Lecture 7

Proportional Hazards Model - Handling Ties and Survival Estimation

Statistics 255 - Survival Analysis

Presented February 4, 2016

Dan Gillen
Department of Statistics
University of California, Irvine
Recall

- If there is one failure at \( t_{(j)} \) and any number of censorings, by convention we assume that censorings immediately follow the failure.

- Suppose at \( t_{(j)} \), there are multiple failures:
  - \( R_{(j)} \) is the set of subjects at risk for failure at \( t_{(j)} \)
  - \( D_{(j)} \) is the set of subjects failing at \( t_{(j)} \)
  - \( d_{(j)} \) is the number of subjects in \( D_{(j)} \)

- The subjects in \( D_{(j)} \) are said to be tied.
Cox Regression Model - Tied Data

Construction of the likelihood

- How to construct the likelihood?
- Recall for *untied data*

\[ L(j) = \Pr\{x_{(j)} \text{ fails at } t_{(j)} \mid \text{ one subject in } R_{(j)} \text{ failed at } t_{(j)} \} \]

where \( x_{(j)} \) is the (column) vector of covariates for the subject that failed at time \( t_{(j)} \)

- There are multiple approaches that have been proposed to account for ties in the partial likelihood that include:
  - Cox’s (1972) "exact discrete" method
  - Breslow’s approximation
  - Efron’s approximation
  - Kalbfleisch and Prentice’s exact method
Cox Regression Model - Tied Data

Cox’s (1972) "exact discrete" method

▶ Cox suggested an approach that presumed true ties could actually occur, which would be in the case of a discrete time model.

▶ In this case we could formulate the contribution to the partial likelihood at failure time \( t(j) \) as follows:

\[
L(j) = \Pr\{\text{all } x_i \in D(j) \text{ fail at } t(j) \mid d(j) \text{ subjects in } R(j) \text{ fail at } t(j)\} \\
= \frac{\Pr\{\text{all } x_i \in D(j) \text{ fail at } t(j)\}}{\Pr\{d(j) \text{ subjects in } R(j) \text{ fail at } t(j)\}}
\]
Cox Regression Model - Tied Data

Cox’s (1972) "Exact Discrete" Method : Simple Example

- Suppose data are:

\[(7, x_1), (9^+, x_2), (18, x_3), (18, x_4), (19^+, x_5)\]

- What is \(L(j)\) at \(t(j) = 18 (j = 2)\)?

\[D(2) = \{x_3, x_4\} \quad \text{and} \quad d(2) = 2\]

\[R(2) = \{x_3, x_4, x_5\}\]

\[L(2) = \frac{\Pr\{x_3 \text{ and } x_4 \text{ fail at } 18\}}{\Pr\{\text{any 2 subjects in } R(2) \text{ fail at } 18\}}\]
Cox Regression Model - Tied Data

Cox’s (1972) "Exact Discrete" Method: Simple Example

**Numerator:**

\[ \Pr\{x_3 \text{ and } x_4 \text{ fail at 18}\} = \Pr\{x_3 \text{ fails at 18}\} \times \Pr\{x_4 \text{ fails at 18}\} = \lambda_3(18)(\Delta t)\lambda_4(18)(\Delta t) \]

\[ \text{...the product of the risks of the failed subjects} \]

**Denominator:**

\[ \Pr\{\text{any 2 subjects in } R(2) \text{ fail at 18}\} = \Pr\{x_3 \text{ fails at } t(2)\} \times \Pr\{x_4 \text{ fails at } t(2)\} \]
\[ + \Pr\{x_3 \text{ fails at } t(2)\} \times \Pr\{x_5 \text{ fails at } t(2)\} \]
\[ + \Pr\{x_4 \text{ fails at } t(2)\} \times \Pr\{x_5 \text{ fails at } t(2)\} \]
\[ = \lambda_3(18)(\Delta t)\lambda_4(18)(\Delta t) \]
\[ + \lambda_3(18)(\Delta t)\lambda_5(18)(\Delta t) \]
\[ + \lambda_4(18)(\Delta t)\lambda_5(18)(\Delta t) \]

\[ \text{...sum of product of risks for all possible pairs of failed subjects} \]
Cox Regression Model - Tied Data

