SESSION 1:
REVIEW AND COX MODEL
FOR ADJUSTMENT AND INTERACTION

Module 8: Survival Analysis for Observational Data
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2016

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

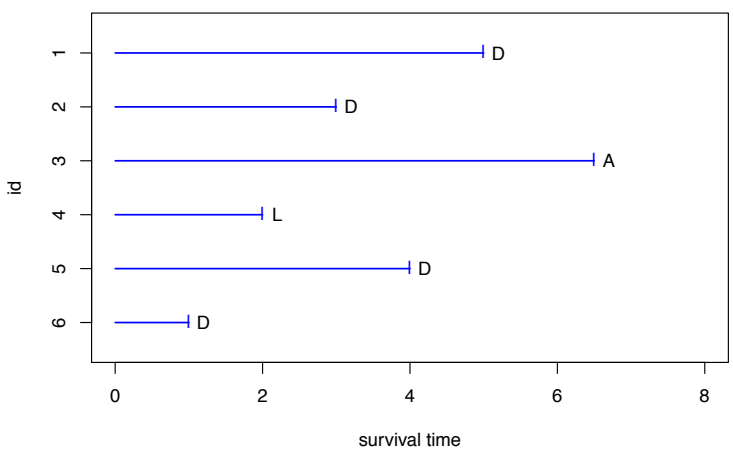
OVERVIEW

- Session 1
  - Quick review of introductory material
  - Adjustment in the Cox model: confounding and precision
  - Effect modification in the Cox model
- Session 2
  - Nonparametric hazard function estimation
  - Competing risks
  - Cumulative Incidence estimation
- Session 3
  - Left entry and left truncation
  - Choice of the time variable
  - Interactions with functions of time
- Session 4
  - Immortal time bias
  - Time-dependent covariates

OUTLINE

- Review of censored data, KM estimation, logrank test and Cox model basics
- Covariate adjustment in Cox model
- Stratification adjustment in Cox model
- Interaction (Effect Modification) in Cox Model
- Precision in Cox model

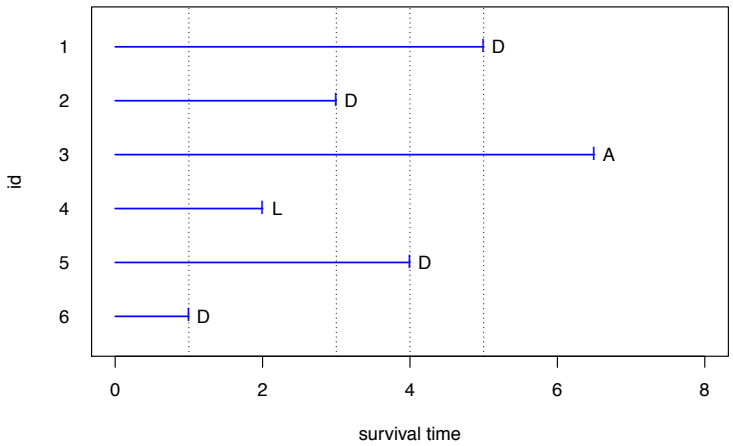
CENSORED DATA



id	Y	δ
1	5	1
2	3	1
3	6.5	0
4	2	0
5	4	1
6	1	1

"Censored" observations give some information about their survival time.

RISK SETS

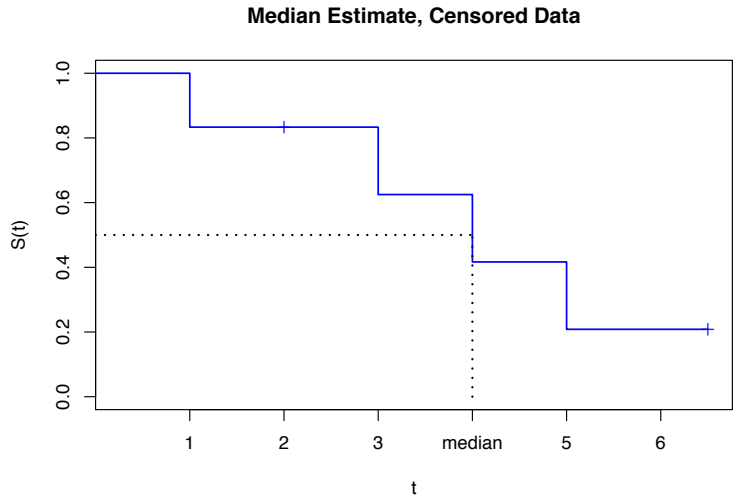


R_1 {1,2,3,4,5,6} R_2 {1,2,3,5} R_3 {1,3,5} R_4 {1,3}

CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.

MEDIAN & SURVIVAL CENSORED DATA



EQUIVALENT CHARACTERIZATIONS

- Any <u>one</u> of the density function(f(t)), the survival function(S(t)) or the hazard function(λ(t)) is enough to determine the survival distribution.
- They are each functions of each other:
  - $S(t) = \int_t^\infty f(s)ds = e^{-\int_0^t \lambda(s)ds}$
  - $f(t) = -\frac{d}{dt}S(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$
  - $\lambda(t) = \frac{f(t)}{S(t)}$

LOGRANK TEST

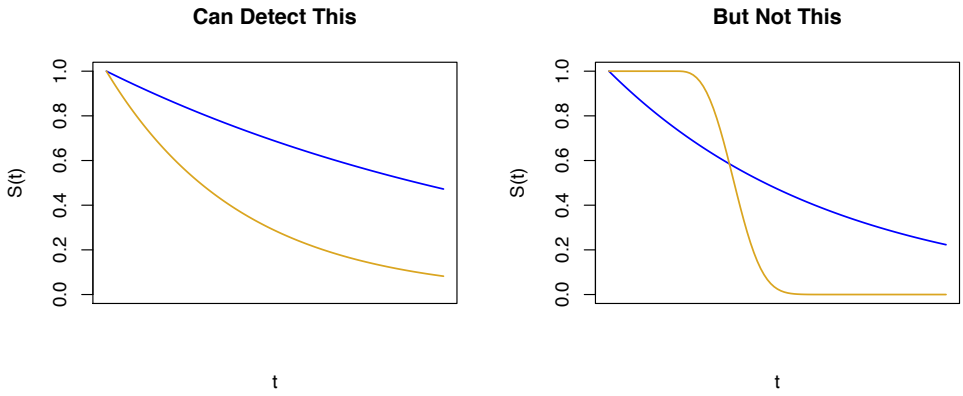
- The test is based on a 2x2 table of group by current status at each observed failure time (ie for each risk set)
- $T_{(j)}$, $j=1,\ldots m$, as shown in the Table below.

Event/Group	1	2	Total
Die	$d_{1(j)}$	$d_{2(j)}$	$D_{(j)}$
Survive	$n_{1(j)}-d_{1(j)}= s_{1(j)}$	$n_{2(j)}-d_{2(j)} = s_{2(j)}$	$N_{(j)}-D_{(j)} = S_{(j)}$
At Risk	$n_{1(j)}$	$n_{2(j)}$	$N_{(j)}$

LOGRANK TEST

- Detects consistent differences between survival curves over time.
- Best power when:
  - H_0: $S_1(t) = S_2(t)$ for all t vs H_A: $S_1(t) = [S_2(t)]^c$, or
  - H_0: $\lambda_1(t) = \lambda_2(t)$ for all t vs H_A: $\lambda_1(t) = c\,\lambda_2(t)$
- Good power whenever survival curve difference is in consistent direction

LOGRANK TEST



Other tests (generalized Wilcoxon and others) can give more weight to early or late differences.

COX REGRESSION MODEL

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

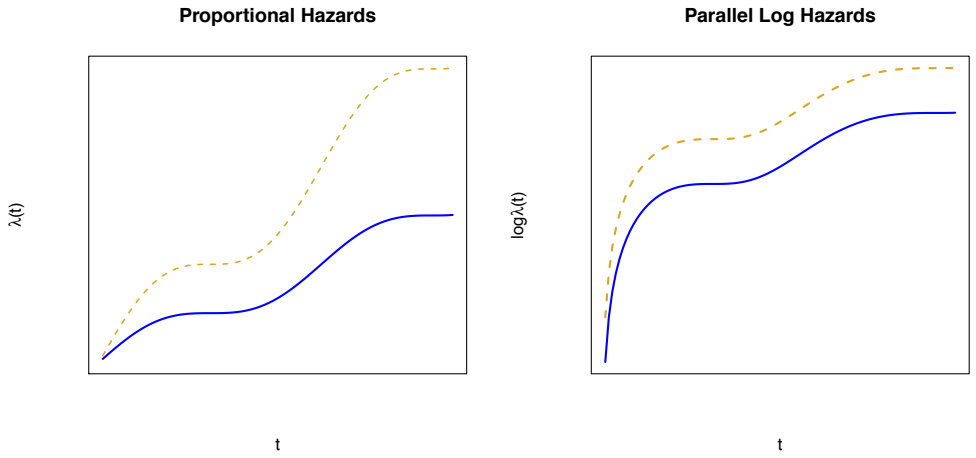
$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

↑ relative risk / hazard ratio

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

↑ intercept

EXAMPLE



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 $x = 1$: $\quad [S_0(t)]^{e^{\beta \cdot 1}} = [S_0(t)]^{e^{\beta}}$

