	−1	−1	1	−1	−1
1.547	1	1	−1	1	−1	−1	1	−1	1	1
1.547	1	1	−1	−1	1	1	1	−1	−1	1
1.491	0	1	−1	−1	−1	1	−1	1	1	−1
0.129	0	−1	1	1	1	−1	1	−1	−1	−1
0.838	0	−1	1	1	−1	−1	−1	1	1	1
0.553	0	−1	1	−1	1	1	−1	1	−1	1
0.046	0	−1	1	−1	−1	1	1	−1	1	−1
0.471	0	−1	−1	1	1	1	−1	−1	1	1
0.448	0	−1	−1	1	−1	1	1	1	−1	−1
0.374	0	−1	−1	−1	1	−1	1	1	1	−1
0.127	0	−1	−1	−1	−1	−1	−1	−1	−1	1

When a main effects model is fitted to the data using the log-linear model of Section 3.1 then infinite estimates result for the intercept and effects A, F and I. The modified estimates of Section 3.1 are given by the values in Table 4.

In reality the infinite estimates are not reported as such in any statistical package but just large and negative and the value depending upon any stopping criteria. Therefore it is often difficult to detect infinite estimates from the resulting estimates. The corresponding modified estimates for the effects with infinite maximum likelihood estimates are negative and have similar values, except for

A. The maximum likelihood and modified estimates for other effects are similar. In Figure 2 we have plotted the data values against the two levels for each of the factors A, F and I. Taking into account the log-linear model used, this explains in part why the A effect is larger in absolute size than the I effect, and, in turn, larger than the F effect.

Table 4. Maximum likelihood and modified estimates for the factorial data of Table 3 using a log-linear model. Parametric simulation inferences based on 100 samples are also given.

Effect	Estimates		Parametric Simulation Results			
	MLE	Modified	Mean	S.E.	95% confidence limit	
constant	$-\infty$	−0.1717	−0.0851	0.3074	−0.6777	0.4485
A	$-\infty$	−1.0627	−1.0184	0.3104	−1.5840	−0.4420
B	0.0435	0.0342	−0.0012	0.3474	−0.8117	0.6155
C	−0.1489	−0.0720	−0.1290	0.3681	−0.7250	0.6929
D	−0.2564	−0.1776	−0.1454	0.3219	−0.8382	0.4093
E	0.0316	0.0094	−0.0089	0.3197	−0.5914	0.5824
F	$-\infty$	−0.4653	−0.4725	0.3548	−1.2047	0.1107
G	−0.5556	−0.3865	−0.3446	0.3384	−1.0210	0.3813
H	−0.2834	−0.2293	−0.1616	0.3485	−0.8040	0.6029
I	$-\infty$	−0.7758	−0.6990	0.3170	−1.3908	−0.2042

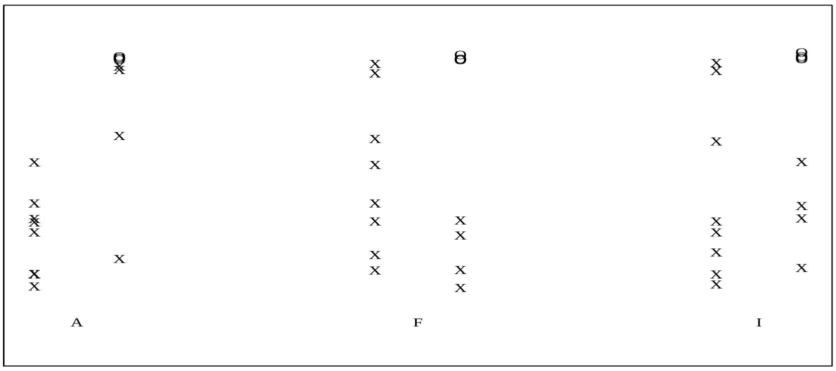

Figure 2. Plot of response on y-axis against factor levels for A, F and I. Note: X denotes uncensored, O censored; y-values are jittered.

In order to obtain standard errors for these estimates we could derive these from the information matrix evaluated at the modified estimates, rather than the infinite maximum likelihood estimate. However, there does not appear to be any theoretical reason for this and it would appear better to use a method based on the bootstrap.