Cox’s (1972) "Exact Discrete" Method : Simple Example

- The $(\Delta t)$’s and baseline hazards cancel and give:

$$L(2) = \frac{e^{\beta^T x_3} e^{\beta^T x_4}}{e^{\beta^T x_3} e^{\beta^T x_4} + e^{\beta^T x_3} e^{\beta^T x_5} + e^{\beta^T x_4} e^{\beta^T x_5}}$$

- In general, let $(R(j); d(j))$ be the set of subsets of $R(j)$ of size $d(j)$.

  - In example above:

$$R(2) = \{3, 4, 5\}$$

$$\{ (R(2); d(2) ) = \{ (3, 4), (3, 5), (4, 5) \}$$
Cox Regression Model - Tied Data

Exact Discrete Partial Likelihood for Ties

- The partial likelihood for the \( j \)th failure time is

\[
L_{(j)}^{ED} = \frac{\prod_{i \in D(j)} \text{risk for subject } i \text{ at } t(j)}{\sum_{l \in (R(j); d(j))} \left\{ \prod_{i \in l} \text{risk for subject } i \text{ at } t(j) \right\}}
\]

- This is the "exact discrete" partial likelihood for tied failure times and is implemented in R using the `coxph()` function with the option `ties="exact"`

- In SAS, this method for handling ties is obtained using the option "discrete"
Cox Regression Model - Tied Data

Kalbfleisch and Prentice’s "exact" method

- Kalbfleisch and Prentice proposed an approach that presumes time is truly continuous

- Under this setting, if there are tied failure times then the tie is due to the imprecise nature of our measurement of time, and that there must be some true ordering of the times

- In this case we can consider the average partial likelihood contribution at time $t_{(j)}$ that arises from breaking the ties in all possible ways
Kalbfleisch and Prentice’s "exact" method: Simple Example

- Returning to the previous example, suppose that at failure time \( t_2 \)

\[
D(2) = \{x_3, x_4\} \quad \text{and} \quad d(2) = 2
\]
\[
R(2) = \{x_3, x_4, x_5\}
\]

Then the average partial likelihood contribution at time \( t_j \) is

\[
L(2) = \frac{1}{2} \left[ \frac{e^{\beta^T x_3}}{e^{\beta^T x_3} + e^{\beta^T x_4} + e^{\beta^T x_5}} \times \frac{e^{\beta^T x_4}}{e^{\beta^T x_4} + e^{\beta^T x_5}} + \frac{e^{\beta^T x_4}}{e^{\beta^T x_3} + e^{\beta^T x_4} + e^{\beta^T x_5}} \times \frac{e^{\beta^T x_3}}{e^{\beta^T x_3} + e^{\beta^T x_4} + e^{\beta^T x_5}} \right]
\]
Kalbfleisch and Prentice’s "exact continuous" method: Simple Example

- In general, if we let
  - $Q(j)$ denote the set of $d(j)!$ permutations of the subjects failing at time $t_j$
  - $P = (p_1, \ldots, p_{d(j)})$ be an element in $Q(j)$
  - $R(j, P, r) = R(j) - \{p_1, \ldots, p_{r-1}\}$

then the average partial likelihood contribution at $t(j)$ is given by

$$
L^{EC}_{(j)} = \frac{1}{d(j)!} \exp\{\beta^T s(j)\} \sum_{P \in Q(j)} \prod_{r=1}^{d(j)} \left\{ \sum_{\ell \in R(j, P, r)} \exp\{\beta^T x_\ell\} \right\}^{-1}
$$

where $s(j) = \sum_{i=1}^{d(j)} x_i$
### Kalbfleisch and Prentice’s "exact continuous" method: Simple Example

- This albkleisch and Prentice's "exact continuous" method for ties is currently not available in R.
- This method is available in SAS via the option "exact" for handling ties.
- Because R and SAS provide different likelihood when a user specifies the "exact" option, this often leads to confusion and explains differential results between the two packages!
Cox Regression Model - Tied Data

**Breslow Method for Ties**

- The denominator in $L_{ED}^{(j)}$ and $L_{EC}^{(j)}$ can be painful to compute if there are many ties.