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

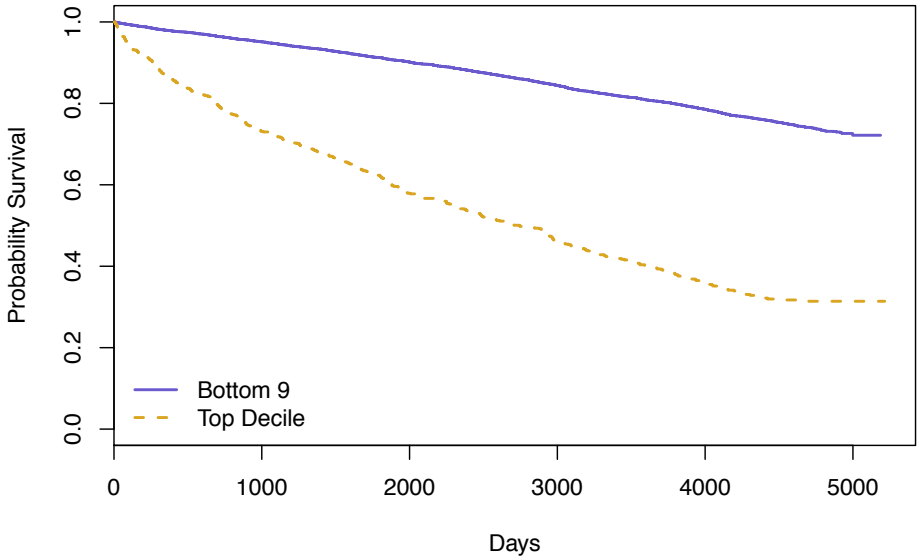
CONFOUNDING

- 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.
  - Age related to Ig levels and risk of death (example next)

SURVIVAL AND IG

- Random subset of the data from A. Dispenzieri, J. Katzmann, R. Kyle, D. Larson, T. Therneau, C. Colby, R. Clark, G. Mead, S. Kumar, L.J. Melton III, and S.V. Rajkumar. Use of monclonal serum immunoglobulin free light chains to predict overall survival in the general population. Mayo Clinic Proc, 87:512–523, 2012.
- Are high free-chain Ig levels associated with survival?
  - Population-based Olmstead County example
  - Men and women 50+ years of age

TOP DECILE FLC



COX REGRESSION

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
topdecileTRUE	1.452639	4.274378	0.0523126	27.7684	0

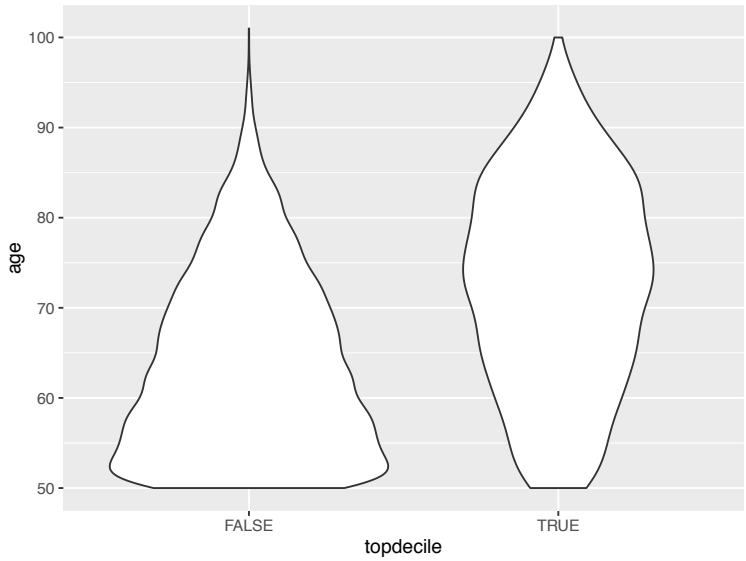
	2.5 %	97.5 %
topdecileTRUE	3.857841	4.735889

ADJUSTED COX REGRESSION

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
topdecileTRUE	0.8012613	2.228350	0.0543721	14.73663	0
age	0.1018649	1.107234	0.0022780	44.71700	0

	2.5 %	97.5 %
topdecileTRUE	2.003096	2.478934
age	1.102301	1.112189

WHY?



ADJUSTMENT MODEL

One binary variable, x_1, with continuous adjustment variable x_2:

$$x_1 = \begin{cases} 1 & \text{Top decile FLC} \\ 0 & \text{Otherwise} \end{cases}$$

$x_2 =$ Age in years

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

Interpretation of e^{β_1}:

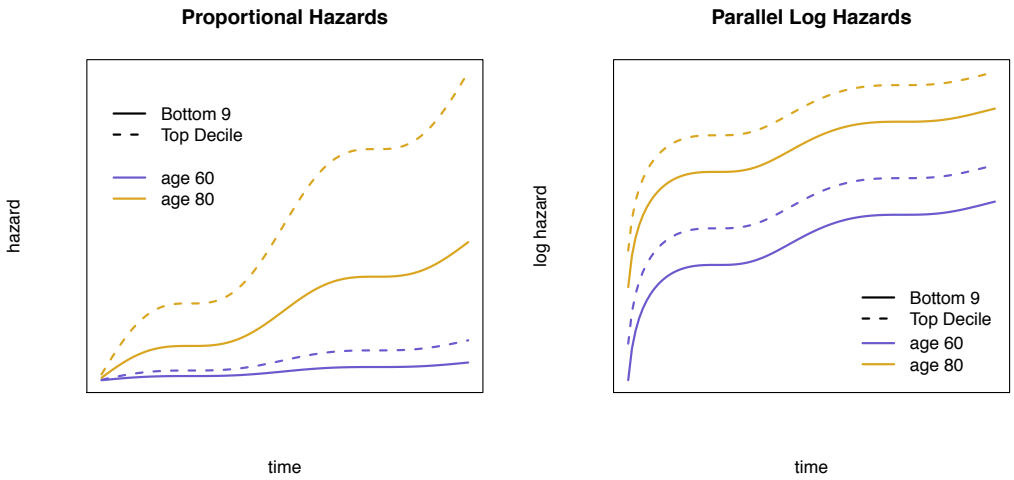
"Relative risk (or hazard ratio) comparing top decile FLC to the rest, among those of the same age".

$\lambda(t)$ for $x_1 = 1$ and x_2: $\lambda_0(t) e^{\beta_1 \cdot 1 + \beta_2 x_2}$

$\lambda(t)$ for $x_1 = 0$ and x_2: $\lambda_0(t) e^{\beta_1 \cdot 0 + \beta_2 x_2}$

ratio: $e^{\beta_1(1-0) + \beta_2(x_2 - x_2)} = e^{\beta_1}$

ADJUSTMENT



RESULTS

- "We found strong evidence that adjusted for age, free light chain(FLC) values in the top decile were associated with the risk of death (P < .0001). Among individuals of the same age, we estimate that having an FLC value in the top decile is associated with 2.23 times the hazard of death (95% CI: 3.86, 4.74)."

STRATIFICATION ADJUSTMENT

One binary variable, x_1, with grouped adjustment variable x_2:

$$x_1 = \begin{cases} 1 & \text{Top decile FLC} \\ 0 & \text{Otherwise} \end{cases}$$

$$x_2 = \begin{cases} 0 & \text{age 50-59} \\ 1 & \text{age 60-69} \\ 2 & \text{age 70-79} \\ 3 & \text{age 80-89} \\ 4 & \text{age 90+} \end{cases}$$

$$\lambda(t) = \lambda_{0x_2}(t)e^{\beta_1 x_1}$$

Interpretation of e^{β_1}:

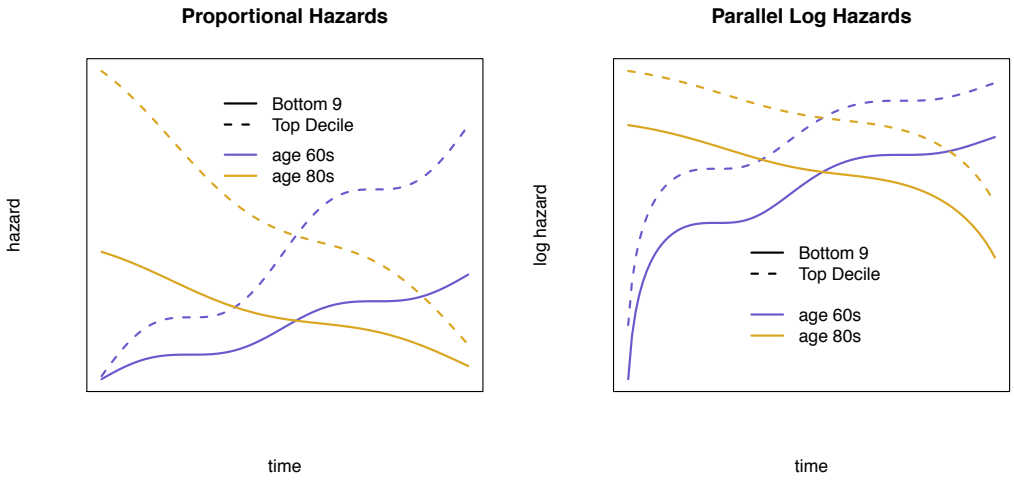
"Relative risk (or hazard ratio) comparing top decile FLC to the rest, among those in the same age group".

$\lambda(t)$ for $x_1 = 1$ and x_2: $\lambda_{0x_2}(t)e^{\beta_1 \cdot 1}$

$\lambda(t)$ for $x_1 = 0$ and x_2: $\lambda_{0x_2}(t)e^{\beta_1 \cdot 0}$

ratio: $\frac{\lambda_{0x_2}(t)}{\lambda_{0x_2}(t)}e^{\beta_1(1-0)} = e^{\beta_1}$

STRATIFICATION ADJUSTMENT



INTERACTION

One binary variable with continuous linear interaction, x_1 and x_2

$$x_1 = \begin{cases} 1 & \text{Top Decile FLC} \\ 0 & \text{Otherwise} \end{cases}$$

$x_2 =$ Age in years

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

Interpretation of e^{β_1}:

"Relative risk (or hazard ratio) comparing top decile FLC to the rest among those with age (= x_2) = zero".

