Analysis of Time to Event Data: Beyond Semi-parametric Models

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics,
University of Washington

Joint with M. Nason, K. Rudser, S. Devlin, M. LeBlanc

February 25, 2014

Classification of Statistical Methods

• Two approaches to data analysis (Breiman)
  – Model based (e.g., regression)
  – Algorithmic (e.g., trees, neural nets)

• Alternative classifications:
  – Variable importance vs prediction
  – Model questions vs model data

• Additional classifications
  – Regression vs ANOVA vs Sample survey
  – Distribution-free vs (Semi)Parametric
  – Frequentist vs Bayesian

Hypothetical Example: Setting

• Consider survival with a particular treatment used in renal dialysis patients
• Extract data from registry of dialysis patients
• To ensure quality, only use data after 1995
  – Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
  – Prevalent cases in 1995: Data from 1995 - 2002
    • Incident in 1994: Information about 2nd – 9th year
    • Incident in 1993: Information about 3rd – 10th year
    • ... 
    • Incident in 1988: Information about 8th – 15th year

Two types of people in the world:

Those who dichotomize everything,

and

Those who don't.
How should we analyze this data?

- Issues with incomplete (missing) data
- Some individuals die before entry to study
  - “Left entry”
- Some individuals are still alive at end of study
  - “Right censoring”

“We have censored data, so we have to use proportional hazards regression.”

---

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  A: 2.07 (logrank P = .0018)
  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)
  D: 0.48 (logrank P = .0018)

  - Lifelines:
    - 50-50? Ask the audience? Call a friend?

---

Who Wants To Be A Millionaire?

- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)
Who Wants To Be A Millionaire?

• Proportional hazards analysis estimates a Treatment : Control hazard ratio of
  
  B: 1.13 (logrank P = .0018)

• Proportional Hazards analysis estimates higher risk of death (on average) for treatment group than control group

Right Censored Data

• Notation:

  Unobserved:
  True times to event: \( \{ T_1, T_2, \ldots, T_n \} \sim F = 1 - S \)
  Censoring Times: \( \{ C_1, C_2, \ldots, C_n \} \sim G \)

  Observed data:
  Observation Times: \( Y_i = \min(T_i, C_i) \)
  Event indicators: \( D_i = \begin{cases} 1 & \text{if } Y_i = T_i \\ 0 & \text{otherwise} \end{cases} \)

Kaplan-Meier Estimates

• Distribution-free estimates of probability of survival

  Ordered distinct observation times: \( 0 = t_0 \leq t_1 \leq \cdots \leq t_k \)

  Time interval: \( (t_{j-1}, t_j] \)
  Number at risk at \( t_j \): \( N_j \)
  Number of events at \( t_j \): \( D_j \)
  Hazard at \( t_j \): \( D_j / N_j \)
  Conditional probability of survival in interval:
  \( \Pr( T > t_j \mid T > t_{j-1}) = 1 - \frac{D_j}{N_j} \)

Cox PH Partial Likelihood

\( \lambda(t \mid X) = \lambda_0(t) \exp\{X \beta\} \quad \Leftrightarrow \quad S(t \mid X) = S_0(t)^{\exp\{X \beta\}} \)

\[ L(\beta) = \prod_{i=1}^{n} \left[ \frac{\exp\{X_i \beta\}}{\sum_{j: j \geq T_i} \exp\{X_j \beta\}} \right]^{D_i} \]

\[ \log L(\beta) = \sum_{i=1}^{n} D_i \left[ X_i \beta - \log \sum_{j: j \geq T_i} \exp\{X_j \beta\} \right] \]
Cox PH Partial Likelihood Based Score

• Depends on number at risk at each time
  – Number at risk a function of survival and censoring

\[
U(\beta) = \frac{\partial}{\partial \beta} \log L(\beta) = \sum_{i=1}^{n} D_i \left( X_i - \sum_{j \in J_i \cap T_i} X_j \exp\{X_j \beta\} \right)
\]

\[
= \sum_{t} \left[ d_{it} - \frac{n_{it} \exp(\beta)}{n_{0t} + n_{it} \exp(\beta)} (d_{0t} + d_{it}) \right]
\]

\[
= \sum_{t} \frac{n_{0t} n_{it}}{n_{0t} + n_{it}} \left[ \hat{\lambda}_{it} - e^\beta \hat{\lambda}_{it} \right]
\]

So What?