A parametric method is to generate residuals ε from the standard exponential distribution, set $y^* = \varepsilon\hat{\mu}$ and censor at c if $y^* > c$. This was done and 95 percent confidence intervals are given in Table 4 using 100 samples. The true values of the parameters are all zero and there are approximately ten uncorrelated estimates. We find two of the confidence intervals do not include zero being for effects A and I which both have infinite maximum likelihood estimates. The constant and effect F have infinite maximum likelihood estimates but the confidence interval includes zero. This demonstrates that some sensible inference can be made using the methods of this paper where maximum likelihood breaks down.

6. Further Comments

Some obvious extensions of the results of Sections 2 and 3 include one to the Weibull distribution and the second to right and left censored data from accelerated failure time studies. Another extension, which is trivial to implement for the results of Section 3.1, is to the case where each individual observation is censored at a different value c_i, say. This would enable the results to be extended to the situation of random censoring provided the censoring did not yield information about the parameters of the lifetime variable (see, for example Cox and Oakes (1984), Section 1.3).

We have used the expected, rather than observed, information form of the modified score equation. Firth (1993) shows that by using the observed information form second order efficiency is improved. The observed information adjustment involves off-diagonal elements of the 'hat' matrix H and it is somewhat more computationally intensive to obtain β^* with the observed version.

The methods are motivated by sampling theory ideas and reduction in bias and we also expect to produce finite estimates under all circumstances. An alternative approach would be Bayesian where a prior distribution is introduced (see Dellaportas and Smith (1993)). Provided the prior is strictly log concave, posterior distributions which are proper result and have finite modes. However, for any data set where the maximum likelihood estimates are infinite the posterior distributions for those parameters would be very sensitive to prior distributions as the likelihood is essentially flat over a wide interval of values. The approach here does allow for objective non-informative prior information to be introduced at least in modifying the score equation and hence finding a posterior mode.

The main motivation of this paper is to find a reasonable method of estimation when infinite maximum likelihood estimates occur with censored data and the modified estimates are shown to have excellent bias reduction and variance properties in small samples.

Acknowledgements

We are particularly grateful to the referees and the Editor who made comments on this paper which led to substantial improvements. The work of Jodie Kelly was supported by an Australian Research Council postgraduate award. The computations referred to in this paper were carried out on computing equipment supplied to the School of Mathematical Sciences, Queensland University of Technology, under the Digital Equipment Agreement ERP No 2057.

Appendix 1.

We give explicit results for regression model considered in Section 3. The adjustment is given by Firth (1993), Section 4.1. For the log-linear model write the loglikehood $\ell(\beta)$ as follows:

$$\ell(\beta) = \sum_{i=1}^{n} \{ - e^{x_i^T\beta} y_i + (1-\delta_i)_i \beta y_i + (1-\delta_i) x_i^T \beta \}.$$

So, with $\eta_i = x_i^T\beta$,

$$U_r = \frac{\partial}{\partial\beta_r}\ell(\beta) = \sum_{i=1}^{n}\sum_{i=1}^{n} \{ - e^{\eta_i} y_i + (1-\delta_i)\} x_{ir}$$

$$U_{rs} = \frac{\partial^2}{\partial\beta_r\partial\beta_s}\ell(\beta) = -\sum_{i=1}^{n} x_{ir}x_{is}e^{\eta_i}y_i$$

$$U_{rst} = \frac{\partial^3}{\partial\beta_r\partial\beta_s\partial\beta_t}\ell(\beta) = -\sum_{i=1}^{n} x_{ir}x_{is}x_{it}e^{\eta_i}y_i$$

and

$$\kappa_{r,s} = \frac{1}{n}E(U_rU_s) = -\frac{1}{n}E(U_{rs}), \quad \kappa_{rst} = \frac{1}{n}E(U_{rst}), \quad \kappa_{r,st} = \frac{1}{n}E(U_rU_{st}).$$

Expectations that are required are, with $\mu_i = e^{-\eta_i}$, $E(y_i) = \mu_i(1 - e^{c/\mu_i})$ and $E[\{\mu_i(1-\delta_i) - y_i\}^2] = \mu_i^2(1 - e^{-c/\mu_i})$.