Interpretation of $e^{\beta_1 + x_2\beta_3}$:

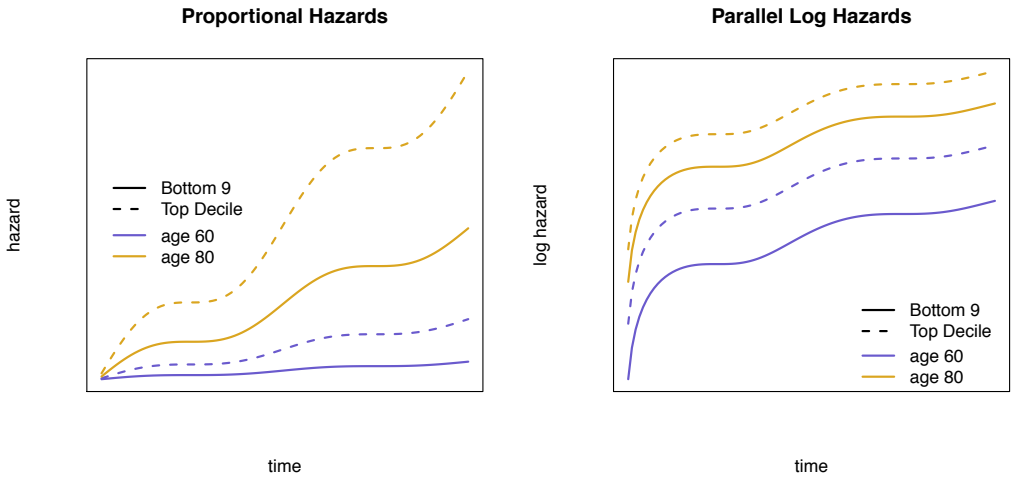
"Relative risk (or hazard ratio) comparing top decile FLC to the rest among those with age = x_2".

$\lambda(t)$ for $x_1 = 1$ and $x_2 = 0$: $\lambda_0(t)e^{\beta_1 \cdot 1}$ $\lambda(t)$ for $x_1 = 1$ and $x_2 \neq 0$: $\lambda_0(t)e^{\beta_1 \cdot 1 + \beta_2 \cdot x_2 + \beta_3 \cdot 1 \cdot x_2}$

$\lambda(t)$ for $x_1 = 0$ and $x_2 = 0$: $\lambda_0(t)e^{\beta_1 \cdot 0}$ $\lambda(t)$ for $x_1 = 0$ and $x_2 \neq 0$: $\lambda_0(t)e^{\beta_1 \cdot 0 + \beta_2 \cdot x_2 + \beta_3 \cdot 0}$

ratio: $e^{\beta_1(1-0)} = e^{\beta_1}$ ratio: $e^{\beta_1(1-0)+\beta_3(x_2-0)} = e^{\beta_1 + x_2\beta_3}$

INTERACTION



INTERACTION

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
topdecileTRUE	2.7312322	15.3517922	0.4154009	6.574930	0.0e+00
age	0.1067648	1.1126726	0.0025185	42.392311	0.0e+00
topdecileTRUE:age	-0.0252304	0.9750852	0.0054342	-4.642936	3.4e-06

	2.5 %	97.5 %
topdecileTRUE	6.8009436	34.6536508
age	1.1071938	1.1181785
topdecileTRUE:age	0.9647549	0.9855261

TOP DECILE HR BY AGE

age	exp(coef)	z	Pr(>|z|)	2.5 %	97.5 %
50	3.897886	8.499784	0.00e+00	2.848328	5.334189
60	3.077554	10.309487	0.00e+00	2.485373	3.810831
70	2.429865	13.162515	0.00e+00	2.128957	2.773302
80	1.918486	10.861243	0.00e+00	1.705679	2.157843
90	1.514729	4.368336	1.25e-05	1.257254	1.824932

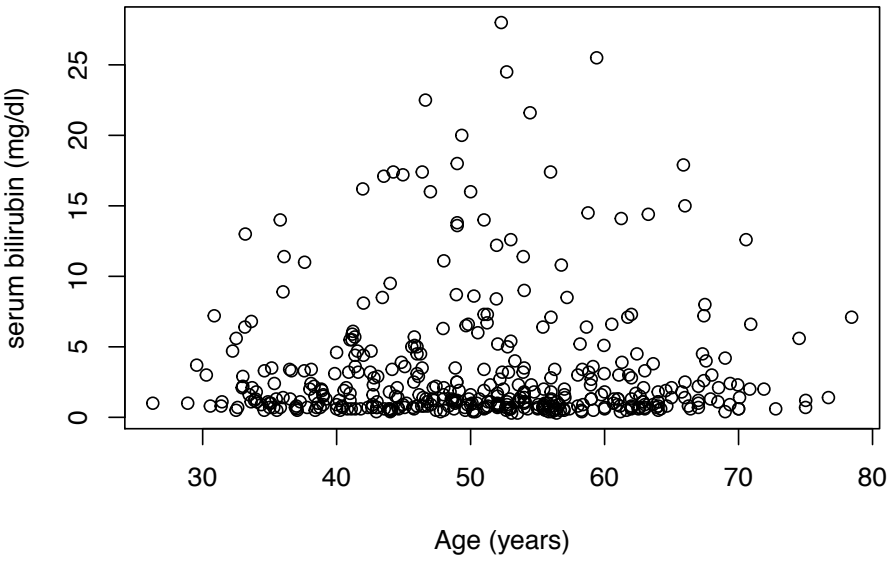
ADJUSTMENT AND PRECISION

- In Cox regression, addition of variables to a model that are associated only with the outcome can improve power.
- There is little effect on the coefficient estimate for other variables (eg treatment) or their standard errors, except when the association between outcome and the added variable is very strong.
- When there is an effect of adding a predictive variable, this is what happens to inference for the treatment variable or other variable of interest:
  - The standard error of its coefficient increases
  - The estimate of the coefficient moves farther from zero
  - The test of whether the coefficient is zero has more power.

PRIMARY BILIARY CIRRHOSIS

- Clinical trial with virtually no treatment effect
- Conducted before widespread use of immune suppressive therapies
- Good data for examining prognostic factors in PBC
- Some patients received liver transplant—treated as censored here
- Serum bilirubin associated with survival
- Treating age as a "precision variable"

AGE-BILIRUBIN ASSOCIATION



PRECISION

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
bili	0.1418533	1.152408	0.0115685	12.26201	0

	2.5 %	97.5 %
bili	1.126572	1.178836

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
bili	0.1436238	1.154450	0.0114189	12.577714	0e+00
age	0.0431303	1.044074	0.0080554	5.354198	1e-07

	2.5 %	97.5 %
bili	1.128899	1.180578
age	1.027719	1.060689

TO WATCH OUT FOR:

- Coefficients in Cox regression are positively associated with risk, not survival.
  - Positive β means large values of x are associated with shorter survival.
- Without certain types of time-dependent covariates (more later), Cox regression does not depend on the actual times, just their order.
  - Can add a constant to all times to remove zeros (which are removed by some software) 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.
- Coefficient interpretation depends on what other variables are in the model and how they are coded (ie. interaction terms, 0/1 vs 1/-1 etc.)

In R

Load packages.

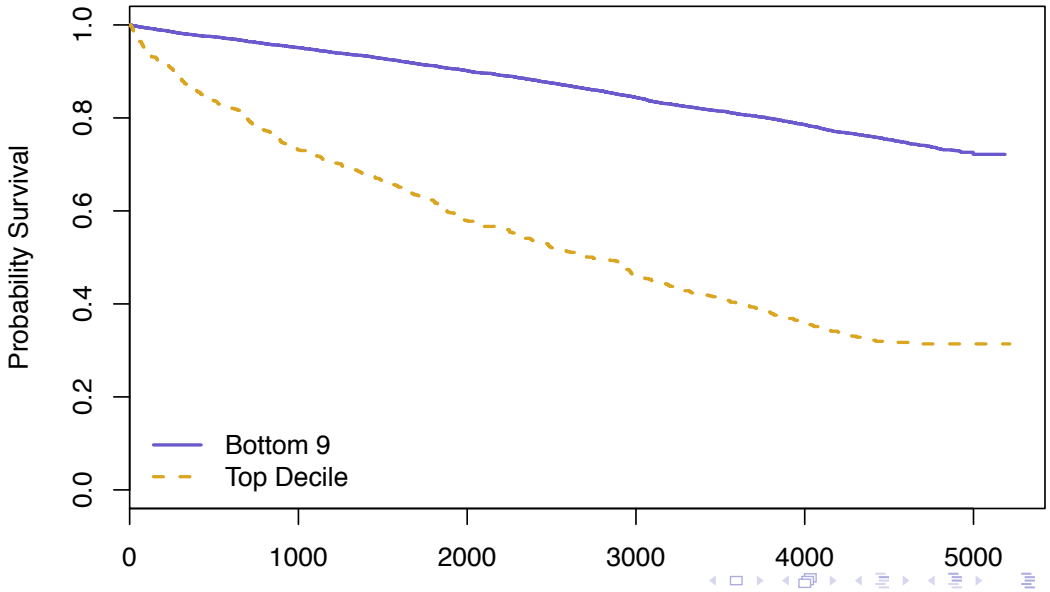
```
library(survival)
library(ggplot2)
```

Get data.

```
df <- flchain[flchain$futime > 7,]
Y <- with(df, Surv(futime, death))
```

Make binary exposure and plot

```
df$topdecile <- df$flc.grp > 9
colors <- c("slateblue", "goldenrod")
plot(survfit(Y ~ topdecile, data = df), xlab = "Days",
       ylab = "Probability Survival", lty = 1:2,
     col = colors, lwd = 2)
legend("bottomleft", legend = c("Bottom 9","Top Decile" ),
       lty = 1:3, col = colors, lwd = 2, bty = "n")
```



Confounding adjustment

```
crude <- coxph(Y ~ topdecile, data = df)
adj.age <- coxph(Y ~ topdecile + age, data = df)
```

