Inference in the Presence of Likelihood Monotonicity for Proportional Hazards Regression

Abstract: Proportional hazards are often used to model event time data subject to censoring. Small samples involving discrete covariates with strong effects can lead to infinite maximum partial likelihood estimates. A methodology is presented for eliminating nuisance parameters estimated at infinity using approximate conditional inference. Conditional higher-order likelihood inference may then be applied to remaining parameter components.

1. Introduction

The proportional hazards regression model [Cox, 1972] is commonly used to model the dependence between time to an event and various covariates. When times to event have continuous distributions, this model constrains the hazard (defined as the derivative of the log of the survival function) to be a baseline function times the exponential of a linear combination of the covariates, and generally the practitioner wants to learn about the coefficients in this linear relationship, while imposing minimal conditions on the baseline hazard function. Often the event times for some experimental subjects are not observed precisely, but are observed to exceed some time, called the censoring time; this phenomenon is known as right censoring.
In order to eliminate the effect of the unknown baseline hazard function from the analysis, inference is often performed by constructing a function from the full likelihood, formed by discarding terms containing information on the gaps between succeeding failures [Cox, 1972]. This resulting function, called the partial likelihood, has many properties in common with the full likelihood, and is often used in the same way that the full likelihood is used; for example, parameters are often estimated by maximizing the partial likelihood, standard errors are often calculated from the second derivative of the log of the partial likelihood functions, and the change in the maximized likelihood as one moves from a larger model to a smaller model nested within it is often used for testing.

This partial likelihood is exactly the same as the likelihood that arises from a multinomial regression model with a data set derived from the proportional hazards data set, to be described below, and, in turn, the multinomial distribution (and hence its likelihood) will be represented exactly by a certain conditional logistic regression model.

Some logistic regression models give rise to likelihood functions that do not have a maximum; instead, there exists one or more contrasts of the parameters such that as this contrast is increased to infinity, the likelihood continues to increase. This monotonicity in the likelihood complicates estimation and testing of logistic regression parameters. At some point during the iterative process of fitting the model, an important numerical calculation (generally first the inversion of the second derivative matrix of the log likelihood) becomes impossible, and the algorithm stops. The most naive response to the
problem of likelihood monotonicity is to report numerical results at the last computable step, and to exit with a warning message. A more sophisticated approach involves regularization, either with a penalty function in frequentist inference, or with an informative prior for Bayesian inference [Firth, 1993].

Because of the association among logistic regression, multinomial regression, and proportional hazards regression models described above, proportional hazards regression methods inherit this difficulty. Present methods for inference for proportional hazards regression in the case of partial likelihood monotonicity include termination of the incomplete iterative solution, as described above for logistic regression, and the application of the Bayesian method of Firth [1993] in this proportional hazards regression context [Heinze and Schemper, 2001]. This method is implemented in both SAS/STAT [SAS Institute, 2019] and in R [Heinze and Ploner, 2018], and always results in finite estimators.

provides a method for diagnosing and adjusting for this monotonicity in conditional logistic regression models. applies these techniques to multinomial regression models. This manuscript extends the methods of and to multinomial regression, to facilitate exact and approximate inference, and to proportional hazards regression to provide approximate inference.

2. The Motivating Example, Revisited

Lee [2017] presents data on Hg concentrations of 133 fish from 20 different species. This paper examines the effect of the proportion of nearby wetlands on Hg levels in 44 largemouth bass. Two fish had Hg levels below detec-

imsart-generic ver. 2010/09/07 file: blinded.tex date: January 20, 2020
Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropWetland</td>
<td>-1.041</td>
<td>0.779</td>
<td>0.18137</td>
</tr>
<tr>
<td>SedMeHgH</td>
<td>-0.372</td>
<td>0.346</td>
<td>0.28263</td>
</tr>
<tr>
<td>SedAVSH</td>
<td>-15.862</td>
<td>0.333</td>
<td>0.00000</td>
</tr>
<tr>
<td>SedLOIH</td>
<td>15.645</td>
<td>0.346</td>
<td>0.00000</td>
</tr>
<tr>
<td>SedMeHgH:SedAVSH</td>
<td>12.954</td>
<td>0.366</td>
<td>0.00000</td>
</tr>
<tr>
<td>SedMeHgH:SedLOIH</td>
<td>-10.820</td>
<td>0.378</td>
<td>0.00000</td>
</tr>
</tbody>
</table>