We then obtain $\kappa_{r,s} = \frac{1}{n}\sum_{i=1}^{n} x_{ir}x_{is}(1-e^{-c/\mu_i})$ or, in matrix notation, $\kappa_{rs} = \frac{1}{n}(X^TWX)_{r,s}$ where W is an $n\times n$ diagonal matrix, $W = \text{diag}(1 - e^{-c/\mu_i})$ and X is the $n \times p$ model matrix. Also

$$\kappa_{rst} = -\frac{1}{n}\sum_i x_{ir}x_{is}x_{it}(1 - e^{-c/\mu_i}),$$

$$\kappa_{r,st} = -\frac{1}{n}\sum_i x_{ir}x_{is}x_{it}(1 - e^{-c/\mu_i} - \frac{c}{\mu_i}e^{-c/\mu_i}),$$

from which $\kappa_{r,s,t} + \kappa_{r,st} = \kappa_{s,rt} - \kappa_{t,rs}$ and equals

$$\frac{1}{n}\sum_i x_{ir}x_{is}x_{it}\Big\{\frac{2c}{\mu_i e^{c/\mu_i}} - (1 - e^{-c/\mu_i})\Big\}$$

and $A_r^E = \frac{1}{2}\sum_{u,v}\kappa^{u,v}(\kappa_{r,u,v} + \kappa_{r,uv})$.

With $W_i = \mathrm{diag}(1 - e^{-c/\mu_i})$ we note that the value h_i of the hat matrix $W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{\frac{1}{2}}$ is given by $\frac{1}{n}\sum_{u,v}\kappa^{u,v}x_{iu}x_{iv}W_i$. Thus

$$A_r^E = \frac{1}{n}\sum_i\sum_{u,v} x_{ir}\kappa^{u,v}x_{iu}x_{iv}W_iW_i^{-1}\Big\{\frac{c}{\mu_i e^{c/\mu_i}} - \frac{1}{2}(1 - e^{-c/\mu_i})\Big\}.$$

This gives

$$A_r^E = \sum_i h_i\Big\{\frac{c}{\mu_i(e^{c/\mu_i} - 1)} - \frac{1}{2}\Big\}x_{ir}$$

as in equation (9).

Appendix 2.

In this appendix we show that the modified estimate is always finite.

Proposition. *Let* $U_r^* = U_r^*(\beta)$ *be defined by the Section* 3.1 *above and* $\tilde U(\beta) = (U_1^*(\beta), \ldots, U_p^*(\beta))^T$. *Let* $B = \{\beta^* : \beta^* = (\beta_1^*, \ldots, \beta_p^*)^T$ *be the solution of* $\tilde U(\beta^*) = 0\}$. *We assume that there does not exist a non-zero vector* $\gamma = (\gamma_1, \ldots, \gamma_p)^T \in R^p$ $(1 \le p \le n)$ *such that* $x_i^T\gamma \le 0$ *holds for all* $1 \le i \le n$. *Then*
(a) *For* $1 < p < n$, *the solution* β^* *is necessarily finite if there exists at least a* β^* *such that* B *is non-empty.*
(b) *For* $p = 1$ *or* $p = n$, *the* B *above is non-empty and bounded. Furthermore, the solution* β^* *is uniquely defined.*

Proof. Without loss of generality, we assume that $c = 1$ throughout this appendix.
(a) For any non-zero vector $\gamma = (\gamma_1, \ldots, \gamma_p)^T \in R^p$ and any $k > 0$, we define the objective function S as follows.

