SESSION 4:
INTRODUCTION TO COX REGRESSION

Module 12: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2017

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

OVERVIEW

• Session 1
  – Introductory examples
  – The survival function
  – Survival Distributions
  – Mean and Median survival time

• Session 2
  – Censored data
  – Risk sets
  – Censoring Assumptions
  – Kaplan-Meier Estimator and CI
  – Median and CI

• Session 3
  – Two-group comparisons: logrank test
  – Trend and heterogeneity tests for more than two groups

• Session 4
  – Introduction to Cox regression
OUTLINE

• Motivation:
  – Confounding in observational studies
  – Stratified randomization designs

• Cox Regression model
  – Coefficient interpretation
  – Estimation and testing
  – Relationship to 2- and K-sample tests
  – Examining non-proportionality

• Examples throughout
CONFOUNDING

• **Observational data**: sometimes observed associations between an explanatory variable and outcome can be due to their joint association with another variable.
  – Age related to both sex and risk of death.
  – Other examples?

PRECISION IN RCTS

• Because of randomization, confounding/imbalance usually not an issue except in small trials.
• As in linear regression, regression models for censored survival data allow group comparisons among subjects with similar values of adjustment or “precision” variables (more later).
• Fairer and possibly more powerful comparison as long as adjustment variables are not the result of treatment.
STRATIFIED RANDOMIZATION

• For strong predictors: concern about possible randomization imbalance
  – Clinic or center
  – Stage of disease
  – Sex
  – Age
• Adjust for stratification variables in analysis
  – More powerful if predictors are strong
  – Same conditioning as the sampling

OUTLINE

• Motivation:
  – Confounding in observational studies
  – Stratified randomization designs
• Cox Regression model
  – Coefficient interpretation
  – Estimation and testing
  – Relationship to 2- and K-sample tests
  – Examining non-proportionality
• Examples throughout
COX REGRESSION MODEL

• Usually written in terms of the hazard function
• As a function of independent variables $x_1, x_2, \ldots, x_k$,

$$
\lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}
$$

↑ relative risk / hazard ratio

$$
\log \lambda(t) = \log \lambda_0(t) + \beta_1 x_1 + \cdots + \beta_k x_k
$$

↑ intercept

RELATIVE RISK / HAZARD RATIO

$$
\lambda(t|x_1, \ldots, x_k) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}
$$

$$
\frac{\lambda(t|x_1, \ldots, x_k)}{\lambda(t|0, \ldots, 0)} = e^{\beta_1 x_1 + \cdots + \beta_k x_k}
$$
REGRESSION MODELS

LS Linear Regression: \( Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon \)

**Linear:** \( Y \sim N(\mu, \sigma^2) \) \( \mu = EY = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \)

**Cox:** \( T \sim S(t) \) \( \lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k} \)

\[ \uparrow \quad \uparrow \]

Distribution of outcome variable \( \uparrow \)
Dependence of distribution on \( x_1, \ldots, x_k \)

PROPORTIONAL HAZARDS MODEL

\[ \lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k} \]

Interpretation of \( e^{\beta_1} \) in general:

"Relative risk (or hazard ratio) associated with a one unit higher value of \( x_1 \), holding \( x_2, \ldots, x_k \) constant".

\[ \lambda(t) \text{ for } x_1 + 1: \quad \lambda_0(t)e^{\beta_1 (x_1 + 1) + \cdots + \beta_k x_k} \]

\[ \lambda(t) \text{ for } x_1: \quad \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k} \]

ratio: \( e^{\beta_1 (x_1 + 1 - x_1)} = e^{\beta_1} \)
EXAMPLE

Single binary $x$:

$$x = \begin{cases} 
1 & \text{Test treatment} \\
0 & \text{Standard treatment}
\end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta x}$$

Interpretation of $e^\beta$:

"Relative risk (or hazard ratio) comparing test treatment to standard".

$$\lambda(t) \text{ for } x = 1: \quad \lambda_0(t)e^{\beta \cdot 1} = \lambda_0(t)e^\beta$$

$$\lambda(t) \text{ for } x = 0: \quad \lambda_0(t)e^{\beta \cdot 0} = \lambda_0(t)$$

ratio: $e^{\beta(1-0)} = e^\beta$
RELATIONSHIP TO SURVIVAL FUNCTION

Single binary $x$:

$$x = \begin{cases} 1 & \text{Test treatment} \\ 0 & \text{Standard treatment} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta x} \implies S(t) = [S_0(t)]^{e^{\beta x}}$$

In terms of $S_0(t)$:

