Infinite Parameter Estimates in Toxicology Studies

John E. Kolassa[1] and Juan Zhang

Rutgers, the State University of New Jersey and Allergan Pharmaceuticals

June 17, 2019

[1]Supported by grant NSF DMS 1712839

Motivating Example

Lee [2017] presents data on Hg concentrations of 133 fish from 20 different species.

Examine effect of nearby wetlands on Hg levels in the 44 largemouth bass.

- 2 fish had Hg levels below detection level.
  - Treat values below detection limit as left censored at limit.
  - Reverse scale, to make more conventional right censoring.
- Covariates:
  - per cent wetlands,
  - High Methyl Hg,
  - High Acid Volatile Sulfide,
  - High Loss on Ignition.

What is the effect of wetlands, controlling for other variables?

Problematic Results

Results of a Proportional Hazards regression in R:

Variable	Estimate	SE	p
PctWetland	-0.01041	0.01437	0.46885
SedAVSH	-9.97778	40.05593	0.80329
SedLOIH	9.76029	40.05606	0.80749
SedMeHgH	-0.37186	0.53300	0.48538
SedAVSH:SedMeHgH	7.06932	40.07296	0.85997
SedLOIH:SedMeHgH	-4.93531	40.08296	0.90201

- Problem: Proportional hazards algorithm fails to converge
- Partial Likelihood appears monotone in a linear combination of interaction variables.

Naive Results

Result without interactions:

Variable	Estimate	SE	p
PctWetland	-0.00383	0.00899	0.67019
SedAVSH	-3.05523	0.85291	0.00034
SedLOIH	3.68796	0.97336	0.00015
SedMeHgH	0.58095	0.34993	0.09688

Outline

Maximum Likelihood

Standard analysis for an exponential family regression model (GLM):

- Independent observations
- $P_{\boldsymbol{\beta}}\left[Y^j = y^j\right] = \exp(\boldsymbol{\beta}^\top \mathbf{x}_j y^j - \psi(\boldsymbol{\beta}) - g(y^j))$
- $\mathbf{x}_j = (x_j^1, \ldots, x_j^d)^\top$
- $\mathbf{U} = \mathbf{X}^\top \mathbf{Y}$
  - $\mathbf{X}$ is the $M \times d$ matrix with rows $\mathbf{x}_j$.
  - $\mathbf{U} = \mathbf{X}^\top \mathbf{Y}$ is the sum of the particular covariate times the response.
- $L(\boldsymbol{\beta}) = \prod_j P_{\boldsymbol{\beta}}\left[Y^j = y^j\right]$
- $\ell(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$
- $\hat{\boldsymbol{\beta}}$ maximizes $\ell(\boldsymbol{\beta}) = \sum_j [\boldsymbol{\beta}^\top \mathbf{x}_j y_j - \psi(\boldsymbol{\beta})] = \boldsymbol{\beta}^\top \mathbf{U} - n\psi(\boldsymbol{\beta})$

Inference for Finite Estimators

Test $H : \beta_j = \beta_j^\circ$ vs. $K : \beta_j \neq \beta_j^\circ$.

- Want to do this using a reference distribution that does not depend on $\boldsymbol{\beta}^{-1} = (\beta^1, \ldots, \beta^{j-1}, \beta^{j+1}, \ldots, \beta^d)$
- P-value calculated with reference to distribution of $U^j | \boldsymbol{U}^{-j}$ for $\boldsymbol{U}^{-j} = U^1, \ldots, U^{j-1}, U^{j+1}, \ldots, U^d$

Standard inference uses an approximation to this distribution.

- Multivariate CLT says $\boldsymbol{U}$ approximately Gaussian
  - with expectation ψ^j,
  - variance calculated from ψ'' using standard multivariate normal conditional variance formula.
  - Derivatives calculated at MLE with constraint $\beta_j = \beta_j^\circ$.
- This fails when this MLE is infinite.