We treat values below the detection level, exhibiting left censorship. As techniques for right-censored data are better developed than for left-censored data, the inverse of Hg level is used as a response variable in a regression. Covariates other than proportion wetlands are dichotomized. Covariates of interest are proportion of wetlands (PropWetland), high sedimentary Methyl Hg (SedMeHgH), high sedimentary Acid Volatile Sulfide (SedAVSH), and high sedimentary Loss on Ignition (SedLOIH). In each case, the prefix Sed indicates that the variable is derived from sediment measurements, and the suffix H indicates that the variable is dichotomized, with high values coded as 1. Two-way interactions for covariates were included, but some were not identifiable. Non-identifiable interactions are omitted from model fits.

An application of proportional hazards regression (in this case, using the software Therneau [2015]), yields the following parameter estimates in Table 1. The interaction between SedMeHgH and SedAVSH is quite large; these lead to large main effects as well. Furthermore, the software provides a warning message indicating that convergence fails. This is generally (but not always) because the true maximizer is at infinity.
Problematically, the proportional hazards algorithm fails to converge, because the partial likelihood is monotone in a linear combination of interaction variables. Removing these interactions gives the results in Table 2, but, as the wetlands estimate changes quite a bit, dropping interesting interactions may lead to bias.

### 2.1. Parallels with Logistic Regression

Analyzing this data set by dichotomizing the response and applying a logistic regression model has the same problem. Consider random variables $Y_j$ taking the value 1 if Hg concentration for fish $j$ exceeds $0.2\mu g/g$, and taking the value 0 otherwise. Model these random variables as independent, with

$$P[Y_j = 1] = \frac{\exp(z_j \gamma)}{1 + \exp(z_j \gamma)}, \quad P[Y_j = 0] = \frac{1}{1 + \exp(z_j \gamma)},$$

for $x_j$ the vectors containing covariates as given in Table 1, and $z_j = (x_j, 1)$. (Note that the intercept term is added as the last, rather than first, parameter, in order to keep the position of other covariates constant). In this case, inference on the proportion of wetlands coefficient, $\gamma_1$, is desired, without specifying values for the other parameters. Sufficient statistics for the parameter $\gamma$ are given by

$$W = Z^\top Y,$$
Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropWetland</td>
<td>-6.308</td>
<td>10.427</td>
<td>0.54521</td>
</tr>
<tr>
<td>SedAVSH</td>
<td>-20.829</td>
<td>6206.281</td>
<td>0.99732</td>
</tr>
<tr>
<td>SedLOIH</td>
<td>20.955</td>
<td>6206.281</td>
<td>0.99731</td>
</tr>
<tr>
<td>SedMeHgH</td>
<td>-1.716</td>
<td>1.122</td>
<td>0.12612</td>
</tr>
<tr>
<td>SedAVSH:SedMeHgH</td>
<td>17.826</td>
<td>9224.451</td>
<td>0.99846</td>
</tr>
<tr>
<td>SedLOIH:SedMeHgH</td>
<td>4.050</td>
<td>8778.815</td>
<td>0.99963</td>
</tr>
<tr>
<td>(Intercept)</td>
<td>1.323</td>
<td>0.653</td>
<td>0.04268</td>
</tr>
</tbody>
</table>

for \( Z \) the matrix with rows \( z_j \). Model (1) forms a natural exponential family, and so the distribution of \( W_1|W_2,\ldots,W_7 \) depends on the parameter of interest \( \gamma_1 \), and not on any of the other model parameters \( \gamma_2,\ldots,\gamma_7 \). Computations to calculate this conditional distribution and complete estimation and testing can be lengthy, and more often, the estimate \( \hat{\gamma} \) maximizing the full likelihood is presented. For sufficiently large samples, the distribution of \( \hat{\gamma} \) is approximately multivariate normal. The approximating distribution has expectation \( \gamma \), and a variance matrix calculated from \( \ell''(\gamma) \), with \( \ell \) the logistic regression log likelihood; since \( \gamma \) is unknown, the estimate \( \hat{\gamma} \) is substituted into the variance formula.