- $S(t)$ for $x = 1$: $[S_0(t)]^{e^{\beta \cdot 1}} = [S_0(t)]^{e^\beta}$
- $S(t)$ for $x = 0$: $[S_0(t)]^{e^{\beta \cdot 0}} = [S_0(t)]^1 = S_0(t)$

PICTURE

Hazard Function

Survival Function
ESTIMATES AND CONFIDENCE INTERVALS

• We estimate \( \beta \) by maximizing the "partial likelihood function"
• Requires iteration on computer
• \( \hat{\beta} \) is a MPLE (Maximum Partial Likelihood Estimator)
• We do not need to estimate \( \lambda_0(t) \) to do this

• Most packages will estimate se(\( \hat{\beta} \)) using the information matrix from this PL.
• 95% CI for \( \beta \): (\( \hat{\beta} - 1.96\text{se}(\hat{\beta}) \), \( \hat{\beta} + 1.96\text{se}(\hat{\beta}) \))
• 95% CI for RR = \( e^\beta : (e^{\hat{\beta}-1.96\text{se}(\hat{\beta})}, e^{\hat{\beta}+1.96\text{se}(\hat{\beta})}) \)

PARTIAL LIKELIHOOD

Data for the \( i^{th} \) subject: \( (t_i, \delta_i, x_{1i}, \ldots x_{ki}) \)
For subject with the \( j^{th} \) ordered failure time : \( (t_{(j)}, 1, x_{1(j)}, \ldots, x_{k(j)}) \)

\[
\text{PL}(\beta_1, \ldots, \beta_k) = \prod_{j=1}^{J} \frac{e^{\beta_1 x_{1(j)} + \cdots + \beta_k x_{k(j)}}}{\sum_{i:t_i \geq t_{(j)}} e^{\beta_1 x_{1i} + \cdots + \beta_k x_{ki}}}
\]

• \( (\hat{\beta}_1, \ldots, \hat{\beta}_k) \) are the values of \( (\beta_1, \ldots, \beta_k) \) that maximize PL(\( \beta_1, \ldots, \beta_k \)). (MLEs)
• Compares x values for the subject who failed at time \( t_{(j)} \) to those of all subjects at risk at time \( t_{(j)} \).
• Does not depend on the values of the \( t_i \), only on their order.
• Does not depend on \( \lambda_0(t) \).
**RISK SET PICTURE**

Risk Sets and Treatment

<table>
<thead>
<tr>
<th>1 vs 0.5</th>
<th>0 vs 0.5</th>
<th>1 vs 0.67</th>
<th>1 vs 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td></td>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FULL LIKELIHOOD**

\[
L(\beta, \lambda_0(t)) = \prod_{i=1}^{n} \frac{\text{Pr}[T = t_i]}{\text{Pr}[T > t_i]} \prod_{i=1}^{n} \lambda(t_i|x_i)S(t_i|x_i) \\
= \prod_{i=1}^{n} \lambda(t_i|x_i)S(t_i|x_i) \\
= \prod_{i=1}^{n} [\lambda(t_i|x_i)]^{\delta_i}S(t_i|x_i) \\
= \prod_{i=1}^{n} [\lambda_0(t_i)e^{\beta x_i}]^{\delta_i}e^{-\int_{0}^{t_i} \lambda_0(s)e^{\beta x}ds}
\]
PARTIAL LIKELIHOOD

Let $H_t$ represent the entire history of failure, censoring and $x$ in the sample before time $t$.

Then the likelihood can be rewritten as follows:

$$L(\beta, \lambda_0(t)) = \prod_{j=1}^{I} \Pr[\text{ith subject fails at } t_{(j)} | H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \cdot \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}]$$

$$= \prod_{j=1}^{I} \frac{\lambda(t_{(j)} | x_{(j)})}{\sum_{l: t_{(l)} \geq t_{(j)}} \lambda(t_{(l)} | x_{(l)})} \cdot \prod_{j=1}^{I} \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}]$$

$$= \prod_{j=1}^{I} \frac{\lambda_0(t_{(j)}) e^{\beta x_{(j)}}}{\sum_{l: t_{(l)} \geq t_{(j)}} \lambda_0(t_{(l)}) e^{\beta x_{(l)}}} \cdot \prod_{j=1}^{I} \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}]$$

Partial Likelihood
Depends only on $\beta$

HYPOTHESIS TESTS

Three tests of $H_0 : \beta = 0$ are possible:

1. Wald test: $\frac{\hat{\beta}}{se(\hat{\beta})}$

2. (Partial) Likelihood ratio test

3. Score test: ($\approx$ logrank test)

Likelihood ratio test is best, but requires fitting full ($\beta = \hat{\beta}$) and reduced ($\beta = 0$) models.
LIKELIHOODS AND TESTS

Hypothesis Tests

- Log likelihood
- Likelihood Ratio Test
- Slope = Score
- Log Likelihood Function
- Wald test

COLON CANCER EXAMPLE

- Clinical trial at Mayo Clinic
- Stage B₂ and C colon cancer patients; adjuvant therapy
- Three arms
  - Observation only
  - Levamisole (stage C only)
  - 5-FU + Levamisole at Mayo Clinic
- Stage C patients only
- Two treatment arms only
COLON CANCER EXAMPLE

Complementary log–log Transformation

SISCR 2017: Module 12 - Intro Survival
Barbara McKnight

Q: Which group has better survival?