• The true survival curves had “crossing hazards”
  – Effect modification by time at risk

• The weighting using the risk sets made no scientific sense
  – The weights were efficient for proportional hazards
  – Statistical precision to estimate a meaningless quantity is meaningless
  – (This highlights the scientific importance of describing the censoring distribution)

• This sort of qualitatively misleading result cannot happen in a RCT, because we do not have “left entry”
  – But a problem remains
Phase III Clinical Trial Results

Event Free Survival (All Regions Combined)

Proportional vs Nonprop Hazards

- Survival curves with proportional hazards, nonproportional ("diverging") hazards with late differences, and nonproportional ("crossing") hazards with early differences (from D. Gillen)

Weighted Logrank Tests

- Fleming and Harrington $G^{\rho\gamma}$ family of weighted logrank tests

$$U(\beta) = \sum_t w_t \left[ d_{it} - \frac{n_{it} e^{\beta_0}}{n_{0t} + n_{1t} e^{\beta_0}} (d_{0t} + d_{1t}) \right]$$

$$= \sum_t w_t \frac{n_{0t} n_{1t}}{n_{0t} + n_{1t}} [\hat{\lambda}_{it} - e^{\beta_0} \hat{\lambda}_{0t}]$$

$$= \sum_t (S(t))^\rho (1 - S(t))^\gamma \frac{n_{0t} n_{1t}}{n_{0t} + n_{1t}} [\hat{\lambda}_{it} - e^{\beta_0} \hat{\lambda}_{0t}]$$

Kaplan Meier Survival Curve

Erlotinib/Gem vs Placebo/Gem (504 deaths)

Erlotinib/Gem: median OS: 6.37 mos (95% CI: 5.84 to 7.33)
Placebo/Gem: median OS: 5.94 mos (95% CI: 5.09 to 6.70)
HR: 0.811 (95% CI: 0.68 to 0.97)
Censoring Distributions

- In a randomized clinical trial (RCT), the censoring distribution is ideally determined by the pattern of accrual
  - Ideal is only “administrative censoring” due to continued survival at time of data analysis
- We can consider accrual over a 3 year period with additional follow-up for one year after accrual stops

Weighted Logrank w/ Nonprop Hazards

- Effect of censoring distribution on inference in presence of late diverging hazard functions

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Uniform</th>
<th>Early</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G_{\alpha} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\alpha} ) (generalized Wilcoxon)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\lambda} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\eta} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\tau} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\kappa} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accrual Pattern</th>
<th>Uniform</th>
<th>Early</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportional/Constant Difference Hazards</td>
<td>( \lambda_{\alpha} ) (Logrank)</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\alpha} ) (generalized Wilcoxon)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\lambda} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\eta} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\tau} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>( G_{\kappa} )</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(Estimated hazard ratio)</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Non-proportional/Non-Constant Difference Hazards</td>
<td>( \lambda_{\alpha} ) (Logrank)</td>
<td>1.00</td>
<td>1.13</td>
</tr>
<tr>
<td>( G_{\alpha} ) (generalized Wilcoxon)</td>
<td>1.00</td>
<td>1.13</td>
<td>0.84</td>
</tr>
<tr>
<td>( G_{\lambda} )</td>
<td>1.00</td>
<td>1.11</td>
<td>0.86</td>
</tr>
<tr>
<td>( G_{\eta} )</td>
<td>1.00</td>
<td>1.08</td>
<td>0.87</td>
</tr>
<tr>
<td>( G_{\tau} )</td>
<td>1.00</td>
<td>1.09</td>
<td>0.87</td>
</tr>
<tr>
<td>( G_{\kappa} )</td>
<td>1.00</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>(Estimated hazard ratio)</td>
<td>0.73</td>
<td>0.69</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Power of Weighted Logrank Tests

- The problem is much greater when we might have crossing survival curves
- Example from Logan, Klein, & Zhang (Biometrics, 2008)
  - Disease free survival after bone marrow transplantation in follicular lymphoma

Crossing Survival Curves

- Figure 1: Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.

- Tests during entire support of censoring distribution
  - Logrank statistic (proportional hazards regression)
  - Weighted logrank
- Tests focusing on hypothesized time $t_0$ near or after crossing of curves
  - KM curves at a single point in time $t$ : $Z_{\text{KM}}(t)$
  - Pseudovalue regression at selected points beyond $t_0$ : $\chi^2_{\text{PSV}}(t)$
  - Combination tests comparing KM at time $t_0$ and weighted logrank beyond time $t_0$
    - Linear combination of test statistics: $Z_{\text{KL}}(t_0)$ and $Z_{\text{WL}}(t_0)$
    - Quadratic combination of test statistics: $\chi^2(t_0)$

Logan, Klein, Zhang (2008): Scenarios

- Null hypothesis curves
- Alternative hypothesis, scenario E
- Alternative hypothesis, scenario F
- Alternative hypothesis, scenario G
- Alternative hypothesis, scenario H