Logistic regression gives the results in Table 3. Note that the logistic regression results contain an additional parameter, labeled Intercept. Note further that large parameter estimates appear in Table 3 in the same places as in Table 1. In this case, no finite maximum likelihood estimators exist, and substituting directly into the variance formula is impossible.

Inferential issues arise in this logistic regression, because the estimates of parameters not of interest are infinite. If all parameter estimates were finite, standard normal approximations to \( p \)-values and confidence intervals could be
employed. Were the only parameter estimated at infinity the interest parameter, \( p \)-value calculations employing a continuity correction may be performed using standard large-sample approximate multivariate normal methods. That is, the conditional \( p \)-value for testing the null hypothesis \( \gamma_1 = \gamma_1^0 \) is

\[
2 \min( P_{\gamma_1^0} [ W_1 \geq w_1^- | W_2, \ldots, W_7 ] , P_{\gamma_1^0} [ W_1 \leq w_1^+ | W_2, \ldots, W_7 ] ),
\]

where \( w_1^+ \) and \( w_1^- \) represent the observed values for \( W_1 \), moved higher and lower by the minimal spacing of values in the conditional sample space of \( W_1 | W_2, \ldots, W_7 \). One of the two probabilities is 1, and for the other, the value \( w_1^+ \) or \( w_1^- \) of \( W_1 \) at which the maximum likelihood estimator must be calculated sits strictly between its largest and smallest possible values, and so a finite maximizer exists.

Furthermore, the relationship between null hypothesis \( \gamma_1^0 \) and the \( p \)-value can be inverted to give a confidence interval. If \( \gamma_1 \) is involved a contrast estimated at infinity, then the observed value of \( W_1 \) is at the end of the conditional sample space for \( W_1 | W_2, \ldots, W_7 \), and hence the confidence interval has one end point \( \pm \infty \), with the other end point determined by solving the equation

\[
P_{\gamma_1^0} [ W_1 \geq w_1^- | W_2, \ldots, W_7 ] = \alpha/2
\]

or the equation

\[
P_{\gamma_1^0} [ W_1 \leq w_1^+ | W_2, \ldots, W_7 ] = \alpha/2.
\]

The appropriate value \( w_1^+ \) or \( w_1^- \) is associated with a maximum likelihood estimator \( \hat{\gamma} \) that is not infinite along any contrast involving \( \gamma_1 \).
Proportional hazards regression models share this same phenomenon of infinite parameter estimates leading to inference problems only when the interest parameter is not involved in the move to infinity. Hence this paper only considers cases in which infinite parameter estimates do not involve the parameter of interest.

? recommends inference in such logistic regression cases by considering the conditional distribution $W_1|W_2, \ldots, W_7$, creating a smaller data set with the same conditional distribution that avoids infinite estimates, and applying normal approximation methods, or higher-order inference to provide more accurate approximation methods, to the conditional distribution. ? extends this technique to multinomial regression. Later sections of this manuscript extend this technique to proportional hazards regression. This extension is achieved only approximately, in that analogs of calculations like (2) fail to construct statistic vectors that are sufficient, and so resulting conditional probability calculations are only approximate.

3. Models beyond Logistic Regression

In this section, we review the multinomial, logistic, and proportional hazards regression models, review the observation by ? that the multinomial regression may be expressed as a conditional logistic regression, and observe that the partial likelihood from the proportional hazards model may be expressed as the likelihood from a multinomial model.
3.1. Multinomial regression

Suppose that \( \mathcal{P} \) multinomial trials are observed; for trial \( m \in \{1, \ldots, \mathcal{P}\} \), an alternative \( D_m \) an element of the set \( \mathcal{A}_m \) of alternatives is observed, with probability

\[
P[D_m = d_m] = \frac{\exp(x_{md_m} \beta)}{\sum_{k \in \mathcal{A}_m} \exp(x_{mk} \beta)}.
\]

(4)

Here \( x_{mj} \in \mathbb{R}^D \) are covariate vectors associated with each of the alternatives. The likelihood is given by

\[
P[\mathcal{Y} = \{y_{mj}\forall m,j\}] = L(\beta) = \prod_{m \in \mathcal{C}} \frac{\exp(x_{mD_m} \beta)}{\sum_{k \in \mathcal{A}_m} \exp(x_{mk} \beta)},
\]

(5)

for

\[
\mathcal{C} = \{1, \ldots, \mathcal{P}\}.
\]

(6)