- For large risk sets, we can approximate:

$$
\sum_{l \in (R_{(j)}; d_{(j)})} \left\{ \prod_{i \in l} \text{risk for subject } i \text{ at } t_{(j)} \right\} \\
\approx \left\{ \sum_{i \in R_{(j)}} \text{risk for subject } i \text{ at } t_{(j)} \right\} d_{(j)}
$$

- This gives the **Breslow approximation** to the partial likelihood for the $j$th failure time:

$$
L_{B}^{(j)} = \frac{\prod_{i \in D_{(j)}} \text{risk for subject } i \text{ at } t_{(j)}}{\left\{ \sum_{i \in R_{(j)}} \text{risk for subject } i \text{ at } t_{(j)} \right\} d_{(j)}}
\approx \frac{\prod_{i \in D_{(j)}} \exp\{\beta^T x_i\}}{\left\{ \sum_{i \in R_{(j)}} \exp\{\beta^T x_i\} \right\} d_{(j)}}
$$
**Cox Regression Model - Tied Data**

**Efron Method for Ties**

- Nearly all software packages default to the Breslow method for handling ties....except R!
- One issue with the Breslow method is that it considers each of the events at a given time as distinct and allows all failed subjects to contribute fully to the risk set.
- The *Efron* approximation allows for partial contributions to the risk set for each of the members that fail at time $t_{(j)}$:

$$L_{Efron}^{(j)} = \frac{\prod_{i \in D_{(j)}} \exp\{\beta^T x_i\}}{\prod_{h=1}^{d_{(j)}} \left\{ \sum_{i \in R_{(j)}} \exp\{\beta^T x_i\} - \frac{h-1}{d_i} \sum_{k \in D_{(j)}} \exp\{\beta^T x_k\} \right\}}$$
Cox Regression Model - Tied Data

Example : Implementation in R

▶ Consider the data on time to death among larynx cancer patients where focus was on the diagnostic utility of disease staging (K & M Section 1.8)

> larynx <- read.table("http://www.ics.uci.edu/~dgillen/STAT255/Data/larynx.txt")

> larynx[1:10,]

```
  stage t2death age year death
1     1    0.6   77   76    1
2     1    1.3   53   71    1
3     1    2.4   45   71    1
4     1    2.5   57   78    0
5     1    3.2   58   74    1
6     1    3.2   51   77    0
7     1    3.3   76   74    1
8     1    3.3   63   77    0
9     1    3.5   43   71    1
10    1    3.5   60   73    1
```

> 

> ##
> ###
> ###### Check for ties
> ##

> sum( duplicated(larynx$t2death[ larynx$death==1 ] ) )

[1] 16
Cox Regression Model - Tied Data

Example : Implementation in R

► Efron Approximation

```r
> ##
> ###### Efron approximation
> ##
> ##
> fit.efron <- coxph( Surv( t2death, death ) ~ age + factor(stage),
> data=larynx )
> summary( fit.efron )

n= 90, number of events= 50

table of coefficients:

  coef exp(coef) se(coef)  z  Pr(>|z|)
  age   0.0190  1.0192  0.0143 1.33 0.182
  factor(stage)2 0.1400  1.1503  0.4625 0.30 0.762
  factor(stage)3 0.6424  1.9010  0.3561 1.80 0.071 .
  factor(stage)4 1.7060  5.5068  0.4219 4.04 5.3e-05 ***

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 1

exp(coef) exp(-coef) lower .95 upper .95
  age  1.02  0.981  0.991  1.05
  factor(stage)2 1.15  0.869  0.465  2.85
  factor(stage)3 1.90  0.526  0.946  3.82
  factor(stage)4 5.51  0.182  2.409 12.59

Likelihood ratio test= 18.3 on 4 df,  p=0.00107
Wald test      = 21.1 on 4 df,  p=0.000296
Score (logrank) test = 24.8 on 4 df,  p=5.57e-05
```
**Cox Regression Model - Tied Data**