Proportional Hazards Regression Model

For subject i, observe

- measurement Y_i
- C_i indicating true measurement (1) or above threshold (censored) (0)
- covariates $\boldsymbol{x}_i$

Model

- subjects as acting independently
- censoring (loss to followup) as not dependent on covariates
- Subjects called "at risk" before censoring, event.
  - $R_i = \{j | Y_j \geq Y_i\}$.
- among those at risk at an observed Y value
  $P\,[\text{subject } i \text{ has the event}] = \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i) / \sum_{j \in R_i} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)$

Proportional Hazards Regression (partial) Likelihood

Leads to partial log likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i|C_i=1} (\boldsymbol{x}_i\boldsymbol{\beta} - \log[\sum_{j\in R_i} \exp(\boldsymbol{x}_j\boldsymbol{\beta})]) \\ &= \boldsymbol{u}\boldsymbol{\beta} - \sum_{i|C_i=1} \log[\sum_{j\in R_i} \exp(\boldsymbol{x}_j\boldsymbol{\beta})]\end{aligned}$$

for $\boldsymbol{u} = \sum_{i|C_i=1} \boldsymbol{x}_i$

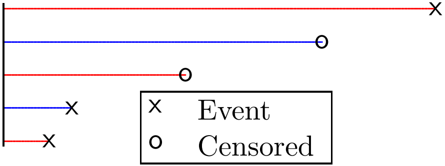
- u_j is sum of covariate value for subjects with event.
- When covariate is 0 or 1, u_j is number of subjects with event and 1 for covariate j.

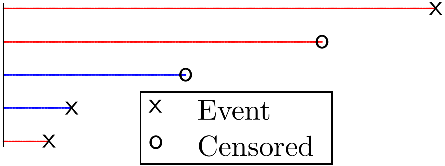
Cox [1972]

Note similarity to multinomial log likelihood.

Partial Likelihood

- Two groups, red and blue.
- Two data sets, having the same partial likelihood.





- Data sets are identical except two consecutive censored items are swapped between groups.
- In a fully-parametric model, right panel is stronger evidence in favor of red.
- Partial likelihood treats them identically.

The Logistic Regression Model

Observe M responses $Y_j \in \{0,1\}$ associated with covariates $\boldsymbol{x}_j \in \Re^d$.

- $P[Y_j = 1] = \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)/(1 + \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j))$
- Independent.
- In this case, if covariates are 0 or 1, then $\boldsymbol{U}$ is the number of subjects with event and with covariate value 1.

In example, $Y = 1$ if Hg concentration exceeds $.2\mu$g/g, 0 otherwise.

	Estimate	Std. Error	Pr(>|z|)
(Intercept)	1.32282	0.65267	0.04268
PctWetland	-0.06308	0.10427	0.54521
SedAVSH	-20.82888	6206.28059	0.99732
SedLOIH	20.95513	6206.28065	0.99731
SedMeHgH	-1.71573	1.12170	0.12612
SedAVSH:SedMeHgH	17.82589	9224.45144	0.99846
SedLOIH:SedMeHgH	4.05001	8778.81457	0.99963

Same phenomenon.

Exact Inference for the LR model for HG data I

- $2^{44} \approx 16 \times 10^{3\times 4} = 1.6 \times 10^{13}$ ways to pick responses $\boldsymbol{Y}$
- $54,600$ choices keeping sufficient statistics for intercept, sediment variables at observed value.
- Various algorithms make this job more tractable.

Exact Inference for the LR model for HG data II

U_2	count	U_2	count	U_2	count
582.0	6	586.6	1540	678.8	1144
583.0	28	587.1	990	679.3	2640
583.5	66	587.3	1760	679.5	264
583.8	132	587.6	880	679.6	1320
584.0	16	588.1	4620	679.8	528
584.5	308	589.1	2640	680.3	5720
584.8	616	676.0	24	680.6	2860
585.3	660	677.0	52	681.1	3960
585.5	176	677.5	264	681.3	2640
585.6	330	677.8	528	681.6	1320
585.8	352	678.0	24	682.1	8580
586.3	3080	678.5	572	683.1	3960

Statistic	P-value Exact
588.1	15560/54600=0.285

cf. normal approximate 2-sided p-value 0.545.

Logistic Regression Solution Using Optimization

Clarkson and Jennrich [1991] detect infinite $\hat{\beta}_j$.
Kolassa [1997] detects observations with extreme probabilities.

- Remove column for interest parameter from $\boldsymbol{X}$ to give $\boldsymbol{Z}$.
- Find vectors $\boldsymbol{r}$, $\boldsymbol{s} \in (\Re_{\geq 0})^d$
  - $s_i > 0$ implies $\hat{\pi}_i = 0$, $r_i > 0$ implies $\hat{\pi}_i = 1$.
  - with the maximal number of positive entries,
  - satisfying linear constraints $\boldsymbol{Mr} + \boldsymbol{Ns} = \mathbf{0}$ for some matrices $\boldsymbol{M}$, $\boldsymbol{N}$.
- Optimization task like linear programing.