The log likelihood for this model is

\[
\ell(\beta) = \beta^\top U - \sum_{m \in \mathcal{C}} \log \left( \sum_{j \in \mathcal{A}_m} \exp(x_{ji} \beta) \right),
\]

(7)

for \( U = \sum_{j=1}^{\mathcal{P}} x_{jiD_i} \), a sufficient statistic for \( \beta \).
3.2. The multinomial regression model as a special case of conditional logistic regression

Consider the the design matrix

\[
\begin{pmatrix}
\mathcal{D} \text{ columns} & \mathcal{P} \text{ columns} \\
\begin{bmatrix}
x_{11} & 1 & 0 & \ldots & 0 \\
x_{12} & 1 & 0 & \ldots & 0 \\
\vdots & 1 & \vdots & \vdots & \vdots \\
x_{1} & 1 & 0 & \ldots & 0 \\
x_{21} & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{2} & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x_{P1} & 0 & 0 & \ldots & 1 \\
x_{PH} & 0 & 0 & \ldots & 1 
\end{bmatrix}
\end{pmatrix}
\]

That is, the first \( \mathcal{D} \) columns are constructed by stacking the \( \mathcal{H}_1 \) covariate vectors for the first set of choices, followed by the \( \mathcal{H}_2 \) covariate vectors associated with with the second set of choices, and on, until the last \( \mathcal{H}_P \) rows contain the last \( \mathcal{H}_P \) covariate vectors. This forms a matrix with \( \mathcal{H}_1 + \cdots + \mathcal{H}_P \) rows and \( \mathcal{D} \) columns. To this matrix is added additional \( \mathcal{P} \) columns on the right. The first of these is a vector with all zeros, except for 1 in the first \( \mathcal{H}_1 \) positions. The second is a vector of all zeros, except for 1 in positions \( \mathcal{H}_1 + 1 \), ...
through $\mathcal{H}_1 + \mathcal{H}_2$; that is, 1 starts just after 1 ended in the previous column. These repeat to fill the remainder of the last $P$ columns.

Let $Z$ denote the matrix in (8). Let $z_j$ be row vector representing row $j$ of $Z$. Let $\tau$ be an arbitrary row vector in $\mathbb{R}^P$, and $\gamma = (\beta, \tau)$. Then, if $Y$ is a random vector of independent responses, satisfying (1), and if

$$
\begin{pmatrix}
V \\
U
\end{pmatrix} = Z^\top Y,
$$

then the distribution of $U$ conditional on $V$ is multinomial.

Throughout this paper, inference for proportional hazards models is related to conditional inference on some associated logistic regression models. The proportional hazards model parameter vector will be denoted $\beta$. The associated logistic regression model has parameters corresponding to the same covariates, plus one or more for one or more intercept terms; this parameter vector is denoted $\gamma$.

### 3.3. Proportional hazards regression partial likelihood in relation to multinomial regression

The proportional hazards regression model considers a set of times $T_1, \ldots, T_K$ at which $K$ patients have an event. Assume that there exist vectors $x_1, \ldots, x_K$ such that

$$
\frac{d}{dt} \log(P[T_j \geq t]) / \frac{d}{dt} \log(P[T_k \geq t]) = \exp((x_j - x_k) \beta) \text{ for all } j, k. \quad (9)
$$

Suppose further that there exist censoring times $C_k$, either fixed, or independent of $T_k$, such that one may observe only $W_k = \min(T_k, C_k)$, and indicators
δ_k, taking the value 1 if T_k ≤ C_k, or 0 if T_k > C_k. Cox [1972] proposed estimating β by maximizing the partial likelihood, also given by

\[ P_\beta [Y_{mj} = y_{mj}\forall m, j] = L(\beta) = \prod_{m \in C} \exp(x_m D_m \beta) / \sum_{k \in A_m} \exp(x_m k \beta), \]  

as in (5), except that C = \{k|δ_k = 1\} (that is, with censored individuals removed), and A_m = \{a \in \{1, \ldots, K\}|T_a \geq T_m\}, and D_m = m.