Logan, Klein, Zhang (2008): Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Method</th>
<th>Equation</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_{\text{KM}}(24)$</td>
<td>(1)</td>
<td>62.4</td>
<td>18.5</td>
<td>21.1</td>
<td>4.7</td>
<td>21.8</td>
<td></td>
</tr>
<tr>
<td>$Z_{\text{KM}}(48)$</td>
<td>(1)</td>
<td>70.1</td>
<td>32.9</td>
<td>65.1</td>
<td>21.5</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>$Z_{\text{KM}}(72)$</td>
<td>(1)</td>
<td>71.2</td>
<td>44.5</td>
<td>85.1</td>
<td>46.1</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>$Z_{\text{KM}}(t_0)$</td>
<td>(2)</td>
<td>75.8</td>
<td>35.0</td>
<td>66.3</td>
<td>20.3</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>$\chi^2_{\text{PSV}}(t_0)$</td>
<td>(3)</td>
<td>74.8</td>
<td>32.0</td>
<td>61.2</td>
<td>16.4</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>$Z_{\text{WL}}(t_0)$</td>
<td>(4)</td>
<td>20.7</td>
<td>36.5</td>
<td>85.4</td>
<td>71.7</td>
<td>82.6</td>
<td></td>
</tr>
<tr>
<td>$Z_{\text{KL}}(t_0)$</td>
<td>(5)</td>
<td>74.7</td>
<td>43.9</td>
<td>84.1</td>
<td>43.4</td>
<td>23.0</td>
<td></td>
</tr>
<tr>
<td>$\chi^2(t_0)$</td>
<td>(6)</td>
<td>70.8</td>
<td>34.5</td>
<td>58.8</td>
<td>26.0</td>
<td>10.7</td>
<td></td>
</tr>
<tr>
<td>Log rank</td>
<td>(7)</td>
<td>67.2</td>
<td>36.7</td>
<td>83.1</td>
<td>61.1</td>
<td>81.1</td>
<td></td>
</tr>
<tr>
<td>Weighted log rank</td>
<td></td>
<td>78.9</td>
<td>28.9</td>
<td>47.0</td>
<td>8.9</td>
<td>22.2</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0, \gamma = 1$</td>
<td></td>
<td>64.7</td>
<td>48.7</td>
<td>93.8</td>
<td>70.9</td>
<td>64.6</td>
<td></td>
</tr>
</tbody>
</table>


- What is the type I error?
- What is the null hypothesis
  - Strong vs weak?
  - Noncrossing vs crossing?
- There seems to be an implicit assumption that failure to reject a difference at time $t_0$ can be interpreted as equivalence
  - If that presumption were valid, then a group with the higher hazard after $t_0$ would have worse survival in the end
Comparing Survival Curves: $S(t) \text{ vs } t$

- Commonly used summary measures can be seen in a plot of survival curves
  - Difference in survival at $t_0$
    - Vertical separation at $t_0$
  - Difference in quantiles
    - Horizontal separation at $p$
  - Difference in means
    - Area between curves
  - Hazard
    - Slope divided by height of the curve
    - (not exactly clear)

More Extensive Scenarios

- The fear is that a naïve user would believe that crossing estimated KM curves and a significantly higher hazard after $t_0$ would correctly identify the better treatment
- In either of the crossing hazards scenarios I can get the probability of making the “wrong” decision arbitrarily close to 50%
  - I can make the probability of a statistically nonsignificant difference in survival at $t_0$ arbitrarily close to 95%
  - I can make the probability of crossing estimated KM curves arbitrarily close to 50%
  - I can make the probability of detecting a higher hazard in the truly “better” group arbitrarily close to 100%

Which Treatment Would You Want?
Motivation from Clinical Trials

- Difference or ratio of
  - Means (arithmetic, geometric, harmonic, ...)
  - Medians (or other quantiles)
  - Proportion exceeding some threshold
  - Odds of exceeding some threshold
  - Time averaged hazard function (instantaneous risk)
  - ...
- Possibly adjusting for covariates
- Distribution-free estimates
  - Avoid assumptions more detailed than the question you are trying to answer
- Frequentist or Bayesian

Criteria for Summary Measure

- We choose some summary measure of the probability distribution according to the following criteria (in order of importance)
  - Scientifically (clinically) relevant
    - Also reflects current state of knowledge
  - Is likely to vary across levels of the factor of interest
    - Ability to detect variety of changes
  - Statistical precision
    - Only relevant if all other things are equal
- In order to avoid multiple comparison issues, the summary measure and analytic model should be prespecified

Typical Observational Setting

- Eligibility criteria for inclusion of observational data
- Some “primary outcome” measurement
  - A “response variable” in regression
- Groups that are homogeneous with respect to the level of some factor(s)
  - Predictor of interest
  - Effect modifiers
  - Confounders
  - Precision variables

But What About Survival Data?

- Time to event data is typically right censored
- We know the exact time of event for some subjects
- For others, we merely know some time that they were observed without an event
  - Clearly nonignorable missing data
Where Am I Going?