$$\begin{aligned} S(\beta + k\gamma, \gamma) &= \tilde U(\beta + k\gamma)^T\gamma = \sum_{r=1}^p U_r^*(\beta + k\gamma)\gamma_r \\ &= \sum_{i=1}^n (1 - \delta_i - 0.5h_i(\beta + k\gamma))x_i^T\gamma I(x_i^T\gamma \ne 0) \\ &\quad + \sum_{i=1}^n \phi(x_i^T\beta + kx_i^T\gamma)h_i(\beta + k\gamma)x_i^T\gamma I(x_i^T\gamma < 0) \\ &\quad + \sum_{i=1}^n \phi(x_i^T\beta + kx_i^T\gamma)h_i(\beta + k\gamma)x_i^T\gamma I(x_i^T\gamma > 0) \end{aligned}$$

$$+\sum_{i=1}^{n} y_i e^{x_i^T\beta} e^{kx_i^T\gamma}(-x_i^T\gamma)I(x_i^T\gamma<0)$$
$$-\sum_{i=1}^{n} y_i e^{x_i^T\beta} e^{kx_i^T\gamma}x_i^T\gamma I(x_i^T\gamma>0)=\sum_{j=1}^{5}J_j(\beta+k\gamma,\gamma), \quad (11)$$

say, where

$$\phi(x)=\frac{e^x}{e^{e^x}-1}>0$$

for every $x\in(-\infty,+\infty)$, and $\phi(-\infty)=1$ and $\phi(+\infty)=0$, $h_i(\beta)$ is the ith diagonal element of the expected 'hat' matrix with the $W(\beta)=\mathrm{diag}(w_1(\beta),\ldots,w_n(\beta))$ and $w_i(\beta)=1-\exp(^{-e^{x_i^T\beta}})$ defined by Section 3.1 above.

Using matrix theory, we know that the trace $\mathrm{tr}(H)=\sum_{i=1}^{p}h_i(\beta)=p$ for any vector $\beta\in R^p$. In this appendix, we assume that the matrices $H(\beta)$ and $G(\beta,\gamma)$ below exist for any given $\beta\in R^p$ and $\gamma\in R^p$. If the inverse matrices involved in $H(\beta)$ and $G(\beta,\gamma)$ do not exist, they can be replaced by the generalized inverses or the Moore-Penrose inverses.

It is obvious that for any non-zero vector γ, as $k\to+\infty$

$$w_i(\beta+k\gamma)=(1-e^{-e^{x_i^T\beta}e^{kx_i^T\gamma}})I(x_i^T\gamma>0)+(1-e^{-e^{x_i^T\beta}e^{kx_i^T\gamma}})I(x_i^T\gamma<0)$$
$$+(1-e^{-e^{x_i^T\beta}})I(x_i^T\gamma=0)$$
$$\to I(x_i^T\gamma>0)+(1-e^{-e^{x_i^T\beta}})I(x_i^T\gamma=0)=v_i(\beta,\gamma) \quad (12)$$

and

$$h_i(\beta+k\gamma)\to g_i(\beta,\gamma) \quad (13)$$

say, where $0<v_i(\beta,\gamma)\le 1$ for all $1\le i\le n$ and all β and $\gamma\in R^p$, $0\le g_i(\beta,\gamma)\le p$ is the ith diagonal element of the matrix $G(\beta,\gamma)=V(\beta,\gamma)^{1/2}X(X^TV(\beta,\gamma)X)^{-1}$ $X^TV(\beta,\gamma)^{1/2}$ with $V(\beta,\gamma)=\mathrm{diag}(v_1(\beta,\gamma),\ldots,v_n(\beta,\gamma))$.

Now, by the conditions of the Proposition we have

$$\lim_{k\to+\infty}J_1(\beta+k\gamma,\gamma)=\sum_{i=1}^{n}(1-\delta_i-0.5g_i(\beta,\gamma))x_i^T\gamma I(x_i^T\gamma\neq 0), \quad (14)$$

$$\lim_{k\to+\infty}J_2(\beta+k\gamma,\gamma)=\sum_{i=1}^{n}g_i(\beta,\gamma)x_i^T\gamma I(x_i^T\gamma<0), \quad (15)$$

$$\lim_{k\to+\infty}J_3(\beta+k\gamma,\gamma)\ge 0, \quad (16)$$

$$\lim_{k\to+\infty} J_4(\beta+k\gamma,\gamma) \geq 0, \tag{17}$$

$$\lim_{k\to+\infty} J_5(\beta+k\gamma,\gamma) = -\infty, \tag{18}$$

and

$$\sum_{j=1}^{5} J_j(\beta+k\gamma,\gamma) = J_5(\beta+k\gamma,\gamma)\Big(1+\frac{\sum_{j=1}^{4} J_j(\beta+k\gamma,\gamma)}{J_5(\beta+k\gamma,\gamma)}\Big) \to -\infty \tag{19}$$