Crude model estimates

```
coef(summary(crude))
```

```
##                    coef exp(coef)   se(coef)       z Pr(>|z|)
## topdecileTRUE 1.452639  4.274378 0.05231265 27.7684        0
```

```
exp(confint(crude))
```

```
##                   2.5 %   97.5 %
## topdecileTRUE 3.857841 4.735889
```

Adjusted estimates

```
coef(summary(adj.age))
```

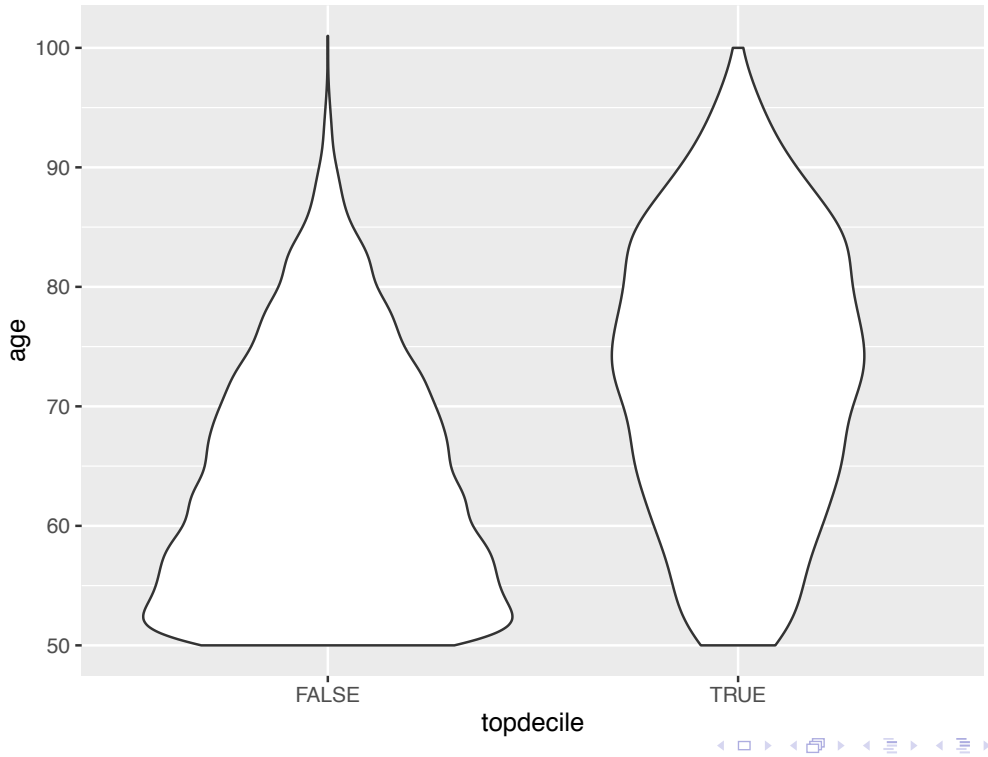
```
##                     coef exp(coef)    se(coef)        z Pr(>|z|)
## topdecileTRUE 0.8012613  2.228350 0.054372073 14.73663        0
## age           0.1018649  1.107234 0.002277991 44.71700        0
```

```
exp(confint(adj.age))
```

```
##                   2.5 %   97.5 %
## topdecileTRUE 2.003096 2.478934
## age           1.102301 1.112189
```

Association with age

```
ggplot(df, aes(x = topdecile, y = age)) + geom_violin()
```



Interaction

```
inter <- coxph(Y ~ topdecile * age, data = df)
```

Interaction

```
coef(summary(inter))
```

```
##                           coef  exp(coef)    se(coef)         z
## topdecileTRUE      2.73123222 15.3517922 0.415400929  6.574930
## age                0.10676484  1.1126726 0.002518495 42.392311
## topdecileTRUE:age -0.02523044  0.9750852 0.005434156 -4.642936
##                       Pr(>|z|)
## topdecileTRUE     4.867595e-11
## age               0.000000e+00
## topdecileTRUE:age 3.434930e-06
```

```
exp(confint(inter))
```

```
##                       2.5 %     97.5 %
## topdecileTRUE     6.8009436 34.6536508
## age               1.1071938  1.1181785
## topdecileTRUE:age 0.9647549  0.9855261
```

HRs at different ages

```
df$age50 <- df$age - 50; df$age60 <- df$age - 60;
df$age70 <- df$age - 70; df$age80 <- df$age - 80;
df$age90 <- df$age - 90
age.50 <- coxph(Y ~ creatinine + topdecile * age50, data = df)
age.60 <- coxph(Y ~ creatinine + topdecile * age60, data = df)
age.70 <- coxph(Y ~ creatinine + topdecile * age70, data = df)
age.80 <- coxph(Y ~ creatinine + topdecile * age80, data = df)
age.90 <- coxph(Y ~ creatinine + topdecile * age90, data = df)
```

HRs at different ages

```
cbind(coef(summary(age.50)), exp(confint(age.50)))
```

```
##                             coef exp(coef)    se(coef)         z
## creatinine           0.24001508 1.2712683 0.029737172  8.071214
## topdecileTRUE        1.39040898 4.0164924 0.160671028  8.653763
## age50                0.10417739 1.1097973 0.002671625 38.994011
## topdecileTRUE:age50 -0.02490311 0.9754044 0.005699952 -4.369003
##                         Pr(>|z|)     2.5 %    97.5 %
## creatinine          6.661338e-16 1.1992919 1.3475645
## topdecileTRUE       0.000000e+00 2.9314569 5.5031376
## age50               0.000000e+00 1.1040013 1.1156238
## topdecileTRUE:age50 1.248150e-05 0.9645681 0.9863624
```

```
age.specific <- rbind(cbind(coef(summary(age.50)),
                            exp(confint(age.50)))[2,],
cbind(coef(summary(age.60)), exp(confint(age.60)))[2,],
cbind(coef(summary(age.70)), exp(confint(age.70)))[2,],
cbind(coef(summary(age.80)), exp(confint(age.80)))[2,],
cbind(coef(summary(age.90)), exp(confint(age.90)))[2,]
)
```

HRs at different ages

```
cbind(age = c(5:9)*10, age.specific[,-c(1,3)])
```

```
##      age exp(coef)         z     Pr(>|z|)    2.5 %   97.5 %
## [1,]  50  4.016492  8.653763 0.000000e+00 2.931457 5.503138
## [2,]  60  3.131080 10.446051 0.000000e+00 2.527484 3.878822
## [3,]  70  2.440851 13.214546 0.000000e+00 2.138266 2.786255
## [4,]  80  1.902779 10.605874 0.000000e+00 1.689497 2.142986
## [5,]  90  1.483323  4.089344 4.325942e-05 1.227905 1.791870
```

Precision

```
data(pbc)
Y2 <- with(pbc, Surv(time, status == 2))
pbc.crude <- coxph(Y2 ~ bili, data = pbc)
precision <- coxph(Y2 ~ bili + age, data = pbc)
```

Precision

```
coef(summary(pbc.crude))
```

```
##           coef exp(coef)   se(coef)        z Pr(>|z|)
## bili 0.1418533  1.152408 0.01156852 12.26201        0
```

```
exp(confint(pbc.crude))
```

```
##         2.5 %   97.5 %
## bili 1.126572 1.178836
```

```
coef(summary(precision))
```

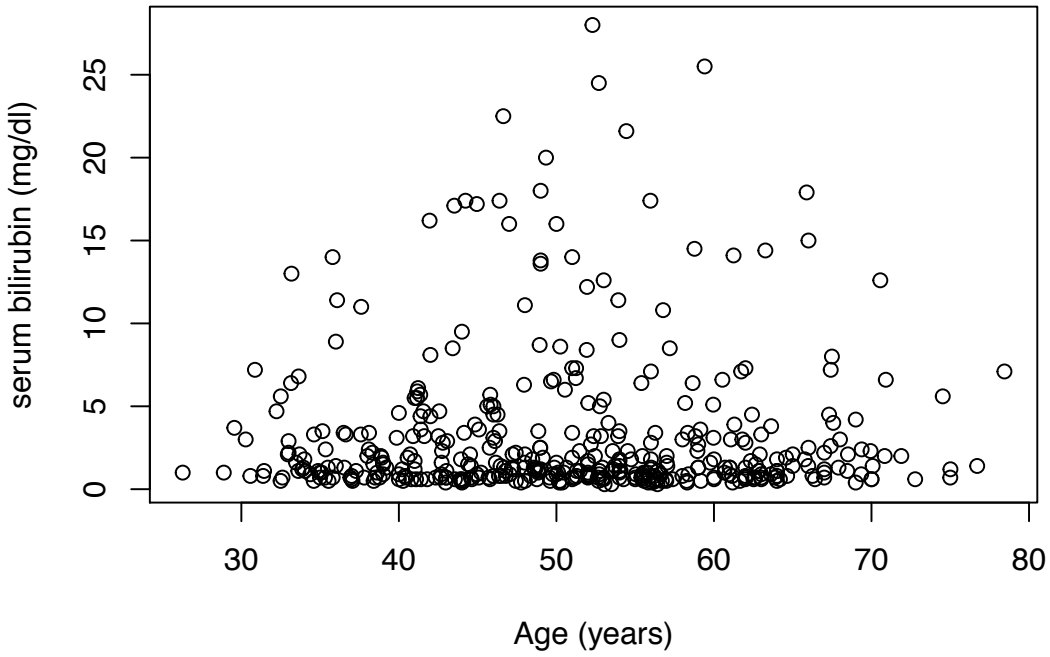
```
##            coef exp(coef)    se(coef)         z     Pr(>|z|)
## bili 0.14362385  1.154450 0.011418915 12.577714 0.000000e+00
## age  0.04313034  1.044074 0.008055425  5.354198 8.593669e-08
```

```
exp(confint(precision))
```

```
##         2.5 %   97.5 %
## bili 1.128899 1.180578
## age  1.027719 1.060689
```

Precision

```
with(pbc,plot(age, bili, xlab = "Age (years)",
              ylab = "serum bilirubin (mg/dl)"))
```



Your turn

Using the flchain data in the survival package:

1. Test whether, adjusted for top decile of free light chain, there is evidence that the HR associated with creatinine depends on age.
2. Create a table of age-specific, top-decile-of-free-light-chain-sadjusted HRs and CIs associated with a 1 mg/dL higher creatinine level for ages 50, 60, 70,80, 90.