4. Infinite Estimates

This section builds techniques for determining components of a proportional hazards parameter vector not having a finite maximum partial likelihood estimator. \( \delta \) determines extreme fitted probabilities may be determined in by the following theorem:

**Theorem 1.** Suppose that matrix Z is of full rank. Take \( u \in \mathbb{R}^K \) and \( r \) and \( s \) are column vectors of non-negative numbers such that

\[ u = (Z^T Z)^{-1} Z^T (s - r), \]  

\[ (u^T (Z^T Z)^{-1} Z^T - n^T) s - u^T (Z^T Z)^{-1} Z^T r = 0, \]  

and

\[ (I - Z(Z^T Z)^{-1} Z^T) (s - r) = 0, \]  

Determine the maximum number of positive entries in \( r \) and \( s \) satisfying (12) and (13). Positive entries in \( r \) and \( s \) indicate fitted all probabilities fixed by the conditioning event \( Z^T Y = u \) at 0 and 1 respectively. Then \( Y | Z^T Y = u \) remains the same if observations with positive value for either \( r \) or \( s \) are removed prior to calculation.
Likelihood Monotonicity

? further suggests the following procedure for inference in logistic regression with parameters not of interest estimated at infinity. Partition $Z$ as $(Z_i, Z_n)$. Here $Z_i$ is the matrix of covariates associated with interest parameters, and $Z_n$ is the matrix of covariates associated with other ("nuisance") parameters. Obtain $r$ and $s$ from Theorem 1 using the reduced design matrix $Z_n$, let $Z_{n1}$ and $Z_{i1}$ be the matrices $Z_n$ and $Z_i$ respectively, with rows corresponding to non-zero entries in $r$ or $s$ omitted, let $Z_{n2}$ be the matrix $Z_n$ with rows corresponding to non-zero entries in $r$ or $s$ retained, and let $Y_1$ be the vector $Y$ with entries corresponding to non-zero entries in $r$ or $s$ omitted. Let $\hat{Y}_2$ be a vector with length equal to the number of non-zero entries in $r$ or $s$, with entries 1 corresponding to positive entries of $s$, and 0 corresponding to positive entries for $r$. Then the conditional distributions

$Z_i^\top Y | Z_n^\top Y$ and $Z_{n1}^\top Y_1 + Z_{n2}^\top \hat{Y}_2 | Z_{n1}^\top Y_1$ are the same.

? notes that the same phenomenon of likelihood monotonicity occurs in multinomial regression, and suggests addressing this problem by constructing design matrix (8), with as many rows as there are alternatives in the original multinomial regression model, and with $K = D + P$ columns, formed by combining the original covariate vectors with new indicator variables. Theorem 1 is then used to indicate which subjects in the logistic regression should be omitted to leave conditional inference for the variables of interest unchanged, and to yield finite conditional maximum likelihood estimators.
5. The Proposed Method

We propose constructing hypothesis tests and confidence intervals for interest parameters in proportional hazards regressions models in which parameters not of direct interest are estimated at infinity, by converting the data set to one yielding approximately equivalent inference and having a likelihood that is non-monotone in all contrasts of parameters.

First, construct a multinomial regression data set from the proportional hazards data set. The new multinomial regression data set has as many trials \( \mathcal{P} \) as there are event times. For trial \( m \), the risk set \( \mathcal{A}_m \) is the set of observations whose event and censoring times are greater than or equal to that under consideration, \( W_m \). The associated alternative covariate vectors are the covariate vectors for these individuals. In the multinomial regression model of §3.1, no relationship need exist among alternatives in the various trials; our proposal is to reuse covariate vectors until the individuals are removed from the risk set \( \mathcal{A}_m \). This identification of the multinomial and proportional hazards regression model is motivated by noting that likelihood contributions multiply for the multinomial model, because of independence, and for the proportional hazards model, because they are defined conditionally. Ideally, frequentist inference for these two models will be different, because they arise from different sampling distributions; we approximate proportional hazards regression by multinomial regression for the purpose of eliminating contrasts estimated at infinity.

This elimination of parameters estimated at infinity is performed as in ?.
We express multinomial regression as conditional logistic regression, analyze
the conditional logistic regression to identify subjects whose fitted probabil-
ities are 0 or 1. This determines which multinomial subjects have selection
probability either 0 or 1, under the conditioning event, and, in turn, implies
which survival subjects are guaranteed either to have or to fail to have the
next event. We remove these from the proportional hazards regression. Stan-
dard statistical software (for example, Therneau [2015]) removes redundant
covariates.