uniformly over the $\{x_i\}$ and $n \geq 1$. The fact that the speed of $\frac{e^x}{x} \to +\infty$ is very fast as $x \to +\infty$ is used in equation (19). Thus, by equations (11) and (19) we find that there exists at least an r such that as $k \to +\infty$

$$U_r^*(\beta+k\gamma) \to +\infty \ (or \ \ -\infty) \tag{20}$$

when $\lim_{k\to+\infty} U_r^*(\beta+k\gamma)$ exists for any non-zero vector γ. If there exists a non-zero vector γ_0 such that some of the limits $\{\lim_{k\to+\infty} U_r^*(\beta+k\gamma_0), 1 \leq r \leq p\}$ do not exist, then it is obvious that $\beta+k\gamma_0$ as $k \to +\infty$ is not the solution of $\tilde{U}(\beta^*) = 0$.

Therefore, the solution β^* of $\tilde{U}(\beta^*) = 0$ is necessarily finite when the B is non-empty. This completes the proof of Proposition (a).

(b) Based on knowing which observations are censored and which are uncensored the likelihood is given by

$$L_i(\beta) = \{\mu_i^{-1}e^{-y_i\mu_i^{-1}}\}^{(1-\delta_i)} \cdot \{e^{-c\mu_i^{-1}}\}^{\delta_i} = e^{(1-\delta_i)x_i^T\beta}e^{-y_ie^{x_i^T\beta}}$$

and

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = e^{\sum_{i=1}^{n}(1-\delta_i)x_i^T\beta}e^{-\sum_{i=1}^{n} y_ie^{x_i^T\beta}}. \tag{21}$$

On the other hand, by Section 3.2 above, we know that the following $p(\beta)$ exists when $p = 1$ or $p = n$

$$p(\beta) = \prod_{i=1}^{n} e^{-0.5x_i^T\beta}(1-e^{-e^{x_i^T\beta}}) = e^{-0.5\sum_{i=1}^{n} x_i^T\beta}\prod_{i=1}^{n}(1-e^{-e^{x_i^T\beta}}). \tag{22}$$

Now, the penalised likelihood function $L^*(\beta)$ and $l^*(\beta) = \log L^*(\beta)$ are

$$L^*(\beta) = L(\beta)p(\beta) = e^{\sum_{i=1}^{n}(0.5-\delta_i)x_i^T\beta}e^{-\sum_{i=1}^{n} y_ie^{x_i^T\beta}}\prod_{i=1}^{n}(1-e^{-e^{x_i^T\beta}}) \tag{23}$$

and

$$l^*(\beta) = \log L^*(\beta) = \sum_{i=1}^{n} c_i x_i^T \beta - \sum_{i=1}^{n} y_i e^{x_i^T \beta} + \sum_{i=1}^{n} \log(1 - e^{-e^{x_i^T \beta}}), \tag{24}$$

where $c_i = 0.5 - \delta_i$.

In the following, we show that

$$m(\beta) = -l^*(\beta) = -\sum_{i=1}^{n} c_i x_i^T \beta + \sum_{i=1}^{n} y_i e^{x_i^T \beta} - \sum_{i=1}^{n} \log(1 - e^{-e^{x_i^T \beta}}) \tag{25}$$

is a strictly convex function with respect to $\beta \in C$, an open convex subset of R^p.

It is obvious that $m(\beta)$ is a twice continuously differentiable real-valued function on C. By Theorem 4.5 of Rockafeller (1970) in order to show that $m(\beta)$ is strictly convex, it suffices to show that the matrix below

$$M(\beta) = (m''_{jk}(\beta))_{1 \le j,k \le p} \tag{26}$$

is positive definite for every $\beta \in C$.