SESSION 2: HAZARD ESTIMATION, COMPETING RISKS, AND CUMULATIVE INCIDENCE

Module 8: Survival Analysis for Observational Data

Summer Institute in Statistics for Clinical Research
University of Washington
July, 2016

Barbara McKnight, Ph.D.

OUTLINE

- Nonparametric hazard function estimation
- Competing risks:
  - Definition: when there is more than one cause of death/failure
  - Cause-specific hazards
  - Effect of removing other types of failure
  - Cumulative functions:
    - Event-free survival
    - Cumulative Incidence estimator

HAZARD FUNCTION

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t | T \geq t]$$

- Instantaneous rate at which death occurs at t in those who are alive at t
- Examples:
  - Age-specific death rate
  - Age-specific disease incidence rate

CUMULATIVE HAZARD FUNCTION

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

$$= \text{area under the hazard function curve between 0 and } t.$$

$$= \text{amount of “hazard" accumulated between 0 and } t.$$

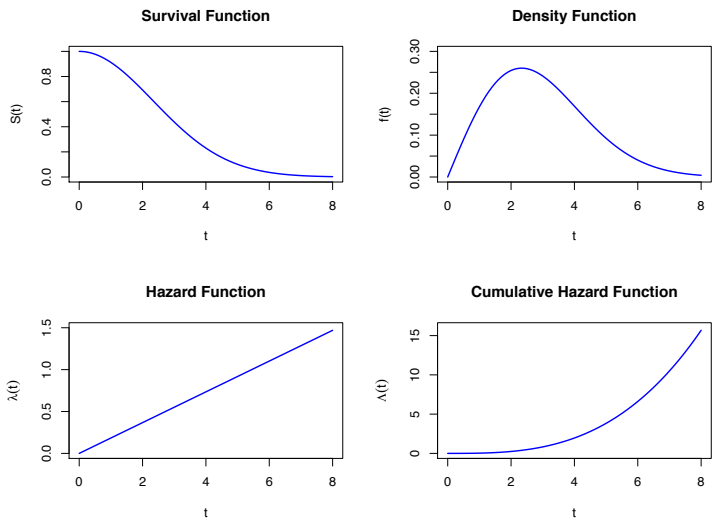
$$= -\log(S(t))$$

Not usually of interest *per se*, but estimates useful for diagnostics.

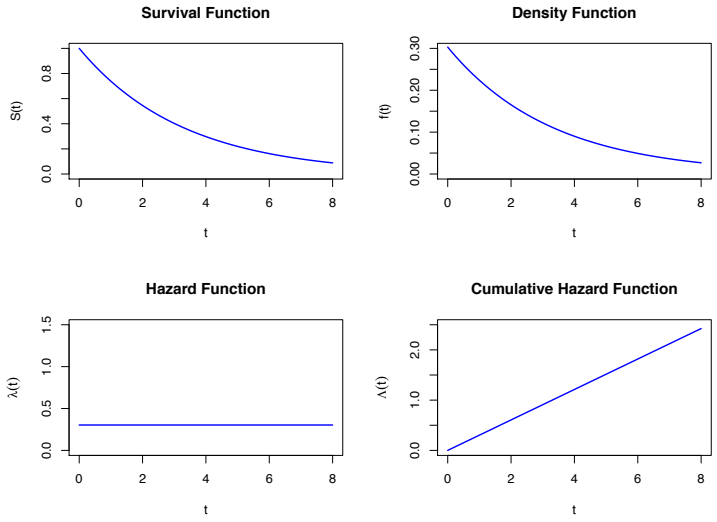
EQUIVALENT CHARACTERIZATIONS

- Any <u>one</u> of these four functions is enough to determine the survival distribution.
- They are each functions of each other:
  - $S(t) = \int_t^\infty f(s)ds = e^{-\int_0^t \lambda(s)ds}$
  - $f(t) = -\frac{d}{dt}S(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$
  - $\lambda(t) = \frac{f(t)}{S(t)}$
  - $\Lambda(t) = \int_0^t \lambda(s)ds = -\log(S(t))$

EQUIVALENT CHARACTERIZATIONS



EQUIVALENT CHARACTERIZATIONS



CUMULATIVE HAZARD

- Nelson - Aalen estimator:

  $\hat{\Lambda}(t) = \sum_{j:t_{(j)} \le t} \frac{D_{(j)}}{N_{(j)}}$

- Variance:

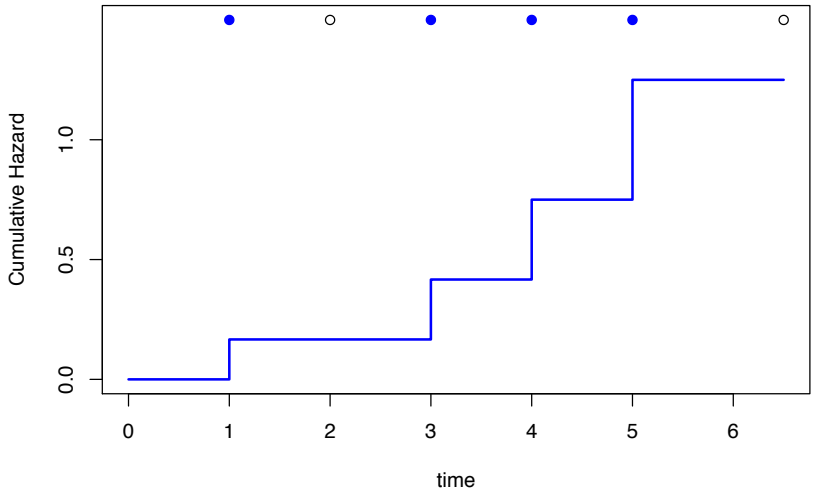
  $\widehat{Var}(\hat{\Lambda}(t)) = \sum_{j:t_{(j)} \le t} \frac{D_{(j)} S_{(j)}}{[N_{(j)}]^3}$

- Standard error:

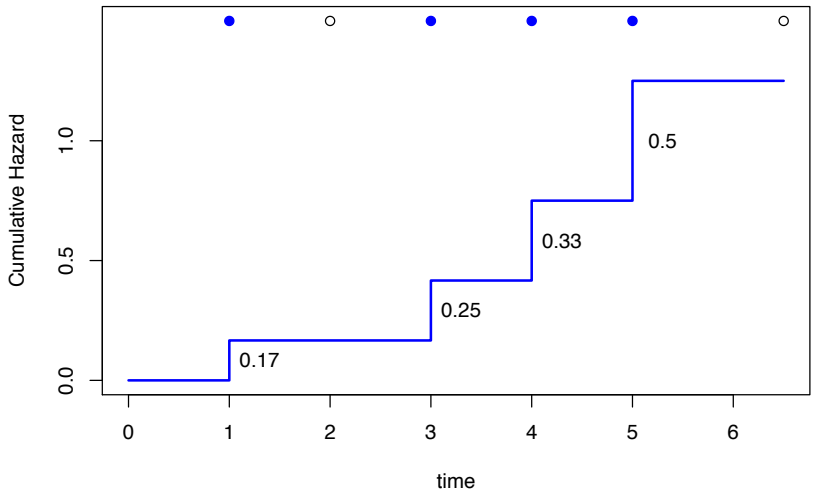
  $\sqrt{\widehat{Var}(\hat{\Lambda}(t))}$ can be used to form pointwise CI's.