A:
TEST COMPARISON

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald’s</td>
<td>8.13</td>
<td>.004</td>
</tr>
<tr>
<td>Score</td>
<td>8.21</td>
<td>.004</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>8.21</td>
<td>.004</td>
</tr>
</tbody>
</table>

Two-sided tests

ANOTHER EXAMPLE

Three groups: use indicators for two

\[
\begin{align*}
  x_1 &= \begin{cases} 
  1 & \text{Levamisole Only} \\
  0 & \text{otherwise}
\end{cases} \\
  x_2 &= \begin{cases} 
  1 & \text{Levamisole + 5FU} \\
  0 & \text{otherwise}
\end{cases}
\end{align*}
\]

Model: \( \lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2} \)

RRs: Levamisole Only vs. Observation \( e^{\beta_1} \)
     Levamisole + 5FU vs. Observation \( e^{\beta_2} \)
     Levamisole + 5FU vs. Levamisole Only \( e^{\beta_2 - \beta_1} \)
HEURISTIC HAZARDS

![Diagram of Proportional Hazards and Parallel Log Hazards]

COOLON CANCER

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Deaths</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation Only</td>
<td>315</td>
<td>168</td>
<td>1.0 (reference)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Levamisole Only</td>
<td>310</td>
<td>161</td>
<td>0.97</td>
<td>(0.78, 1.21)</td>
<td>0.81</td>
</tr>
<tr>
<td>Levamisole + SFU</td>
<td>204</td>
<td>123</td>
<td>0.69</td>
<td>(0.55, 0.87)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Q: Which group has best survival?

A:
TEST COMPARISON

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald’s</td>
<td>11.56</td>
<td>.003</td>
</tr>
<tr>
<td>Score</td>
<td>11.68</td>
<td>.003</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>12.15</td>
<td>.002</td>
</tr>
</tbody>
</table>

Same hypothesis as 3-group heterogeneity test. Score test is same in large samples.

COLON CANCER TRIAL DATA

Colon Cancer Trial: All Three Groups

![Survival Probability vs Days from Diagnosis](image)
**TREND**

- When there are several groups, it is sometimes of interest to test whether risk increases from one group to the next:
  - Several dose groups
  - Other ordered variable
  - Example: tumor differentiation

- For $x = \begin{cases} 1 & \text{well differentiated} \\ 2 & \text{moderately differentiated} \\ 3 & \text{poorly differentiated} \end{cases}$

  Model: $\lambda(t) = \lambda_0(t)e^{\beta x}$

- Score test is the same as the trend test
- Could use other values for $x$ (actual dose levels)

---

**TREND**

For $x = \begin{cases} 1 & \text{well differentiated} \\ 2 & \text{moderately differentiated} \\ 3 & \text{poorly differentiated} \end{cases}$

Model: $\lambda(t) = \lambda_0(t)e^{\beta x}$

**Interpretation of $e^\beta$:** HR associated with the comparison of one worse differentiation group to one better:

- poorly differentiated to moderately differentiated, or
- moderately differentiated to well differentiated

**Q:** What is HR comparing poorly differentiated to well differentiated?

**A:**
TREND WITH DIFFERENTIATION

One presentation based entirely on trend ("grouped linear") model:

<table>
<thead>
<tr>
<th>One category worse differentiation (well, moderately, poor)</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4</td>
<td>(1.1, 1.8)</td>
<td></td>
</tr>
<tr>
<td>P = .003 (trend)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

I prefer presenting hazard ratios and CI's based on dummy variable model, and providing P-value for trend.
TREND WITH DIFFERENTIATION

My preferred presentation based on dummy variable mode with trend P-value:

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Deaths</th>
<th>Hazard Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well differentiated</td>
<td>66</td>
<td>26</td>
<td>1.0 (reference)</td>
<td>--</td>
</tr>
<tr>
<td>Moderately differentiated</td>
<td>434</td>
<td>196</td>
<td>1.2</td>
<td>(0.80, 1.8)</td>
</tr>
<tr>
<td>Poorly differentiated</td>
<td>98</td>
<td>54</td>
<td>1.8</td>
<td>(1.2, 3.0)</td>
</tr>
</tbody>
</table>

P = .003 (trend)

I usually would not present this for an a priori trend hypothesis, but for comparison here, the heterogeneity P-value (2 df) is 0.009.