**Example: Implementation in R**

- **Breslow Approximation**

```r
> ##
> ###### Breslow approximation
> ##
> fit.breslow <- coxph( Surv( t2death, death ) ~ age + factor(stage),
>                      data=larynx, method="breslow" )
> summary( fit.breslow )

n= 90, number of events= 50

| coef   | exp(coef) | se(coef) | z     | Pr(>|z|) |
|--------|-----------|----------|-------|----------|
| age    | 0.0189    | 1.0191   | 0.0143| 1.33     | 0.185    |
| factor(stage)2 | 0.1386 | 1.1486   | 0.4623| 0.30     | 0.764    |
| factor(stage)3 | 0.6383 | 1.8934   | 0.3561| 1.79     | 0.073 .  |
| factor(stage)4 | 1.6931 | 5.4361   | 0.4222| 4.01     | 6.1e-05 *** |

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 1

<table>
<thead>
<tr>
<th>exp(coef)</th>
<th>exp(-coef)</th>
<th>lower .95</th>
<th>upper .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1.02</td>
<td>0.981</td>
<td>0.991</td>
</tr>
<tr>
<td>factor(stage)2</td>
<td>1.15</td>
<td>0.871</td>
<td>0.464</td>
</tr>
<tr>
<td>factor(stage)3</td>
<td>1.89</td>
<td>0.528</td>
<td>0.942</td>
</tr>
<tr>
<td>factor(stage)4</td>
<td>5.44</td>
<td>0.184</td>
<td>2.376</td>
</tr>
</tbody>
</table>

Likelihood ratio test= 18.1 on 4 df,  p=0.0012
Wald test        = 20.8 on 4 df,  p=0.000344
Score (logrank) test = 24.3 on 4 df,  p=6.87e-05

---

Estimating \( \Lambda_0(t) \) and \( S_0(t) \)

- Interpretation of \( S_0(t) \)
- Ex: Catheter infection
- Ex: Duration of breastfeeding
- Review of univariate case
- Incorporation of covariates via the PH model
Cox Regression Model - Tied Data

Example: Implementation in R

▶ Cox Exact Discrete Method

```R
> ##
> ####### Exact discrete (Cox) method
> ##
> fit.exact <- coxph( Surv( t2death, death ) ~ age + factor(stage),
>                     data=larynx, method="exact" )
> summary( fit.exact )

n= 90, number of events= 50

coef exp(coef) se(coef)  z  Pr(>|z|)
age  0.0193  1.0195  0.0144  1.34 0.181
factor(stage)2 0.1410  1.1515  0.4648  0.30 0.762
factor(stage)3 0.6475  1.9108  0.3587  1.81 0.071 .
factor(stage)4 1.7346  5.6669  0.4294  4.04 5.4e-05 ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

exp(coef) exp(-coef) lower .95 upper .95
age  1.02  0.981  0.991  1.05
factor(stage)2 1.15  0.868  0.463  2.86
factor(stage)3 1.91  0.523  0.946  3.86
factor(stage)4 5.67  0.176  2.442 13.15