- or could use $\hat{\Lambda}(t) = -\log(\hat{S}(t))$

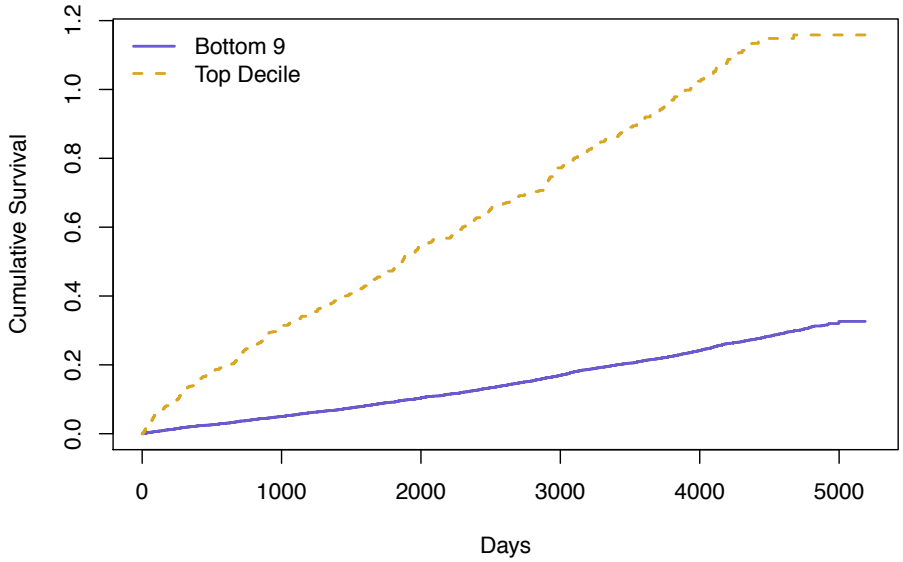
CUMULATIVE HAZARD FUNCTION



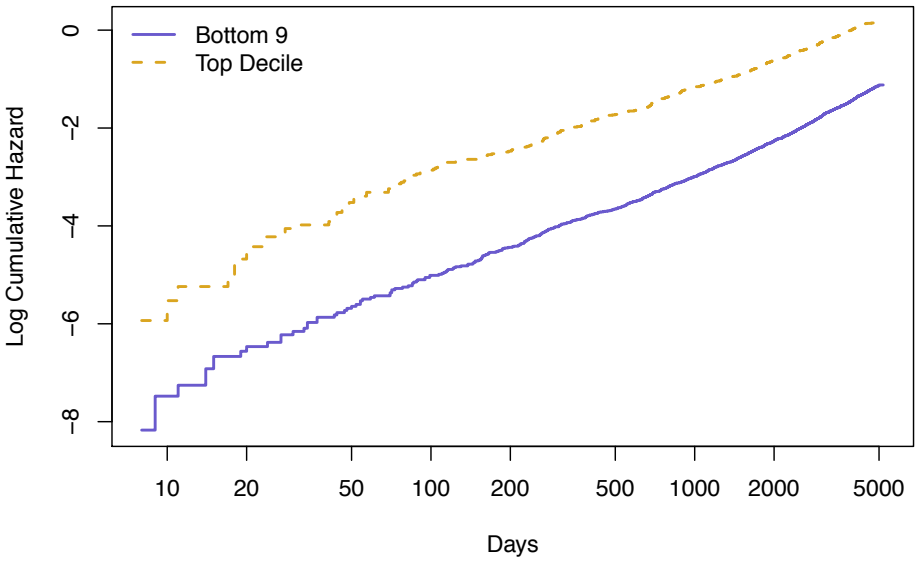
CUMULATIVE HAZARD FUNCTION



FLC EXAMPLE



FLC EXAMPLE



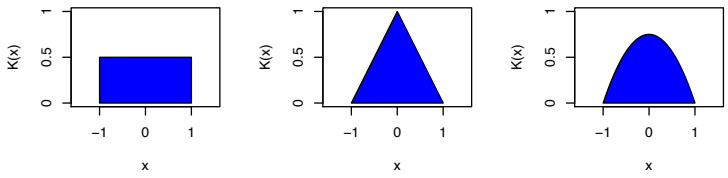
HAZARD FUNCTION

IDEA:

At each time t, let estimate of $\lambda(t)$ be a weighted average of jumps in $\hat{\Lambda}(t)$ at nearby times.

Steps:

1. Choose a "bandwidth" $\pm b$ outside of which observations are not averaged.
2. Choose a "kernel" or weight function $K(\cdot)$.

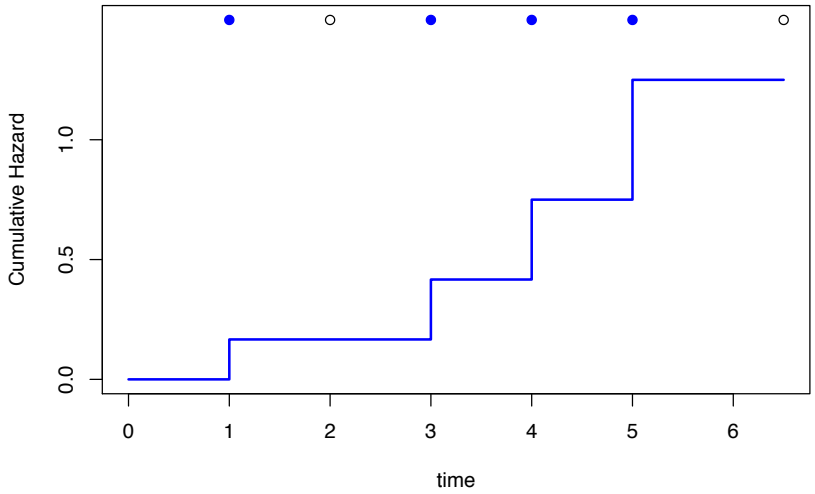


HAZARD FUNCTION

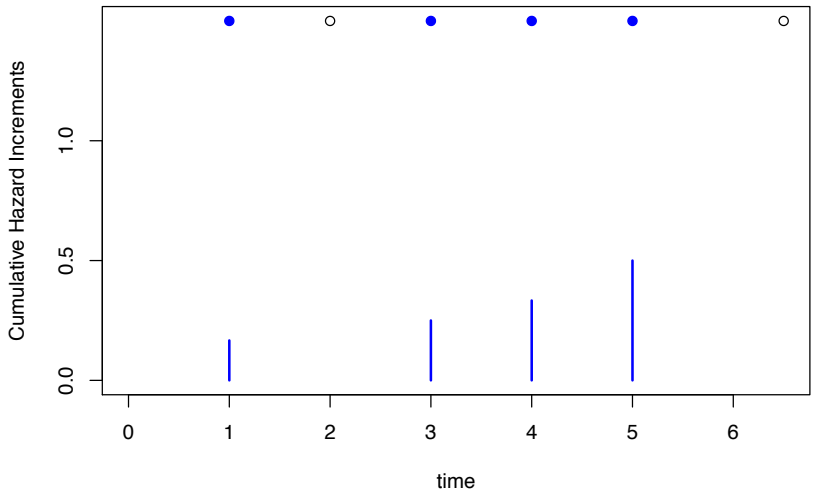
3. Calculate the estimate:

- $\hat{\lambda}(t) = \frac{1}{b}\sum_{j=1}^{J} K(\frac{t-t_{(j)}}{b})\frac{D_{(j)}}{N_{(j)}}$
- $se(\hat{\lambda}(t)) = \frac{1}{b}\left\{\sum_{j=1}^{J} K^2(\frac{t-t_{(j)}}{b})\frac{D_{(j)}}{N_{(j)}^2}\right\}^{\frac{1}{2}}$