• Censored data results from a particular choice of sampling scheme
• Usually such a sampling scheme is necessary due to logistical constraints
• There is nothing inherent in the mere presence of censored data that need alter the question which is deemed scientifically most important

Common Practice

• The overwhelming majority of statistical inference is based on means
  – Means of continuous random variables
    • t test, linear regression
  – Proportions (means of binary random variables)
    • chi square test (t test)
  – Rates (means) for count data
    • Poisson analyses
• I want to be able to do this with censored data in regression settings as well

Statistical Models for Survival: Options

• Parametric models
  – Weibull, lognormal, etc.
• Semiparametric models
  – Proportional hazards, etc.
• Distribution-free
  – Weighted rank tests: logrank, Wilcoxon, etc.
  – Comparison of Kaplan-Meier estimates

Parametric Models

• \( F \) is known up to some finite dimensional parameter vectors
\[
F(t) = \Psi(t, \Phi)
\]
where:
\[
\Psi(\cdot, \cdot) \quad \text{has known form}
\]
\[
\Phi \quad \text{is finite dimensional and unknown}
\]
Parametric Inference

- Parametric inference generally proceeds through likelihood methods
  - MLE found by Newton-Raphson iteration
  - Asymptotic distributions from theory of regular problems

\[
L(\Phi; T, D) = \prod_{i=1}^{n} \left( f(T_i; \Phi) (1 - G(T_i)) \right)^{D_i} \left( S(T_i; \Phi) g(T_i) \right)^{1-D_i}
\]

Parametric Models: Issues

- Advantages
  - Can estimate any of the summary measures
  - Can handle sparse data
- Disadvantages
  - Not robust to other distributions
    - Likelihood depends explicitly on the shape of the distribution through the censored observations
    - Parametric estimates do not generally have easy distribution-free interpretation
      - E.g., lognormal model is not particularly robust
  - Little reason to suggest particular distribution
    - But motivation does exist for Weibull and Gamma

Semiparametric Models

- Exact form of within group distributions are unknown, but related to each other by some finite dimensional parameter vector
  - Full inference only for comparing distributions
  - One group's distn can be found from another group's and a finite dimensional parameter
  - (Most often: Distributions equal under $H_0$)

(My definition of semiparametric models is a little stronger than some statisticians', but agrees with commonly used semiparametric survival models)

Semiparametric Models: Notation

For group $k$: $F_k(t) = \Psi(t; \Phi_k)$

where:
- $\Psi(\cdot; \cdot)$ has unknown form (in $t$)
- $\Phi_0 = \bar{\Phi}$ for identifiability of $\Psi(\cdot; \cdot)$
- $\Phi_k$ is finite dimensional and unknown ( estimable by comparing two or more groups)
Semiparametric Survival Models

Accel failure: \( F_{\hat{\theta}}(t) = F_0(t \theta_k) \)

Prop hzd: \( S_{\hat{\theta}}(t) = [S_0(t)]^{\theta_k} \)

where in a regression problem

\[ g(\theta_k) = \hat{X}_i \hat{\beta} \]

for some link function \( g() \)

Semiparametric Models: Issues

- Advantages
  - Can handle sparse data
  - More robust than any single parametric model

- Disadvantages
  - Not easily interpreted when semiparametric model does not hold
  - Little reason to suggest a given risk factor would affect distribution in only one way
  - Generally inference possible only with “fundamental” functional

Problem with (Semi)parametrics

- Many mechanisms would seem to make it likely that the problems in which a fully parametric model or even a semiparametric model is correct constitute a set of measure zero
- Treatments are often directed to outliers
- Treatments are often only effective in subsets
- Factors affect rates; outcomes measure cumulative effects

Inflammatory Assertion

- (Semi)parametric models are not typically in keeping with the state of knowledge as an experiment is being conducted
- The assumptions are more detailed than the hypothesis being tested, e.g.,
  - Question: How does the intervention affect the first moment of the probability distribution?
  - Assumption: We know how the intervention affects the 2nd, 3rd, ..., \( \infty \) central moments of the probability distribution.
Robustness Results

From Jon Wellner’s notes for STAT 581:

7 Limit theory for the statistical agnostic
In the preceding sections we studied the limit behavior of the MLE \( \hat{\theta}_n \) (or ELE \( \tilde{\theta}_n \)) under the assumption that the model \( P \) is true; i.e. assuming that the data \( X_1, \ldots, X_n \) were governed by a probability distribution \( \theta \in P \). Frequently however we are in the position of not being at all sure that the true \( P \) is an element of the model \( P \), and it is natural to ask about the asymptotic behavior of \( \hat{\theta}_n \) (or \( \tilde{\theta}_n \)) when, in fact, \( P \not\in P \). This point of view is implicit in the robustness literature, and especially in the work of Huber (1964), (1967), and White (1982).

if we suppose that the infimum is achieved at \( \theta = \theta_0(P) \). Thus it is natural to expect that (under reasonable additional conditions)

\[
\hat{\theta}_n = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} p(X_i; \theta)
\]

\[
\rightarrow \arg\max_{\theta} \left( E \log p(X_i) - K(P, P_0) \right) = \theta_0(P) = \arg\min_{\theta \in \Theta} K(P, P_0).
\]