By a simple calculation, we obtain for all $1 \le j, k \le p$

$$m''_{jk}(\beta) = \sum_{i=1}^{n} y_i e^{x_i^T \beta} x_{ij} x_{ik} + \sum_{i=1}^{n} q(x_i^T \beta) x_{ij} x_{ik} = \sum_{i=1}^{n} r(x_i^T \beta) x_{ij} x_{ik}, \tag{27}$$

where $r(x_i^T \beta) = y_i e^{x_i^T \beta} + q(x_i^T \beta)$ and

$$q(x) = \frac{1}{(e^{e^x} - 1)^2} \cdot e^x [e^{e^x} \{e^x - 1\} + 1]. \tag{28}$$

It is obvious that $q(-\infty) = q(+\infty) = 0$. In the following, we only need to justify that $q(x) > 0$ for every $x \in (-\infty, +\infty)$. In order to prove this, it suffices to show (note (8)), that

$$q_1(x) = e^{e^x} \{e^x - 1\} + 1 > 0 \tag{29}$$

for every $x \in (-\infty, +\infty)$, which follows from

$$q_1(-\infty) = 0 \text{ and } q_1'(x) = e^{e^x} e^{2x} > 0 \tag{30}$$

for every $x \in (-\infty, +\infty)$.

Thus, we get for all $1 \le i \le n$ and every $\beta \in R^p$

$$r(x_i^T \beta) > 0. \tag{31}$$

Hence, for any non-zero vector $b = (b_1, \ldots, b_p)^T$ we have

$$b^T M(\beta) b = \sum_{j=1}^{p} \sum_{k=1}^{p} \sum_{i=1}^{n} r(x_i^T \beta) x_{ij} x_{ik} b_j b_k = \sum_{i=1}^{n} r(x_i^T \beta)(x_i^T b)^2 > 0, \tag{32}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^T$.

Therefore the matrix $M(\beta)$ is positive definite for every $\beta \in C$. Furthermore, the continuity of $m(\beta)$ implies that $m(\beta)$ is lower semicontinuous. Thus, $m(\beta)$ is a proper closed convex function of $\beta \in R^p$.

In order to apply Theorem 27.2 of Rockafeller (1970) to show that the conclusion of the Proposition holds, it suffices to show that for any non-zero vector $\gamma = (\gamma_1, \ldots, \gamma_p)^T \in R^p$ and $k \to +\infty$

$$m(\beta + k\gamma) \to +\infty. \tag{33}$$

By the conditions of Proposition (a), we have $x_i^T\gamma > 0$ for some i and every non-zero vector $\gamma \in R^p$. Now observe that

$$\begin{aligned}
m(\beta + k\gamma) &= -\sum_{i=1}^n c_i x_i^T(\beta + k\gamma) + \sum_{i=1}^n y_i e^{x_i^T(\beta+k\gamma)} - \sum_{i=1}^n \log(1 - e^{-e^{x_i^T(\beta+k\gamma)}}) \\
&= -\sum_{i=1}^n c_i x_i^T\beta + \sum_{i=1}^n y_i e^{x_i^T\beta} I(x_i^T\gamma{=}0) - \sum_{i=1}^n \log(1 - e^{-e^{x_i^T\beta}}) I(x_i^T\gamma{=}0) \\
&\quad -k\sum_{i=1}^n c_i x_i^T\gamma I(x_i^T\gamma \neq 0) + \sum_{i=1}^n y_i e^{x_i^T\beta} e^{kx_i^T\gamma} I(x_i^T\gamma > 0) \\
&\quad +\sum_{i=1}^n y_i e^{x_i^T\beta} e^{kx_i^T\gamma} I(x_i^T\gamma < 0) \\
&\quad -\sum_{i=1}^n \log(1 - e^{-e^{x_i^T\beta}}) I(x_i^T\gamma > 0) - \sum_{i=1}^n \log(1 - e^{-e^{x_i^T\beta}}) I(x_i^T\gamma < 0) \\
&= \sum_{l=1}^8 m_l(\beta + k\gamma),
\end{aligned} \tag{34}$$

say, where $\{m_i(\beta + k\gamma), i = 1, 2, 3\}$ are independent of k.