Construction of a New Data Set without Infinite Estimates

Kolassa [1997] shows that removing observations with extreme probabilities leads to same conditional distribution for interest parameter.

- Remove $\boldsymbol{X}$ rows corresponding to positive entries in $\boldsymbol{r}$, $\boldsymbol{s}$.
- Remove redundant columns of $\boldsymbol{X}$.
- Recalculate sufficient statistics.

Use standard saddlepoint inference on reduced data set.

The Multinomial Regression Model

Multinomial Regression Model:

- N decisions to make
- For decision $j \in \{1, \ldots, N\}$, potential choice $D_j \in \{1, n_j\}$, $P[D_j = i] \propto \exp(\boldsymbol{x}_{ji}\boldsymbol{\beta})$

Log likelihood is $\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{\beta} - \sum_{j=1}^{N} \log(\sum_{i=1}^{n_j} \exp(\boldsymbol{x}_{ji}\boldsymbol{\beta}))$

- for $\boldsymbol{u} = \sum_{j=1}^{N} \boldsymbol{x}_{jD_j}$

As a special case of Conditional Logistic Regression I

- Design matrix for Conditional Logistic Regression

$$
\begin{array}{r}
\\
n_1 \text{ rows like this} \\
\\
n_2 \text{ rows like this} \\
\vdots \\
n_N \text{ rows like this}
\end{array}
\quad
\begin{array}{c}
N \text{ columns} \qquad d \text{ columns} \\
\begin{pmatrix}
1 & 0 & \dots & 0 & \boldsymbol{x}_{11} \\
1 & 0 & \dots & 0 & \boldsymbol{x}_{12} \\
\vdots & 0 & \dots & \vdots & \vdots \\
1 & 0 & \dots & 0 & \boldsymbol{x}_{1n_1} \\
0 & 1 & \dots & 0 & \boldsymbol{x}_{21} \\
\vdots & \vdots & \dots & \vdots & \vdots \\
0 & 1 & \dots & 0 & \boldsymbol{x}_{2n_2} \\
\vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & \dots & 1 & \boldsymbol{x}_{N1} \\
\vdots & \vdots & \dots & \vdots & \vdots \\
0 & 0 & \dots & 1 & \boldsymbol{x}_{Nn_N}
\end{pmatrix}
\end{array}
$$

As a special case of Conditional Logistic Regression II

- Response is column of $n_1 + \cdots + n_N$ zeros
  - except for 1 in slots corresponding to choice.
- Condition on sufficient statistics for the indicators.
- Kolassa [2016] applies Kolassa [1997] to this conditional logistic regression.

The Plan for Inference

Convert data set to one with non-monotone likelihood.

- Treat partial likelihood as product of multinomial likelihoods
  - Observations are not independent
  - Multiply probabilities because they are conditional.
- Express multinomial events as conditional logistic regression
- Analyze conditional logistic regression to identify subjects whose fitted probabilities are zero or one.
- Implies which multinomial subjects have selection probability either 0 or 1.
- This implies which survival subjects are guaranteed either to have or to fail to have the next event.
- Remove these from the Cox regression.
- Remove redundant covariates (coxph in R does this.)

Apply saddlepoint approximation.

Competitors

Run standard algorithm until it fails numerically.

- Advantage: Generally close to "best" answer.
- Disadvantage: Performance, stability not guaranteed.

Regularization Heinze and Schemper [2001]:

- Advantage: Numerically stable
- Disadvantage: Depends on regularization choice.

Identification of parameters estimated at infinity
Clarkson and Jennrich [2000]:

- Advantage: Solves a simpler optimization problem than recommended here.
- Disadvantage: Does not facilitate inference.

Bayesian approach

- Zhang and Kolassa [2008] provide matching prior argument.

Pluses and Minuses of the New Approach

- Pluses:
  - Subjects removed from the multinomial regression are removed consistently.
  - No tuning is required for regularization.
- Minuses:
  - Slow.
  - Conditioning argument removing nuisance parameters is only approximate, because it ignores censoring and sequential aspects of the problem.

Smallmouth Bass Result

Result removing extreme subjects:

```
                      coef  se(coef) Pr(>|z|)
PctWetland        -0.01753  0.01344  0.19207 (cf -0.0104
SedAVSH           -4.80506  1.47348  0.00111 from the
SedLOIH            5.57092  1.51177  0.00023 ``Let it Run''
SedMeHgH          -0.36040  0.54185  0.50597 approach)
SedAVSH:SedMeHgH   1.49744  0.80577  0.06311
SedLOIH:SedMeHgH        NA  0.00000       NA
```

- Different from the failed-to-converge result.
- Different from result dropping offending covariate
- NA for covariate indicates removal because of nonidentifiability.