The Problem

• Incorrect assumptions can lead to incorrect statistical inference
• Efficiency is overstated
• Precision of estimators can be over- or understated
  – Hypothesis tests do not attain the nominal size
• Hypothesis tests can be inconsistent
  – Even an infinite sample size may not detect the alternative
• Interpretation of estimators can be wrong
• Under the strong null, estimators can be inconsistent across studies

Foundational Issues: Null

• Which null hypothesis should we test?
  • Strong: Intervention has no effect whatsoever
    – Permutation tests can be used
      \[ H_0 : F(t) = G(t), \forall t \]
  • Weak: Intervention has no effect on specified summary measure of the distribution
    – “Permutation tests considered harmful”
      \[ H_0 : \theta = \theta_0 \]

(Foundational Issues: Alternative)

• What do we mean by an alternative?
  – (How do we model mean-variance relationship?)
• Parametric counterfactual?
  – A hypothetical beneficial treatment, would have specified shape
• Empirical?
  – Observed data is in someway misleading re true shape
  – With finite dimensional nuisance parameters, we would use best estimate of the nuisance parameter under the alternative
  – Distribution-free approaches to estimate infinite dimensional nuisance parameter
    • Bootstrap tilting (assumes same support as observations)
    • Empirical likelihood (assumes same support as observations)
    • Data augmentation with above
Example: Semiparametric ROC Curves

- ROC curve: Sensitivity vs 1 – Specificity as a function of criteria used in medical diagnostic testing

Diagnostic test based on $T_i > c$

Sensitivity: Positive in Disease

$S_{\text{Sens}} = \Pr(T_i > c \mid D) = S_H(c)$

Specificity: Negative in Health

$S_{\text{Spec}} = 1 - \Pr(T_i > c \mid H) = 1 - S_H(c)$

Receiver Operating Characteristic (ROC) curve

$ROC(t) = S_H^{-1}(S_H^{-1}(t))$

Example: Power Family of ROC Curves

- A semiparametric model exhibiting proportional hazards

$ROC(t) = S_H^{-1}(S_H^{-1}(t)) = t^\theta \Leftrightarrow S_D(y) = [S_H(y)]^\theta$

- Possible analytic models
  - Parametric: Exponential (or Weibull) regression
  - Semiparametric Proportional Hazards
    - Cox PH estimating equation: Semiparametric efficient
    - ROC-GLM (Pepe, 2000)

Inference With Proportional Hazards

- Efficient parametric estimating equation would lead to different estimates under the strong null $H_0: ROC_1(t) = ROC_2(t)$
  - Semiparametric estimating equations are ok; Cox most efficient

Table 2: Comparison of exponential, Cox proportional hazards and ROC-GLM estimates of $\theta$ when the true ROC curve is a power curve and $\theta = -1.2$. Estimates were calculated as the mean and standard deviation of 3,000 replications.

<table>
<thead>
<tr>
<th>Regression Type</th>
<th>Exponential</th>
<th>Normal</th>
<th>LogNorm.</th>
<th>Exponential</th>
<th>Normal</th>
<th>LogNorm.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_{\text{Exp}}$</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
</tr>
<tr>
<td>$\theta_{\text{Normal}}$</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
</tr>
<tr>
<td>$\theta_{\text{LogNorm.}}$</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
<td>-0.126 (0.089)</td>
<td>-0.124 (0.120)</td>
<td>-0.124 (0.109)</td>
</tr>
</tbody>
</table>
Inference With Non Proportional Hazards

- Semiparametric efficient Cox PH estimating equation would lead to different estimates under the strong null $H_0: \text{ROC}_1(t) = \text{ROC}_2(t)$
  - ROC-GLM is consistent under strong null

Table 3: Comparison of exponential, Cox proportional hazards and ROC-GLM estimates of $\theta$ when the true ROC curve is binormal. Estimates were calculated as the mean and standard deviation of 5,000 replications.

<table>
<thead>
<tr>
<th>Regression Type</th>
<th>$F_0$</th>
<th>300:300</th>
<th>100:500</th>
<th>500:300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Exp</td>
<td>-1.197 (0.002)</td>
<td>-1.281 (0.155)</td>
<td>-1.089 (0.161)</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>-0.033 (0.004)</td>
<td>-0.033 (0.004)</td>
<td>-0.033 (0.007)</td>
</tr>
<tr>
<td></td>
<td>Log/Norm</td>
<td>2.648 (0.294)</td>
<td>2.783 (0.441)</td>
<td>2.600 (0.280)</td>
</tr>
<tr>
<td>Cox PH</td>
<td></td>
<td>-0.909 (0.052)</td>
<td>-0.986 (0.095)</td>
<td>-0.928 (0.139)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.929 (0.060)</td>
<td>-0.886 (0.083)</td>
<td>-1.055 (0.140)</td>
</tr>
<tr>
<td>ROC-GLM $T_{rcc}$</td>
<td></td>
<td>-0.792 (0.060)</td>
<td>-0.728 (0.088)</td>
<td>-0.788 (0.150)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.792 (0.064)</td>
<td>-0.728 (0.087)</td>
<td>-0.788 (0.150)</td>
</tr>
<tr>
<td>ROC-GLM $T_{ps}$</td>
<td></td>
<td>-0.741 (0.090)</td>
<td>-0.741 (0.090)</td>
<td>-0.741 (0.155)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.741 (0.090)</td>
<td>-0.718 (0.097)</td>
<td>-0.717 (0.060)</td>
</tr>
<tr>
<td>ROC-GLM $T_{rs}$</td>
<td></td>
<td>-0.729 (0.095)</td>
<td>-0.728 (0.086)</td>
<td>-0.784 (0.154)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.729 (0.090)</td>
<td>-0.728 (0.097)</td>
<td>-0.784 (0.150)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.729 (0.092)</td>
<td>-0.728 (0.085)</td>
<td>-0.784 (0.150)</td>
</tr>
</tbody>
</table>