Since $y_i > 0$ for all $1 \le i \le n$, we know that as $k \to +\infty$

$$\begin{aligned}
&\sum_{l=1}^5 m_l(\beta + k\gamma) \\
&= \sum_{i=1}^n y_i e^{x_i^T\beta} e^{kx_i^T\gamma} I(x_i^T\gamma > 0)\Big(1 + \frac{\sum_{l=1}^4 m_l(\beta + k\gamma)}{\sum_{i=1}^n y_i e^{x_i^T\beta} e^{kx_i^T\gamma} I(x_i^T\gamma > 0)}\Big) \to +\infty
\end{aligned} \tag{35}$$

uniformly over $n \ge 1$, which is because the speed of $e^x/x \to +\infty$ is very fast as $x \to +\infty$.

Also, we obtain

$$\lim_{k\to+\infty} m_6(\beta + k\gamma) = \lim_{k\to+\infty} \sum_{i=1}^n y_i e^{x_i^T\beta} e^{kx_i^T\gamma} I(x_i^T\gamma < 0) \ge 0 \tag{36}$$

uniformly over $n \geq 1$.

On the other hand, we get

$$\lim_{k\to+\infty} m_7(\beta + k\gamma) = -\lim_{k\to+\infty} \sum_{i=1}^{n} \log(1 - e^{-e^{x_i^T\beta} e^{kx_i^T\gamma}}) I(x_i^T\gamma > 0) \geq 0 \qquad (37)$$

uniformly over $n \geq 1$, and as $k \to +\infty$

$$m_8(\beta + k\gamma) = -\sum_{i=1}^{n} \log(1 - e^{-e^{x_i^T\beta} e^{kx_i^T\gamma}}) I(x_i^T\gamma < 0) \to +\infty \qquad (38)$$

uniformly over $n \geq 1$.

This concludes the proof of (33). Thus $m(\beta)$ does not have a direction of recession which implies that B is a non-empty bounded set by applying Theorem 27.2 of Rockafeller (1970). The uniqueness of β^* follows from the strict convexity of $m(\beta)$.

References

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429-436.

Clarkson, D. and Jennrich, R. (1991). Computing extended maximum likelihood estimates for linear parameter models. *J. Roy. Statist. Soc. Ser. B* **53**, 417-426.

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

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.* **42,** 443-459.

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

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist Soc. Ser. B* **54**, 657-699.

Haberman, S. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Hamada, M. and Tse, S. K. (1988). A note on the existence of maximum likelihood estimates in linear regression models using interval-censored data. *J. Roy. Statist. Soc. Ser. B* **50**, 293-296.

Hamada, M. and Tse, S. K. (1989). MLECHK: A FORTRAN program for checking the existence of maximum likelihood estimates from censored, grouped, ordinal and binary data from designed experiments. IIQP Research Report 89-09, University of Waterloo, The Institute for Improvement in Quality and Productivity, Waterloo, Ontario, Canada.

Hamada, M. and Tse, S. K. (1992). On estimability problems in industrial experiments with censored data. *Statist. Sinica* **2**, 381-391.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* Second edn. Chapman and Hall, London.

Pettitt, A. N. (1996). Infinite estimates with fractional factorial experiments. *Statist.* **45**, 197-206.

Powell, M. J. D. (1970). A hybrid method for nonlinear algebraic equations. In *Numerical Methods for Nonlinear Algebraic Equations* (Edited by P. Rabinowitz). Grodon and Breach Science Publishers, London, New York.

Rockafeller, R. T. (1970). *Convex Analysis.* Princeton University Press, Princeton.

Silvapulle, M. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression model for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48,** 100-106.

The Numerical Algorithms Group Limited (1993). *NAG Fortran Library Manual-Mark 16. Volume 1,* first edn. Oxford, U.K.

Verbeek, A. (1989). The compactification of generalised linear models. In *Statistical Modelling* (Edited by A. Decorli, B. Frances, R. Gilchrist and G. Seebeer), 314-327. Springer, New York.

School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland, 4001, Australia.

E-mail: a.pettitt@fsc.qut.edu.au

E-mail: j.kelly@fsc.qut.edu.au

E-mail: j.gao@fsc.qut.edu.au

(Received December 1995; accepted June 1997)