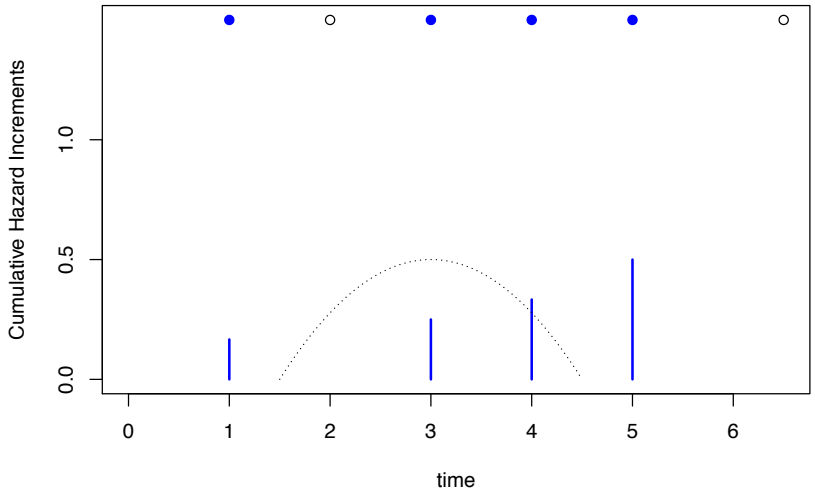
CUMULATIVE HAZARD FUNCTION



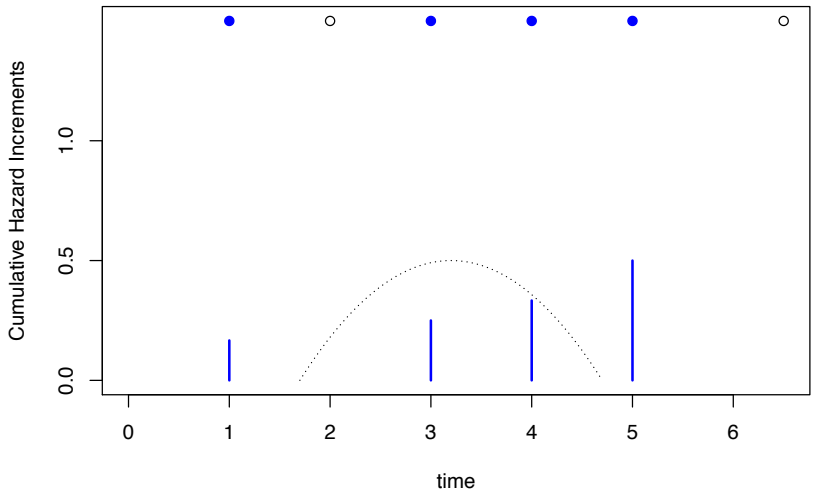
KERNEL HAZARD ESTIMATE



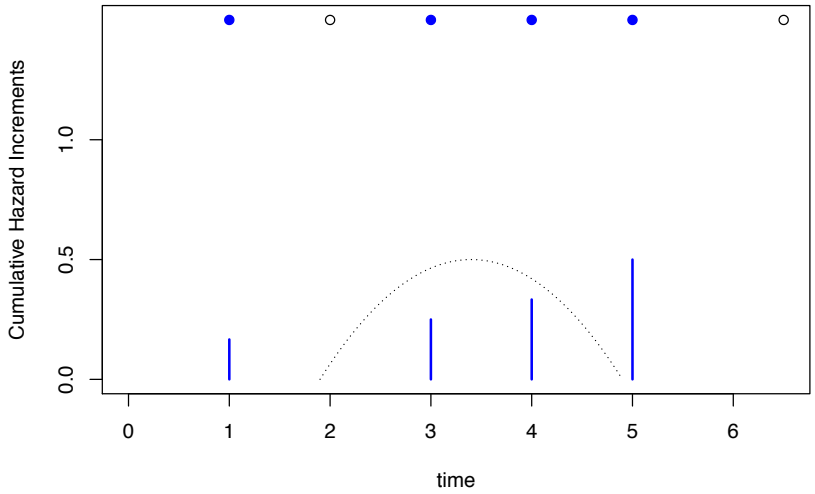
KERNEL HAZARD ESTIMATE



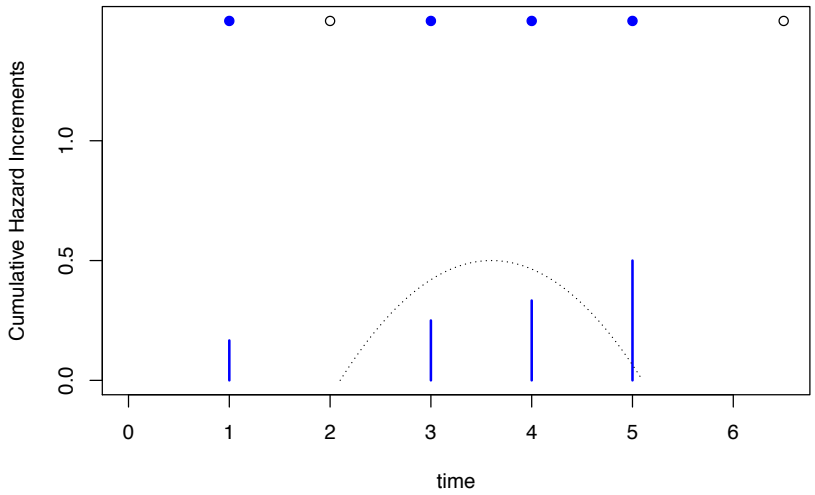
KERNEL HAZARD ESTIMATE



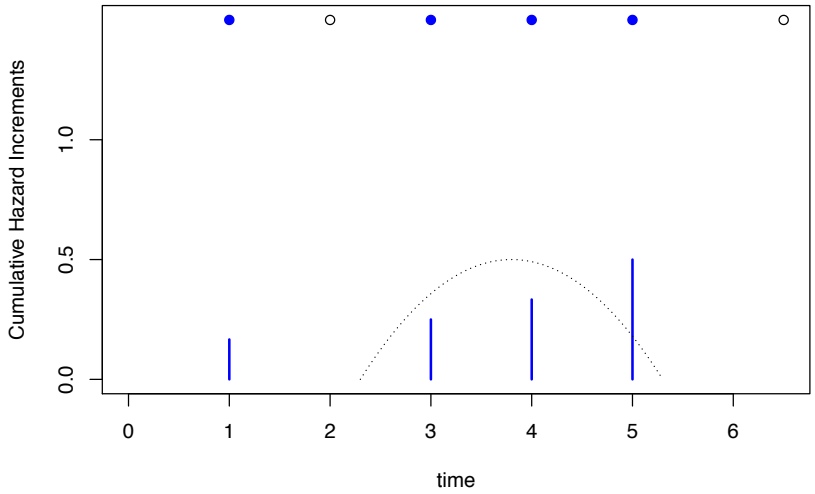
KERNEL HAZARD ESTIMATE



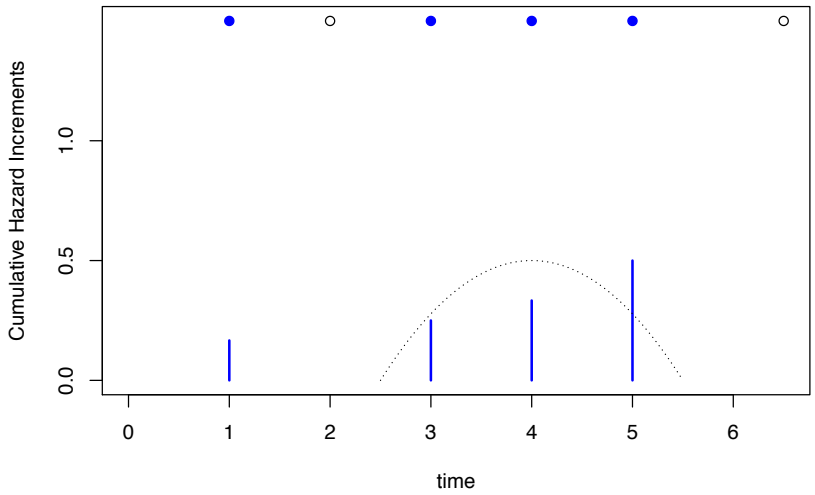
KERNEL HAZARD ESTIMATE



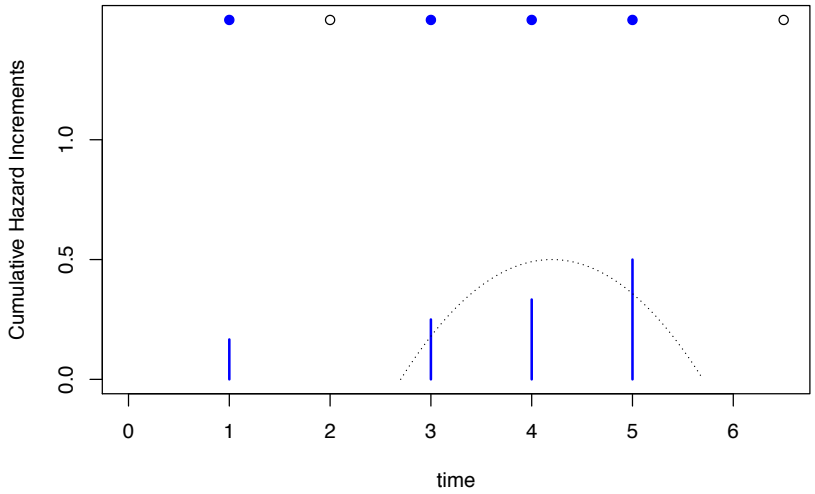
KERNEL HAZARD ESTIMATE



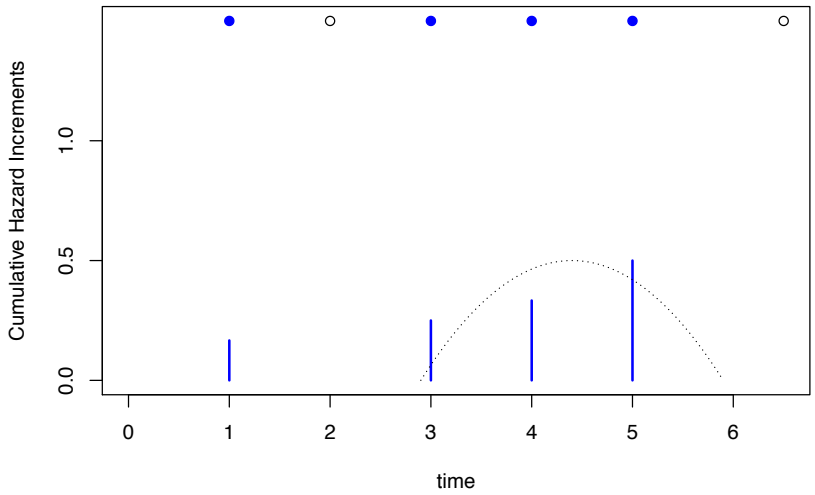
KERNEL HAZARD ESTIMATE



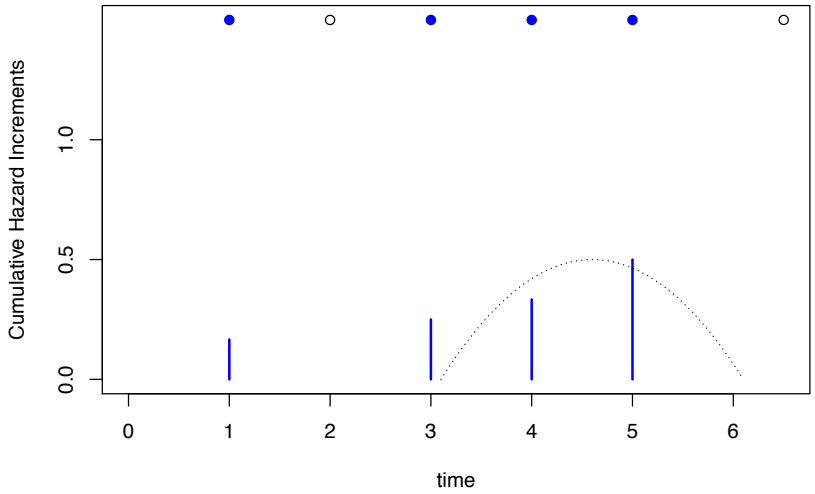
KERNEL HAZARD ESTIMATE



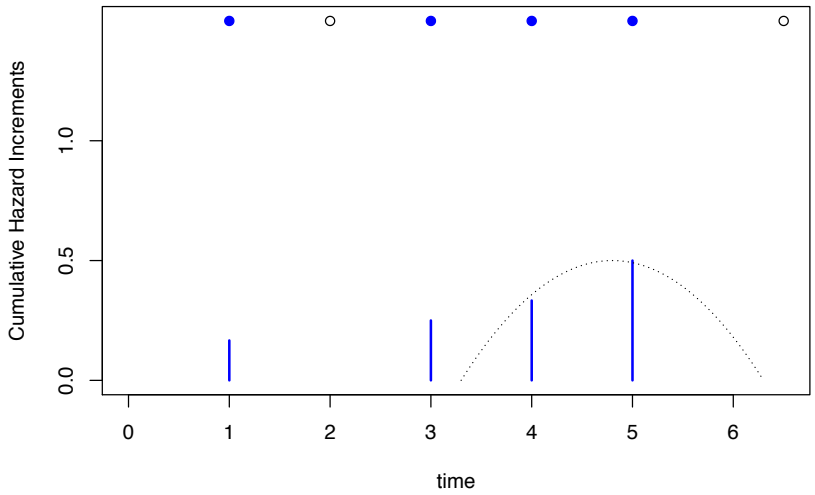
KERNEL HAZARD ESTIMATE



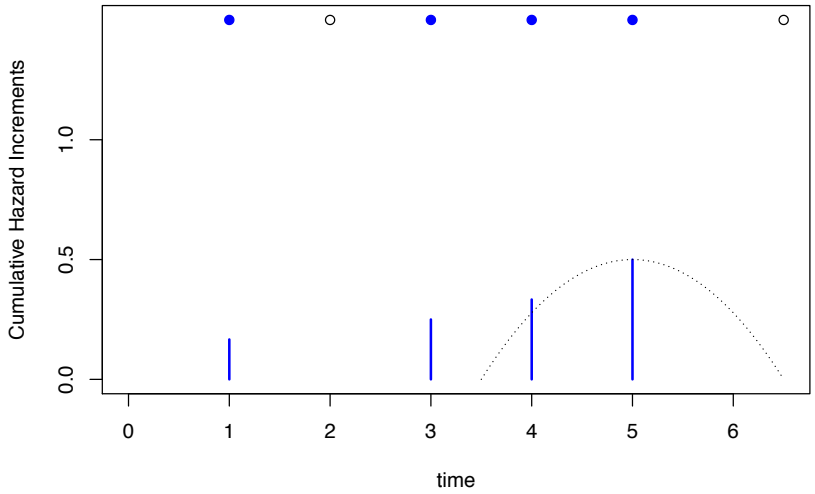
KERNEL HAZARD ESTIMATE



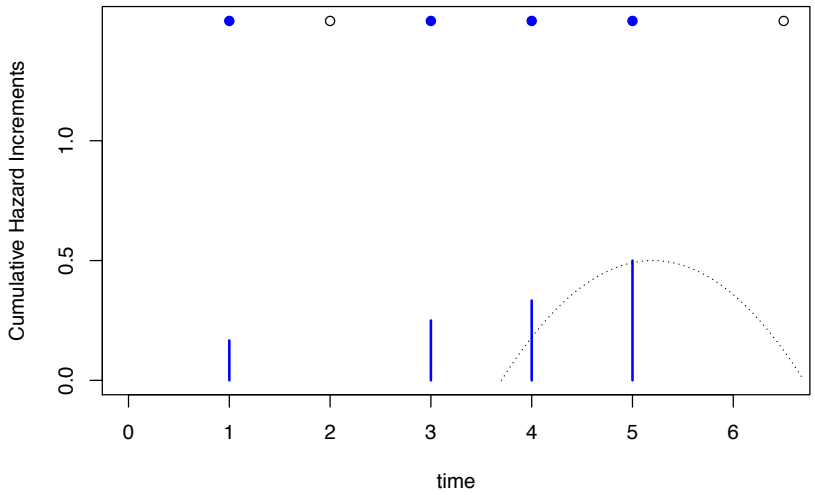
KERNEL HAZARD ESTIMATE



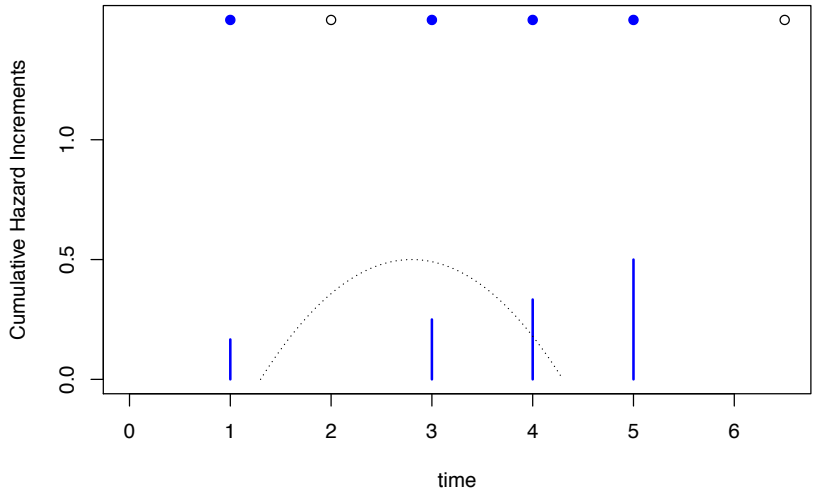
KERNEL HAZARD ESTIMATE



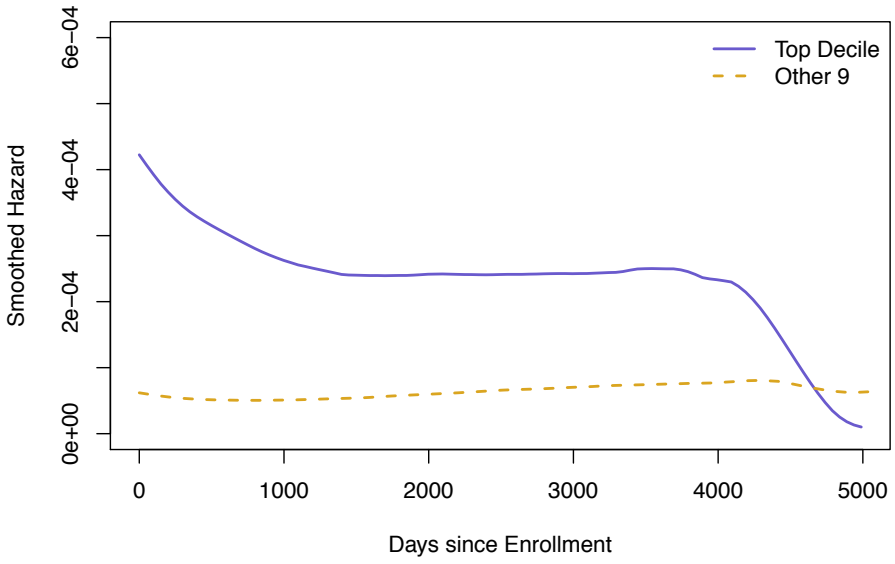
KERNEL HAZARD ESTIMATE



KERNEL HAZARD ESTIMATE



FLC EXAMPLE



COMPETING RISKS

- When there is more than one cause of failure:
  - Recurrence or death before recurrence
  - MI, stroke, PE or death from other causes
- The different types of failure are called "competing risks".
  - They "compete" to be the first to make subjects experience an event
- Suppose we are interested in just one of the competing risks
  - Recurrence
  - MI

COMPETING RISKS

- A random variable T = time to event of interest, doesn't exist for subjects who die from another cause before the event of interest occurs.
  - T = time to MI (before other CVD event) doesn't exist for those who have a stroke, PE or die of other causes first.
- A distribution for T described by $S(t), \Lambda(t),$ and $\lambda(t)$ does not make sense.
  - Not defined for those who fail of a competing cause before t.
- Some think the difficulty can be solved by defining $T = \infty$ for events that never happen, but this does not resolve the conceptual issues.

FUNCTIONS OF TIME

- Some functions that make sense:
  - Cause-specific hazard function
  - Probability of being alive and event-of-interest free at time t (progression-free survival)
  - Probability event of interest will happen by time t (Cumulative incidence function)
    - Not the same as 1 – S(t)

CAUSE-SPECIFIC HAZARD FUNCTION

The "cause-specific hazard" does make sense.