A Non-Solution: Model Checking

- Frequentist issues
  - Over fitting
  - Low power
  - Multiple testing of strong null

- Bayesian issues
  - No statistical foundations

Distribution-free Models

- Form of F is completely arbitrary and unknown within groups
- The summary measure measuring factor effect is just some difference between distributions
- The summary measure is estimated nonparametrically
  - (preferably within groups and then compared across groups)

Typical Chronology

- Parametric models
  - Preferably justifiable scientifically
    - Bernoulli, Poisson, normal, lognormal, exponential, gamma, Weibull
- Evaluate semiparametric behavior of estimators
  - Scientifically
    - Designed studies evaluating behavior
    - Families of parametric distributions for skewness, kurtosis...
  - Evolutionarily
    - What we haven’t been burned by
- Distribution-free use
Example: Linear Regression

- **Parametric**
  - i.i.d. normal errors
- **Semiparametric**
  - CLT for i.i.d. errors under regular conditions
- **Distribution-free (nearly)**
  - Gauss-Markov: BLUE
  - Weighted and general LS
  - Smoothing residuals to avoid systematic errors

Data Modeling Approach

- Least squares regression
  \[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]
  \[ \hat{\beta}_1 = \frac{\sum (X_i - \bar{X})y_i}{\sum (X_i - \bar{X})^2} \]
  \[ \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \]
  \[ \beta_i = \frac{\sum (X_i - \bar{X})\mu_i}{\sum (X_i - \bar{X})^2} \]

Censored Data Problem

- It is hard to “break” linear regression, because every complete observation estimates its mean unbiasedly and consistently
- With censored data, we need to find some other consistent estimator of the functional of interest
- Use Kaplan-Meier estimated survival distribution for each individual

Distribution-free Regression

- Purpose of regression models
  - Borrow data to address sparse data
  - Compute contrasts across groups to assess associations between variables
- General Goal: Separate these tasks
  - Blend of regression, ANOVA, sample survey
  - Borrow information with distribution-free predictive models
  - Regression contrasts on “imputed” functionals to assess associations (variable importance)
**Distribution-free Approach**

- Generalization using predictive model

Borrow info to obtain $\hat{\theta}_i$

Contrast $\hat{\beta}_i = \sum c_i \hat{\theta}_i$ (\(\sum c_i = 0\))

- Inference by resampling entire procedure

---

**Regression Trees**

- Recursive partitioning
  - Divide sample into two groups that differ the most
    - Criteria for division based on one of the variables
    - Two sample test(s) of strong null (your choice)
      - “Cocktail” of weighted log rank tests
    - Repeat until no further divisions possible
    - Each “leaf” is then regarded as a homogeneous group

- Bagging, boosting $\Rightarrow$ smoother models

---

**Borrowing Information: Approaches**

- (Semi)parametric models with added flexibility
  - Splines, time varying parameters

- Separate quantile regressions

- Kernel smoothers

- Regression trees

---

**CARTscans**

- Graphical presentation of complex predictive models
  - Analogy with CT scans

- Goal: Visual assessment of variable importance

- Two dimensional slices through higher dimensional space four variables at a time

- An array of plots arranged to facilitate reconstruction of data

- Colors used to reflect summary measure of response variable
  - (Colors that I can see)
Trees and CARTscans

Regression Tree on FEV Data

CARTscan of FEV tree

Example: FEV and Smoking

- FEV data on 654 healthy children
  - Ages 3 – 19
  - Heights appropriate for age
  - Both sexes
  - 65 smokers
- Analyze mean FEV across predictor groups
CARTscan of Regression Model

Variable Importance

- Distribution-free regression: Least squares
  - Data modeling approach:
    \[ Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i \]
  - Distribution-free approach
    \[ \hat{\beta}_i = \frac{\sum (X_i - \bar{X})Y_i}{\sum (X_i - \bar{X})^2} \]
- Distribution-free approach
  Borrow info to obtain
  Contrast
  \[ \hat{\theta}_i = \sum w_i \hat{\theta}_i \]