OVARIAN CANCER SCREENING TRIAL

![Graph of ovarian cancer screening trial results]

MMS vs no screening HR 0.85 (95% CI 0.70-1.03); p=0.10
USS vs no screening HR 0.89 (95% CI 0.73-1.07); p=0.21

Number at risk

<table>
<thead>
<tr>
<th>No screening</th>
<th>MMS</th>
<th>USS</th>
</tr>
</thead>
<tbody>
<tr>
<td>101299</td>
<td>50624</td>
<td>50623</td>
</tr>
<tr>
<td>100720</td>
<td>50343</td>
<td>50383</td>
</tr>
<tr>
<td>99662</td>
<td>49846</td>
<td>49838</td>
</tr>
<tr>
<td>98238</td>
<td>49176</td>
<td>49192</td>
</tr>
<tr>
<td>96632</td>
<td>48345</td>
<td>48363</td>
</tr>
<tr>
<td>75582</td>
<td>37758</td>
<td>37668</td>
</tr>
<tr>
<td>25252</td>
<td>12592</td>
<td>12689</td>
</tr>
</tbody>
</table>
PROPORTIONAL HAZARDS

• One way to examine evidence against proportional hazards is to look at plots of scaled Schoenfeld residuals and perform tests based on them.
• For each failing subject there is a Schoenfeld residual for each x variable in the model.
• At the subject's failure time, the residual measures how the value of x for the subject who fails differs from a weighted average of x values for those still at risk. (Weights depend on estimated HR for each subject at risk).
• If consistently high or low over an interval of time, this is evidence that the hazard at that time is even higher (lower) for the subject with that x than the model indicates.

SCHOENFELD RESIDUALS

Formula for Schoenfeld residuals

Let \( r_i(t) = e^{\hat{\beta}x_i(t)} \) be the estimated hazard ratio for the \( i^{th} \) subject at \( t \) compared to \( x(t) = 0 \).

Then for \( \bar{x}(\hat{\beta}, t) = \frac{\sum_{\text{at risk at } t} r_i(t)x_i(t)}{\sum_{\text{at risk at } t} r_i(t)} \),

The Schoenfeld residual for the \( k^{th} \) subject failing at time \( t \) is given by \( x_k(t) - \bar{x}(\hat{\beta}, t) \).

The scaled Schoenfeld residual is the Schoenfeld residual divided by a variance estimate.
SCHOENFELD RESIDUALS

- Grambsch and Therneau (1994) showed that the scaled Schoenfeld residual measures the deviation of a time-dependent log hazard ratio $\beta(t)$ from time-constant $\hat{\beta}$.

- Can use linear regression comparing scaled Schoenfeld residuals to functions of time to examine evidence for lack of constant hazard ratio over time.


COLON CANCER TRIAL DATA

| Observation Arm Omitted | $\hat{\beta}$ | $\exp(\hat{\beta})$ | se($\hat{\beta}$) | z   | Pr($>|z|$) |
|------------------------|---------------|---------------------|-------------------|-----|------------|
| 5FU + Lev              | -0.34         | 0.71                | 0.12              | -2.83| 0.0064     |
| 4+ Nodes Pos           | 0.98          | 2.67                | 0.12              | 8.08| <0.0001    |

$e^{\hat{\beta}_{RX}}$ CI: (0.5629, 0.9008)

LRT: 8.098 on 1 df, $P = 0.0044$
FOR NODE 4 POSITIVITY

FOR TREATMENT
TEST FOR NON-PROPORTIONALITY

<table>
<thead>
<tr>
<th>Variable</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>node4</td>
<td>0.158</td>
</tr>
<tr>
<td>txLev+5FU</td>
<td>0.560</td>
</tr>
</tbody>
</table>

No strong evidence for non-proportionality based on scaled Schoenfeld residuals correlation with “time” $S(t)$.

TO WATCH OUT FOR:

- Coefficients in Cox regression are positively associated with risk, not survival.
  - Positive $\beta$ means large values of $x$ are associated with shorter survival.
- Without certain types of time-dependent covariates, Cox regression does not depend on the actual times, just their order.
  - Can add a constant to all times to remove zeros (some packages remove observations with time = 0) without changing inference
- For LRT, nested models must be compared based on same subjects.
  - If some values of variables in larger model are missing, these subjects must be removed from fit of smaller model.
- Hazards may not always be proportional