Likelihood ratio test= 18.4 on 4 df,  p=0.00102
Wald test      = 21  on 4 df,  p=0.000317
Score (logrank) test = 24.7 on 4 df,  p=5.9e-05
```
Summary

- Why are data tied?
  - Is time truly discrete or do we have a crude measure of a continuous outcome?
  - Are the data interval censored? If so, might consider another method (later in course)

- Questions:
  - Are there any times with many failures (e.g. more than 10% of risk set fails at a given time)?
  - Are there many times with a few ties each?
  - If so, do sensitivity analysis: – use Efron (or Breslow) method for model building – check some model fits (including final model) with exact algorithm relative to the true nature of the outcome (discrete or continuous)

- Efron method best for everyday use.
Estimating the $\Lambda_0(t)$ and $S_0(t)$

Recall

- In the PH Model

\[ \lambda_i(t \mid x_i) = \lambda_0(t) \exp(\beta^T x_i) \]

and

\[ \Lambda_i(t \mid x_i) = \Lambda_0(t) \exp(\beta^T x_i) \]

- Thus

\[ S_i(t \mid x_i) = e^{-\Lambda_i(t \mid x_i)} \]

\[ = e^{-\Lambda_0(t)} \exp(\beta^T x_i) \]

\[ = \{e^{-\Lambda_0(t)}\} \exp(\beta^T x_i) \]

\[ = \{S_0(t)\} \exp(\beta^T x_i) \]

- $\lambda_0(t)$, $\Lambda_0(t)$, and $S_0(t)$ are the functions corresponding to $x_i = 0 \ldots$ they are called the baseline hazard, cumulative hazard and survival functions
Estimating the $\Lambda_0(t)$ and $S_0(t)$

**Example**

- In patients with renal insufficiency, catheters are placed surgically (1) or percutaneously (2)

- A study to compare the two methods was mounted with the interest of reducing the rate of catheter exit site infections (K & M Section 1.4)
  - Time origin: placement of catheter
  - Failure event: occurrence of exit site infection
  - Primary reason for censoring: catheter failure (is censoring independent?)

Estimating the $\Lambda_0(t)$ and $S_0(t)$

**Example**

- Fit PH model to renal insufficiency data, with covariate `cathpla` and estimate survival for the placement groups.

```r
# # # # #
# # # # # # Kidney catheter example from K & M (1.4)
> #
> catheter <- read.table( "http://www.ics.uci.edu/~dgillen/STAT255/Data/kidneycatheter.txt" )
> names( catheter ) <- c( "time", "infect", "cathpla" )
> catheter[1:5,]
  time infect cathpla
  1 1.5 1 1
  2 3.5 1 1
  3 4.5 1 1
  4 4.5 1 1
  5 5.5 1 1

> #
> #
> # # # # # # Fit Cox model for type of placement
> fit <- coxph( Surv( time, infect ) ~ cathpla, data=catheter )
> estSurv <- survfit( fit, newdata=data.frame( cathpla=c(1,2) ) )
```
Estimating $\Lambda_0(t)$ and $S_0(t)$

Example

- Now, plot fitted survival functions for $\text{cathpla} = 1$ (surgically placed) and $\text{cathpla} = 2$ (percutaneously placed)

```r
> plot( estSurv, lty=1:2, las=1, ylab="Infection-free Survival", xlab="Time to Infection (mths)"
> legend( 5, .4, lty=1:2, legend=c("Surgical Placement", "Percutaneous Placement"), bty="n" )
```

![Survival function graph](image-url)
Estimating $\Lambda_0(t)$ and $S_0(t)$

**Example**

- **Note:** The previous curves are *fitted* curves (subject to PH assumption). The “raw” KM survival fits are below

```r
> kmFit <- survfit( Surv( time, infect ) ~ cathpla, data=catheter )
> kmPlot( kmFit, xscale=1, ylab="Infection-free Survival",
+ xlab="Time to Infection (mths)",
+ groupLabels=c("Surgical Placement",
+ "Percutaneous Placement") )
> legend( 2, .3, lty=1:2, legend=c("Surgical Placement",
+ "Percutaneous Placement"), bty="n" )
```

![Survival Curve Diagram](image.png)

Time to Infection (mths)
Infection-free Survival
0.0 3.0 6.0 9.0 12.0 15.5 19.0 22.5 26.0 29.5

0.0 0.2 0.4 0.6 0.8 1.0

- Surgical Placement
- Percutaneous Placement

<table>
<thead>
<tr>
<th></th>
<th>Surgical Placement</th>
<th>Percutaneous Placement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to Infection (mths)</td>
<td>43 (0)</td>
<td>76 (0)</td>
</tr>
<tr>
<td></td>
<td>22 (8)</td>
<td>27 (10)</td>
</tr>
<tr>
<td></td>
<td>9 (13)</td>
<td>11 (11)</td>
</tr>
<tr>
<td></td>
<td>0 (15)</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Total</td>
<td>119 (0)</td>
<td>49 (18)</td>
</tr>
<tr>
<td></td>
<td>20 (24)</td>
<td>1 (26)</td>
</tr>
</tbody>
</table>
Estimating $\Lambda_0(t)$ and $S_0(t)$

**Interpretation of $S_0(t)$**

- $\hat{S}_0(t)$ is the *fitted* survival function:
  1. For $x_i = 0$ and
  2. Under the PH assumption

- *If* PH assumption is correct, $\hat{S}_0(t)$ can be though of as an *adjusted* survival function, where the adjustment is to the case where *all subjects* have $x_i = 0$

- **What if *no subjects* have $x_i = 0$?**
  - Then $\hat{S}_0(t)$ is a *meaningless extrapolation* (e.g. if $x_i = \text{age}_i$ and all subjects over 65 years)

- Therefore: make sure you construct your covariates so that $x_i = 0$ is representative of a subject who could appear in your sample...another good reason to *center* your covariates
**Estimating $\Lambda_0(t)$ and $S_0(t)$**

**Example**

- Breast feeding duration among new mothers (K & M Section 1.14)

- Data are available on time (weeks) from birth to weaning for 927 first born children

- Here, we will consider the "survival" experience, adjusting for birth year ($y_{ob}$)...