Distribution-free Summary Measures

- Estimate functionals within each leaf

Using Kaplan - Meier survival estimate \( \hat{S}(t) \)

Mean:
\[ \hat{\theta} = \int_0^\infty \hat{S}(u) du \]

Median:
\[ \hat{\theta} = \hat{S}^{-1}(0.5) \]

Proportion above threshold:
\[ \hat{\theta} = \hat{S}(a) \]

Proportion above threshold:
\[ \hat{\theta} = \hat{S}(a) \]

Weighted average of hazard:
\[ \hat{\theta} = \int_0^\infty w(u) \hat{\lambda}(u) du \]

Distribution-free Summary Measures

- Depending on the censoring scheme, not all summary measures are estimable
  - The support of the censoring distribution may preclude estimation of the mean and some quantiles
  - Can instead use the mean of the truncated distribution
    - "Average increase in days alive during first 5 years"

Mean of truncated distribution:
\[ \hat{\theta} = \int_0^a \hat{S}(u) du \]
“Proof of Concept”

- Compare three predictive models
  - Regression tree, Buckley-James, PH
- Compare three probability models
  - Null hazard, null median, PH
- Evaluate contrasts of functionals
  - Restricted mean, 75th percentile, median
  - RMSE, coverage of 90% CI
- Setting
  - Moderately correlated covariate distribution

Features of Null Hazard Curves

True Distributions

- True Means
- True Medians
### RMSE: Null Hazard

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Approach</th>
<th>No Censoring Bilirubin</th>
<th>No Censoring Protime</th>
<th>20% Censoring Bilirubin</th>
<th>20% Censoring Protime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Tree</td>
<td>0.044</td>
<td>0.143</td>
<td>0.038</td>
<td>0.132</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.150</td>
<td>0.381</td>
<td>0.100</td>
<td>0.317</td>
</tr>
<tr>
<td>Median</td>
<td>Tree</td>
<td>0.119</td>
<td>0.319</td>
<td>0.108</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.148</td>
<td>0.376</td>
<td>0.101</td>
<td>0.323</td>
</tr>
<tr>
<td>75%ile</td>
<td>Tree</td>
<td>0.131</td>
<td>0.402</td>
<td>0.119</td>
<td>0.365</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.164</td>
<td>0.389</td>
<td>0.106</td>
<td>0.310</td>
</tr>
</tbody>
</table>

### RMSE: Null Median

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Approach</th>
<th>No Censoring Bilirubin</th>
<th>No Censoring Protime</th>
<th>20% Censoring Bilirubin</th>
<th>20% Censoring Protime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Tree</td>
<td>0.038</td>
<td>0.138</td>
<td>0.035</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.068</td>
<td>0.154</td>
<td>0.070</td>
<td>0.147</td>
</tr>
<tr>
<td>Median</td>
<td>Tree</td>
<td>0.083</td>
<td>0.319</td>
<td>0.078</td>
<td>0.285</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.072</td>
<td>0.220</td>
<td>0.073</td>
<td>0.163</td>
</tr>
<tr>
<td>75%ile</td>
<td>Tree</td>
<td>0.129</td>
<td>0.183</td>
<td>0.125</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.067</td>
<td>0.169</td>
<td>0.066</td>
<td>0.146</td>
</tr>
</tbody>
</table>

### RMSE: Proportional Hazards

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Approach</th>
<th>No Censoring Bilirubin</th>
<th>No Censoring Protime</th>
<th>20% Censoring Bilirubin</th>
<th>20% Censoring Protime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Tree</td>
<td>0.020</td>
<td>0.109</td>
<td>0.055</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.018</td>
<td>0.104</td>
<td>0.052</td>
<td>0.105</td>
</tr>
<tr>
<td>Median</td>
<td>Tree</td>
<td>0.031</td>
<td>0.133</td>
<td>0.068</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.022</td>
<td>0.113</td>
<td>0.062</td>
<td>0.106</td>
</tr>
<tr>
<td>75%ile</td>
<td>Tree</td>
<td>0.079</td>
<td>0.129</td>
<td>0.101</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.072</td>
<td>0.087</td>
<td>0.096</td>
<td>0.095</td>
</tr>
</tbody>
</table>
Coverage of 90% CI, Power: Strong Null

<table>
<thead>
<tr>
<th>θ</th>
<th>Approach</th>
<th>Bilirubin</th>
<th>Protome</th>
<th>Bilirubin</th>
<th>Protome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Tree</td>
<td>93.0</td>
<td>92.5</td>
<td>7.0</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>89.6</td>
<td>89.4</td>
<td>10.4</td>
<td>10.6</td>
</tr>
<tr>
<td>Median</td>
<td>Tree</td>
<td>93.2</td>
<td>93.7</td>
<td>6.8</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>89.5</td>
<td>89.4</td>
<td>10.5</td>
<td>10.6</td>
</tr>
<tr>
<td>75%ile</td>
<td>Tree</td>
<td>94.8</td>
<td>94.1</td>
<td>5.2</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>90.0</td>
<td>89.4</td>
<td>10.0</td>
<td>10.6</td>
</tr>
</tbody>
</table>

