Proportional Hazards Analysis of Survival Data with Tied Survival Times: Theory and Best Practices

Muhammed Y. Idris (myi100@psu.edu)
Christopher Zorn (zorn@psu.edu)
Pennsylvania State University

Abstract

We examine the dominant approaches for dealing with tied survival times in the Cox proportional hazards model. The accuracy of the various approximations depends critically on both the relative frequency of tied survival times and on the degree of concentration of the data on small numbers of times. Our findings suggest a set of best practices for applied researchers for dealing with tied survival time data.

Tied Survival Times in Cox's (1972) Proportional Hazards Model

- Cox's (1972) proportional hazards model is:

$$h_i(t) = h_0(t)\ \exp\left(\mathbf{X}_i\beta\right)$$

- In the absence of tied survival times, the log-partial-likelihood of Cox's model is equal to:

$$\ln L = \sum_{j=1}^{J}\left\{\sum_{i\in D_j}\mathbf{X}_i\beta - d_j \ln\left[\sum_{r\in R_j}\exp(\mathbf{X}_r\beta)\right]\right\}.$$

  where J is the set of all survival times, D_j is the set of d_j observations failing at time t_j, and R_j denotes the set of observations "at risk" for failure at t_j.
- Cox's original model requires that all survival times be distinct; i.e., that no two observations experience the event of interest (or censoring) at the same time (are "tied").
- This is because the Cox model uses only the composition of the risk set and the relative ordering of the event times to inform the partial likelihood.

Methods for Handling Tied Survival Times

There are three widely-implemented methods for dealing with ties in the Cox context:

- Breslow / Peto (1972):

$$\ln L_{\text{Breslow}}(\beta) = \sum_{j=1}^{J}\sum_{i\in D_j}\left\{\mathbf{X}_i\beta - \ln\left[\sum_{\ell\in R_j}\exp(\mathbf{X}_\ell\beta)\right]\right\}$$

- Efron (1974):

$$\ln L_{\text{Efron}}(\beta) = \sum_{j=1}^{J}\sum_{i\in D_j}\left\{\mathbf{X}_i\beta - \frac{1}{d_j}\sum_{k=1}^{d_j-1}\ln\left[\sum_{\ell\in R_j}\exp(\mathbf{X}_\ell\beta) - k\left(\frac{1}{d_j}\sum_{\ell\in D_j}\exp(\mathbf{X}_\ell\beta)\right)\right]\right\}$$

- Exact Partial Likelihood (Cox 1972):

$$\ln L_{\text{Exact}}(\beta) = \sum_{j=1}^{J}\left\{\sum_{i\in R_j}\delta_{ij}(\mathbf{X}_i\beta) - \ln[f(r_j, d_j)]\right\},\ \text{where}$$

$$f(r,d) = g(r-1,d) + g(r-1,d-1)\exp(\mathbf{X}_k\beta),$$

$$k = r\text{th observation in } R_j,$$

$$r_j = \text{cardinality of } R_j,\ \text{and}$$

$$g(r,d) = \begin{cases} 0 \text{ if } r < d, \\ 1 \text{ if } d = 0 \end{cases}$$

The Prevalence and Dispersion of Ties

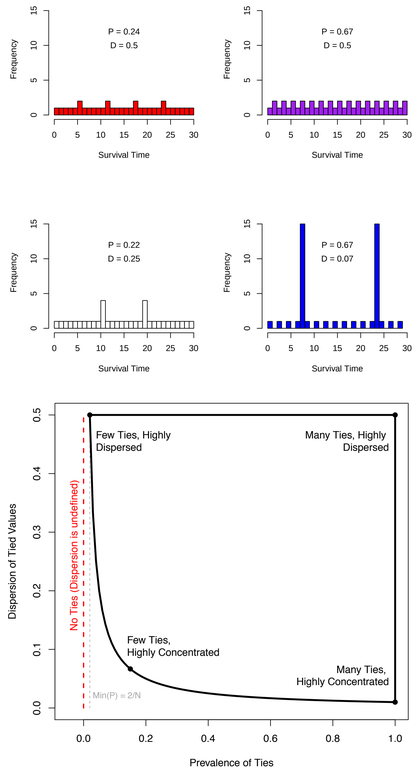
- Let N denote the number of (non-censored) events in the data, and N_t the number of observations who share a survival time with at least one other observation; similarly, define J as the total number of unique survival times in the data, and J_t as the corresponding count of "shared" survival times.
- The **prevalence** (P) of ties is simply the proportion of observations with shared survival times:

$$P = \frac{N_t}{N} \in [0, 1].$$

- The **dispersion** (D) of tied survival times reflects the extent to which ties are grouped in the data:

$$D = \frac{J_t}{N_t} \in \left[\frac{1}{N_t}, 0.5\right].$$

- Higher values of P denote data with greater relative numbers of tied survival times.
- Survival data with higher values of D contain relatively few observations for each (tied) survival time; lower dispersion corresponds to circumstances where tied observations are "concentrated" in a few shared survival times.