After the proportional hazards model is converted to one giving approxi-
mately equivalent inference on interest parameters, and one for which maxi-
mum likelihood estimates are finite, standard normal theory, or higher-order,
inference can be applied. Without loss of generality, assume that inference on
the first covariate is desired. If \( \ell \) as in (7), with observations determined as
above removed, and with covariates making the covariate matrix no longer
of full rank removed, let \( \hat{\beta} \) be the maximizer of \( \ell(\beta) \), and let \( \tilde{\beta} \) represent
the maximizer of \( \ell(\beta) \) subject to \( \beta_1 = 0 \). Let \( \hat{\sigma}_1^2 \) represent the inverse of
the first diagonal entry of \( \ell''(\hat{\beta})^{-1} \). Then the Wald statistic for testing the
null hypothesis \( \beta_1 = 0 \) is \( \hat{z} = \hat{\beta}_1 / \hat{\sigma}_1 \), and the signed root of the likelihood
ratio statistic is \( \hat{w} = \sqrt{2[\ell(\tilde{\beta}) - \ell(\hat{\beta})]} \); both of these statistics have a null hy-
pothesis distribution that is approximately standard normal. Davison [1988]
argues that, in a simpler context, a saddlepoint approximation, calculated
from \( \hat{z} \) and \( \hat{w} \), may be used to provide an approximation to the conditional
probabilities in (3); the result, without a continuity correction, did not show
an improvement over the signed root of the likelihood ratio statistic.
Clarkson and Jennrich [2000] provide a related technique for detecting the presence of parameters estimated at infinity, but this technique does not provide for inference after detection.

Alternative inferential techniques in this context are to run the standard algorithm until it fails numerically. An advantage of this approach is that it produces confidence intervals and $p$-values generally, although not always, close to our approach suggested here. Unfortunately, performance and stability not guaranteed.

One might also employ regularization, potentially using a proper Bayesian prior. An advantage of regularization is numerical stability. Dependence on the choice of regularization is a drawback. Zhang and Kolassa [2008] propose matching priors.

6. Examples

After converting the proportional hazards regression task into a conditional logistic regression task whose likelihood is exactly the same as the original partial likelihood, identifying observations whose selection probabilities are zero or one based on covariates other than proportion wetlands, removing these subjects, performing proportional hazards regression on the reduced data set provides Table 4.

Note that the effect of proportion wetlands differs in Tables 1, 2, and 4. The result of the proposed method differs only minimally from the failed-to-converge result (-1.008 vs -1.041), but differs substantially from the result dropping offending covariate (-1.008 vs. -0.383, about one standard error
Table 4
Results of Proportional Hazards Regression for Inverse Hg Levels, after Removing Individuals With Deterministic Selection Probabilities

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropWetland</td>
<td>-1.008</td>
<td>1.287</td>
<td>0.434</td>
</tr>
<tr>
<td>SedAVSH</td>
<td>-5.030</td>
<td>3.005</td>
<td>0.094</td>
</tr>
<tr>
<td>SedLOIH</td>
<td>4.800</td>
<td>2.892</td>
<td>0.097</td>
</tr>
<tr>
<td>SedMeHgH</td>
<td>-0.369</td>
<td>2.889</td>
<td>0.898</td>
</tr>
<tr>
<td>SedAVSH:SedMeHgH</td>
<td>2.125</td>
<td>2.734</td>
<td>0.437</td>
</tr>
<tr>
<td>SedLOIH:SedMeHgH</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 5
Results of Proportional Hazards Regression for Inverse Hg Levels With One Covariate Relabeled

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>PropWetland</td>
<td>-1.041</td>
<td>0.779</td>
<td>0.181</td>
</tr>
<tr>
<td>SedMeHgL</td>
<td>0.372</td>
<td>0.346</td>
<td>0.283</td>
</tr>
<tr>
<td>SedAVSH</td>
<td>-2.908</td>
<td>0.333</td>
<td>0.000</td>
</tr>
<tr>
<td>SedLOIH</td>
<td>4.825</td>
<td>0.346</td>
<td>0.000</td>
</tr>
<tr>
<td>SedMeHgL:SedAVSH</td>
<td>-12.954</td>
<td>0.491</td>
<td>0.000</td>
</tr>
<tr>
<td>SedMeHgL:SedLOIH</td>
<td>10.820</td>
<td>0.491</td>
<td>0.000</td>
</tr>
</tbody>
</table>

different). One covariate was removed because of non-identifiability. See the R package PHInfiniteEstimates.