Summary

The problem of infinite estimates in proportional hazards regression may be addressed by:

- Re-expressing Cox regression as a conditional logistic regression
- Using the method of Kolassa [1997] to remove subjects whose probability of event is estimated at 0 or 1
- Performing the standard analysis on the reduced data set.

The above procedure does not adversely affect test size.

Bibliography I

Douglas B. Clarkson and Robert I. Jennrich. Computing extended maximum likelihood estimates for linear parameter models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):417–426, 1991. ISSN 00359246. URL http://www.jstor.org/stable/2345752.

Douglas B. Clarkson and Robert I. Jennrich. Computing extended maximum likelihood estimates for cox proportional-hazards models. In F. Thomas Bruss and Lucien Le Cam, editors, *Game theory, optimal stopping, probability and statistics*, volume Volume 35 of *Lecture Notes–Monograph Series*, pages 205–217, Beachwood, OH, 2000. Institute of Mathematical Statistics. doi: 10.1214/lnms/1215089754. URL https://doi.org/10.1214/lnms/1215089754.

Bibliography II

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):pp. 187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

Georg Heinze and Michael Schemper. A solution to the problem of monotone likelihood in cox regression. *Biometrics*, 57(1):pp. 114–119, 2001. ISSN 0006341X. URL http://www.jstor.org/stable/2676848.

John E. Kolassa. Infinite parameter estimates in logistic regression, with application to approximate conditional inference. *Scandinavian Journal of Statistics*, 24(4):523–530, 1997. doi: 10.1111/1467-9469.00078. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9469.0

John Edward Kolassa. Inference in the presence of likelihood monotonicity for polytomous and logistic regression. *Advances in Pure Mathematics*, 6(5):331–341, 2016. doi: 10.4236/apm.2016.65024.

Bibliography III

Lopaka Lee. *NADA: Nondetects and Data Analysis for Environmental Data*, 2017. URL `https://CRAN.R-project.org/package=NADA`. R package version 1.6-1.

Juan Zhang and John Kolassa. A practical procedure to find matching priors for frequentist inference. *Communication in Statistics- Theory and Methods*, 42, 05 2008.

Naive Results

Results of a Proportional Hazards regression in R:

Variable	Estimate	SE	p
⋮			
Tilapia	-5.267×10^{-1}	1.217	0.665108
Walleye	4.846	1.283	0.000158
WhiteBass	$3.329\times10^{+1}$	4.572×10^{5}	0.999942
WhiteCrappie	-5.190×10^{-1}	1.733	0.764530
PctWetland	2.017×10^{-2}	4.816×10^{-3}	2.82×10^{-5}

- Problem: Proportional hazards algorithm fails to converge
- Partial Likelihood monotone in parameter for White Bass.

Logistic Regression Linear Constraints

Kolassa [1997] detects observations with extreme probabilities.

- Remove column for interest parameter from $\boldsymbol{X}$ to give $\boldsymbol{Z}$.
- Find vectors $\boldsymbol{r}$, $\boldsymbol{s} \in (\Re_{\geq 0})^d$ with the maximal number of positive entries, satisfying
  - $(\boldsymbol{T}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top - \mathbf{1})\boldsymbol{s} - \boldsymbol{T}^\top(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{r} = \mathbf{0}$
  - $(\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top)(\boldsymbol{s} - \boldsymbol{r}) = \mathbf{0}$

Logistic Regression Solution Using Linear Programming

- Find vectors $\boldsymbol{r}$, $\boldsymbol{s} \in (\Re_{\geq 0})^d$
  - with the maximal number of positive entries,
  - satisfying linear constraints $\boldsymbol{Mr} + \boldsymbol{Ns} = \mathbf{0}$ for some matrices $\boldsymbol{M}$, $\boldsymbol{N}$.
- Maximize $\sum_j c_j(r_j + s_j)$: Linear programming.
  - Start with $c_j = 1 \forall j$.
- This optimization is not exactly the same as maximizing the number of positive entries.
  - After finding $r_j + s_j > 0$, reset $c_j = 0$ and restart.
  - If LP is solved using simplex method, can hot start next iteration.