```r
# Breastfeeding example from K & M (1.14)

> names( bfeed ) <- c( "duration", "icompbf", "racemom", "poverty", + "momsmoke", "momdrink", "momage", "yob", "momeduc", "precare" )
```
Estimating $\Lambda_0(t)$ and $S_0(t)$

Example

- Estimate and plot the baseline survival curve after "centering" birth year at 1980

```r
> ##
> ###### (Roughly) center year of birth to 1980
> bfeed$yob.80 <- bfeed$yob - 80
> 
> ##
> ###### Fit Cox model with year of birth as single
> ###### covariate and plot baseline survival
> fit.80 <- coxph( Surv( duration, icompbf ) ~ yob.80, data=bfeed )
> estSurv <- survfit( fit.80, newdata=data.frame( yob.80=c(0) ) )
> plot( estSurv, ylab="Proportion Breastfeeding",
> xlab="Time from Birth (wks)"
> 
> legend( 50, .8, lty=1:2,
>        legend=c("Baseline Survival Estimate (1980 birth year)",
>                   "95% Pointwise CI"), bty="n" )
```
Estimating $\Lambda_0(t)$ and $S_0(t)$

Example

- Resulting baseline survival curve after “centering” birth year at 1980
Estimating $\Lambda_0(t)$ and $S_0(t)$

Example

- Compare with baseline survival curve without “centering" birth year at 1980