The phenomenon in Table 1 is easier to understand if the effect of Sediment Methyl Mercury is reversed; define a new variable SedMeHgL representing values of this variable at or below the median. Then the naive non-convergent model fit corresponding to Table 1 is as given in Table 5.

This motivates division of the observations into four groups:

- A. Those with either SedMeHgL equal to zero (that is, SedMeHgH=1), or both SedAVSH and SedLOIH equal to zero. Observations in this group have both interactions zero, and so the infinite parameter estimates do not have an effect.

- B. Those with all of SedMeHgL=1 (that is, SedMeHgH=0), SedAVSH=1,
and SedLOIH=1. For these observations, the large parameter estimates appear to cancel, leaving a finite linear predictor.

- C. Those with SedMeHgL=1 (that is, SedMeHgH=0), SedAVSH=1, and SedLIOH=0. These observations will always have their selection probabilities estimated at zero.

- D. Those with SedMeHgL=1 (that is, SedMeHgH=0), SedLIOH=1, and SedAVSH=0. These observations will always have their selection probabilities estimated at one.

Table 6 gives the data values, sorted by response, and labeled in these groups. Recall that because the response variable in the model is inverse mercury level, the proportional hazards regression fits the probability of lying higher in the table. Recall also that the first two observations were censored at their detection level.

Group D is empty. Observations in group C make up the lowest three observations in the data set; because the model is rich enough to allow for a separate model fit for these observations, this fit diverges to negative infinity. Note also that groups A and B are scattered throughout the data.

One should check that the proposed procedure, with certain observations chosen for removal, does not target this removal in such a way as to adversely effect the distribution of \( p \)-values under the null hypothesis. Figure 1 shows the portion of the quantile plot for these \( p \)-values. Ideal frequentist inference would have this curve linear through \((0,0)\) and with slope 1. We see that the approximations to \( p \) values given by standard normal approximations to
## Table 6

**Labeled Data Values**

<table>
<thead>
<tr>
<th>SedMeHgL</th>
<th>SedAVSH</th>
<th>SedLOIH</th>
<th>PropWetland</th>
<th>Group</th>
<th>Hg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.000</td>
<td>C</td>
<td>0.100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.030</td>
<td>C</td>
<td>0.100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.000</td>
<td>C</td>
<td>0.117</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.010</td>
<td>B</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.127</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.129</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.130</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.010</td>
<td>A</td>
<td>0.140</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.163</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.181</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.950</td>
<td>B</td>
<td>0.200</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.300</td>
<td>A</td>
<td>0.207</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.217</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.241</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.247</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.010</td>
<td>B</td>
<td>0.249</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.010</td>
<td>B</td>
<td>0.249</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.256</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.370</td>
<td>A</td>
<td>0.262</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.262</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.271</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.018</td>
<td>A</td>
<td>0.313</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.300</td>
<td>A</td>
<td>0.334</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.018</td>
<td>A</td>
<td>0.351</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.500</td>
<td>A</td>
<td>0.370</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.010</td>
<td>B</td>
<td>0.373</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.390</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.020</td>
<td>B</td>
<td>0.393</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.421</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>0.471</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.200</td>
<td>A</td>
<td>0.545</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
<td>A</td>
<td>0.551</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.160</td>
<td>A</td>
<td>0.572</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.190</td>
<td>A</td>
<td>0.575</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.160</td>
<td>A</td>
<td>0.626</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.500</td>
<td>A</td>
<td>0.630</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.020</td>
<td>A</td>
<td>0.633</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.500</td>
<td>A</td>
<td>0.660</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.800</td>
<td>A</td>
<td>0.820</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
<td>A</td>
<td>1.119</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.800</td>
<td>A</td>
<td>1.420</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.020</td>
<td>A</td>
<td>1.803</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.160</td>
<td>A</td>
<td>1.803</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.800</td>
<td>A</td>
<td>4.220</td>
</tr>
</tbody>
</table>
the distribution of Wald and log likelihood ratio statistics are close to the nominal values, but a bit high.

Fig 1. Observed and Nominal p-values

References


John Edward Kolassa. Inference in the presence of likelihood monotonicity