Coverage of 90% CI, Power: Null Hazard

<table>
<thead>
<tr>
<th>θ</th>
<th>Approach</th>
<th>Bilirubin</th>
<th>Protome</th>
<th>Bilirubin</th>
<th>Protome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Tree</td>
<td>90.85</td>
<td>95.45</td>
<td>92.95</td>
<td>70.05</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>2.50</td>
<td>22.30</td>
<td>24.80</td>
<td>13.70</td>
</tr>
<tr>
<td>Median</td>
<td>Tree</td>
<td>91.60</td>
<td>94.85</td>
<td>88.30</td>
<td>67.15</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.20</td>
<td>6.25</td>
<td>24.30</td>
<td>12.70</td>
</tr>
<tr>
<td>75%ile</td>
<td>Tree</td>
<td>96.25</td>
<td>97.70</td>
<td>100.00</td>
<td>98.50</td>
</tr>
<tr>
<td></td>
<td>Cox</td>
<td>0.00</td>
<td>0.85</td>
<td>21.50</td>
<td>14.75</td>
</tr>
</tbody>
</table>

Basic Results: Estimation

- Distribution-free approach can "find" contrasts across summary measures when semiparametric models don't hold
- Distribution-free approach does not do too badly in RMSE when semiparametric models do hold
- Overfitting data at first stage is desirable

Basic Results: Inference

- More work needed on resampling based estimated SEs
  - Sensitivity to small sample size, number of events in leaves
  - Sensitivity to boundaries between leaves
  - Better behavior with bagging
- Augmented data approaches are of interest
Extensions: ROC Curve Regression

- ROC curve is composition of survival curve on inverse survival curve
- (Semi)parametric models
  - Binormal model
  - ROC GLM (Pepe)
  - AUC GLM (Dodd, Pepe)
- Distribution-free models:
  - Estimate Healthy survival; use Diseased “placement values”
  - Estimate Healthy, Diseased survival curves separately
  - ROC trees grown under strong null
    - “Cocktail” of wAUCs

Total vs Ratio PSA by Age, Time to Dx

- ROC Trees
- ROC-GLM
- AUC-GLM

So? Depends on Perspective

- News item (circa 1987):
  - Mother and 3 year old visit neighbor, who owns a pool and a dog
  - Adults lost track of toddler “for just a few minutes”
  - Find toddler lying on side of pool
  - Toddler’s clothes soaking wet and with tooth marks
- Local headlines
  - Seattle Times: “Dog Saves Toddler From Drowning”
  - UW Daily: “Toddler Foils Dog’s Attempt at Murder”
- JSM corollary
  - New method “Inventor”: “Do you see what a difference it makes?”
  - Old method “Defender”: “So there is really no major difference.”
So? Depends on Perspective

- ROC Trees provide less smoothing and do so adaptively
  - Description: Useful to see departures from smooth trends
  - Inference: May not provide sufficient smoothing for efficiency
  - And ultimately the target of inference is a smoothed quantity

### Table 3: Contrasts of the AUC as estimated by ROC-GLM, AUC-GLM and ROC-Trees for the PSA example. All models were fit using a GEE with an independence working covariance matrix, and quantile-based confidence intervals are estimated using 1,000 bootstrap samples.

<table>
<thead>
<tr>
<th></th>
<th>ROC-Trees</th>
<th>ROC-GLM</th>
<th>AUC-GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (per decade)</td>
<td>-0.075 (-0.124, 0.021)</td>
<td>-0.028 (-0.046, 0.005)</td>
<td>-0.068 (-0.18, 0.047)</td>
</tr>
<tr>
<td>Marker</td>
<td>0.121 (0.093, 0.151)</td>
<td>0.120 (0.091, 0.140)</td>
<td>0.097 (0.049, 0.234)</td>
</tr>
<tr>
<td>Time to Dx (per year)</td>
<td>-0.021 (-0.041, 0.001)</td>
<td>-0.021 (-0.042, 0.000)</td>
<td>-0.020 (-0.044, 0.005)</td>
</tr>
<tr>
<td>Marker x Time</td>
<td>-0.007 (-0.029, 0.011)</td>
<td>-0.032 (-0.029, 0.003)</td>
<td>-0.012 (-0.036, 0.004)</td>
</tr>
</tbody>
</table>

Final Comments: Hobbies into the Future

- Censored survival data:
  - Distribution-free approach using trees is still of interest to me both descriptively and inferentially
  - But much more to be done

- ROC curves
  - Distribution-free approach using trees of interest descriptively
  - ROC GLM with binormal ROC curve is impressively robust
  - It was extremely hard to break
  - Two parameters do a pretty good job modeling a variety of curves that monotonically increase from (0.0) to (1.1)