Simulation Procedure

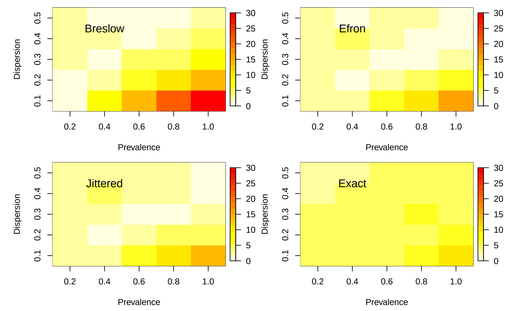
Exponential survival times:

$$T_i = \text{Exponential}(0 + 1.0X_i) \qquad N \in \{60, 600, 6000\}$$
$$X_i \sim i.i.d.\ \text{Bernoulli}(0.5) \qquad P \in [0.2, 1]$$
$$D \in [0.1, 0.5]$$

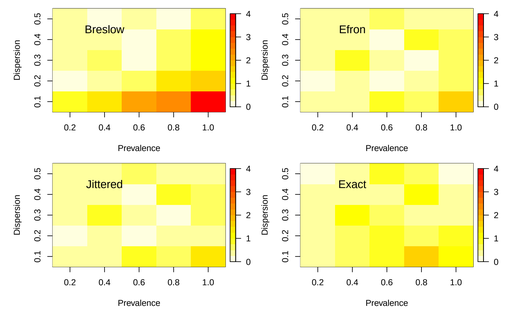
- "Jitter": adding $\lambda = 1000$ exponential noise to each observation.
- $K = 1000$ replications for each permutation of $\{N, P, D\}$
- Percent Absolute Bias$_{N,P,D} = \left|\frac{\sum_{k=1}^{K}\hat{\beta}_k - 1}{1000}\right| \times 100$

Simulation Results

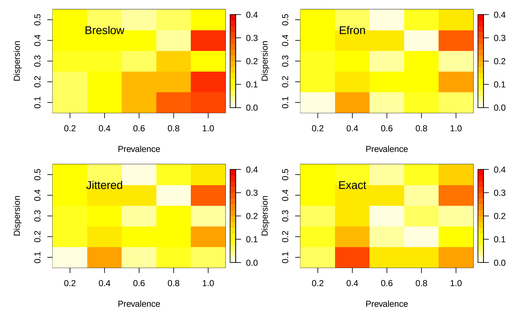
Percent Absolute Bias in $\hat{\beta}$, $N = 60$



Percent Absolute Bias in $\hat{\beta}$, $N = 600$



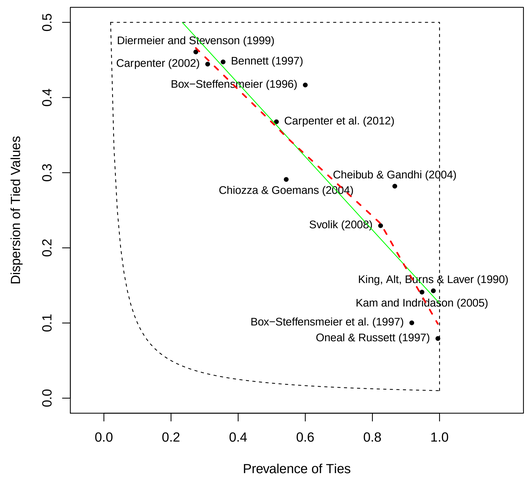
Percent Absolute Bias in $\hat{\beta}$, $N = 6000$



Existing Studies

- Existing studies in political science are characterized by high tie prevalence and varying levels of dispersion.
- In general, higher tie prevalence is associated with lower dispersion / greater concentration of ties among a few values.

Values of P and D for Existing Studies



Implications and Future Work

Implications

- All methods perform equally well when $P < 0.5$.
- For $P > 0.5$, all methods perform equally well if dispersion is high (i.e., when there are large numbers of survival times with relatively few tied observations per time).
- If P is high and D is low, the exact method is strongly recommended.
- For all approximations, the degree of bias decreases in the sample size at roughly $O(1/N)$.

Future Work

1. Examine standard error estimates and effective coverage rates.
2. Assess sensitivity to "asymmetrically" tied data.
3. Develop simple R and Stata routines to aid model selection.

References

Cox, D.R. 1972. "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B* 34(2): 187-220.
Efron, Bradley (1974). "The Efficiency of Cox's Likelihood Function for Censored Data." *Journal of the American Statistical Association* 72 (359): 557-565.
Hertz-Picciotto, Irva, and Beverly Rockhill. 1997. "Validity and Efficiency of Approximation Methods for Tied Survival Times in Cox Regression." *Biometrics* 53(3):1151-1156.

Poster prepared for PolMeth XXX, July 18-20, 2013, Charlottesville, VA. This work is ongoing; comments are very, very welcome.

This project is supported in part by the PSU College of Liberal Arts and the National Science Foundation.