```r
> ##
> ####### Compare with baseline estimate for "non-centered" fit
> fit.00 <- coxph( Surv( duration, icompbf ) ~ yob, data=bfeed )
> estSurv <- survfit( fit.00, newdata=data.frame( yob=c(0) ) )
> plot( estSurv, ylab="Proportion Breastfeeding",
    xlab="Time from Birth (wks)"
> legend( 10, .4, lty=1:2,
    legend=c("Baseline Survival Estimate (1900 birth year)",
    "95% Pointwise CI"), bty="n" )
```
Estimating $\Lambda_0(t)$ and $S_0(t)$

Example

- Resulting baseline survival curve without “centering" birth year at 1980
Estimating $\Lambda_0(t)$ and $S_0(t)$

### Estimation of $\Lambda_0(t)$ and $S_0(t)$

- **Recall in the univariate case:**

\[
\hat{S}_{KM}(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)
\]

\[
\Lambda_{NA}(t) = \sum_{j: t_j \leq t} \frac{d_j}{n_j}
\]

- **For 1 subject alive at $t_j - \Delta t$:**

\[
\Pr\{\text{failure in } (t_j - \Delta t, t_j] \mid \text{alive at } t_j - \Delta t\} \approx \lambda(t_j)(\Delta t)
\]

- **For $n_j$ subjects alive at $t_j - \Delta t$ (i.e. in $R_j$):**

\[
d_j = \sum_{i \in R_j} I\{\text{subject } i \text{ failed at } t_j\}
\]
Estimating $\Lambda_0(t)$ and $S_0(t)$

**Estimation of $\Lambda_0(t)$ and $S_0(t)$**

▶ Then:

$$E\{ \# \text{ failures in } (t_j - \Delta t, t_j] \mid n_j \text{ alive at } t_j - \Delta t \} = E\{ d_j \mid R_j \}$$

▶ But:

$$E\{ d_j \mid R_j \}$$

$$= \sum_{i \in R_j} \Pr\{ \text{subj } i \text{ fails in } (t_j - \Delta t, t_j] \mid \text{subj } i \text{ alive at } t_j - \Delta t \}$$

$$\approx \sum_{i \in R_j} \lambda(t_j)(\Delta t)$$

$$= n_j \lambda(t_j)(\Delta t)$$
Estimating $\Lambda_0(t)$ and $S_0(t)$

<table>
<thead>
<tr>
<th>Estimation of $\Lambda_0(t)$ and $S_0(t)$</th>
</tr>
</thead>
</table>
| Now, estimate $\lambda(t)$ by replacing $E\{d_j \mid R_j\}$ with $d_j$ (this is like setting expected equal to observed):

$$d_j = n_j \hat{\lambda}(t_j)(\Delta t)$$

or

$$\hat{\lambda}(t) = \begin{cases} 
\frac{d_j}{n_j(\Delta t)} & , \quad t \in (t_j - \Delta t, t_j] \\
0 & , \quad \text{o.w.}
\end{cases}$$

Ex: Catheter infection
Ex: Duration of breastfeeding
Review of univariate case
Incorporation of covariates via the PH model
Estimating $\Lambda_0(t)$ and $S_0(t)$

This gives rise to the Nelson-Aalen estimator:

$$\hat{\Lambda}(t) = \int_0^t \hat{\lambda}(u) \, du$$

$$= \sum_{j: t_j \leq t} \hat{\lambda}(t_j)(\Delta t)$$

$$= \sum_{j: t_j \leq t} \frac{d_j}{n_j}$$

and we can think of the Kaplan-Meier estimator as:

$$\hat{S}(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

$$= \prod_{j: t_j \leq t} \left(1 - \hat{\lambda}(t_j)(\Delta t)\right)$$
Estimating $\Lambda_0(t)$ and $S_0(t)$

**Estimation of $\Lambda_0(t)$ and $S_0(t)$**

- Now, suppose we introduce covariates (via the PH model):

  \[ \lambda_i(t) = \lambda_0(t) \exp(\beta^T x_i) \]

  Then:

  \[
  E\{d_j \mid R_j\} = \sum_{i \in R_j} \Pr\{\text{ind } i \text{ fails in } (t_j - \Delta t, t_j) \mid \text{ind } i \text{ alive at } t_j - \Delta t\} \\
  \approx \sum_{i \in R_j} \lambda_i(t_j)(\Delta t) \\
  = \sum_{i \in R_j} \lambda_0(t_j) \exp(\beta^T x_i)(\Delta t) \\
  = \lambda_0(t_j)(\Delta t) \sum_{i \in R_j} \exp(\beta^T x_i)
  \]
Estimating $\Lambda_0(t)$ and $S_0(t)$

Again, estimate $\lambda_0(t)$ by replacing $E\{d_j \mid R_j\}$ with $d_j$

$$d_j = \hat{\lambda}_0(t_j)(\Delta t) \sum_{i \in R_j} \exp(\beta^T x_i)$$

or

$$\hat{\lambda}_0(t) = \begin{cases} 
  \frac{d_j}{(\Delta t) \sum_{i \in R_j} \exp(\beta^T x_i)} & , \quad t \in (t_j - \Delta t, t_j] \\
  0 & , \quad \text{otherwise}
\end{cases}$$

Interpretation of $S_0(t)$
Ex: Catheter infection
Ex: Duration of breastfeeding
Review of univariate case
Incorporation of covariates via the PH model
Estimating $\Lambda_0(t)$ and $S_0(t)$

So the baseline cumulative hazard can be estimated as

$$\hat{\Lambda}_0(t) = \sum_{j: t_j \leq t} \frac{d_j}{\sum_{i \in R_j} \exp(\beta^T x_i)}$$

and the baseline survival function as

$$\hat{S}_0(t) = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{\sum_{i \in R_j} \exp(\beta^T x_i)}\right)$$