Let

T = time to first "failure" of any type
c = 1 failure is of the 1^{st} type
= 2 failure is of the 2^{nd} type
⋮
= K failure is of the K^{th} type.

Cause-specific hazard function for k^{th} cause:

$$\lambda^{(k)}(t) = \lim_{\Delta t \to 0} \Pr[T \in [t, t+\Delta t), c = k | T \geq t]/\Delta t$$

INTERPRETATION

Cause-specific hazard function:

$$\lambda^{(k)}(t) = \lim_{\Delta t \to 0} \Pr[T \in [t, t+\Delta t), c = k | T \geq t] / \Delta t$$

Interpretation:

Instantaneous risk of a type-k failure at time t, among those who have not yet failed due to any cause by just before time t.

Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The Analysis of Failure Times in the Presence of Competing Risks. Biometrics. 1978 Dec 1;34(4):541–554.

EXAMPLE 1

$$\lambda^{(k)}(t) = \lim_{\Delta t \to 0} \Pr[T \in [t, t+\Delta t), c = k | T \geq t] / \Delta t$$

Example:

T = time to recurrence or death (whichever first)

$$c = \begin{cases} 1 & \text{recurrence} \\ 2 & \text{death, no recurrence} \end{cases}$$

$\lambda^{(1)}(t)$ = recurrence rate at t among those alive at t.

EXAMPLE 2

$$\lambda^{(k)}(t) = \lim_{\Delta t \to 0} \Pr[T \in [t, t+\Delta t), c = k | T \geq t] / \Delta t$$

Example:

T = time to first CVD event or death (whichever first)

c = $\begin{cases} 1 & \text{MI} \\ 2 & \text{stroke} \\ 3 & \text{PE} \\ 4 & \text{Death, no CVD event} \end{cases}$

$\lambda^{(1)}(t)$ = MI rate at t among those alive and CVD event-free at t.

$\lambda^{(2)}(t)$ = stroke rate at t among those alive and CVD event-free at t.

$\lambda^{(3)}(t)$ = PE rate at t among those alive and CVD event-free at t.

PROPERTIES

T = time to first "failure" of any type

$$\lambda^{(k)}(t) = \lim_{\Delta t \to 0} \Pr[T \in [t, t+\Delta t), c = k | T \geq t] / \Delta t$$

- The different events defined by c must be mutually exclusive
- The different events defined by c must be exhaustive
- The hazard function for the distribution of T is given by :

$$\lambda(t) = \sum_{k=1}^{K} \lambda^{(k)}(t)$$

INFERENCES FOR CAUSE-SPECIFIC HAZARD

- Estimation of $\Lambda^{(k)}(t)$ and $\lambda^{(k)}(t)$
- P - sample tests for heterogeneity and trend in the $\lambda^{(k)}(t)$, including the logrank test
- Fits of the Cox regression model $\lambda^{(k)}(t) = \lambda_0^{(k)}(t)e^{\beta_1 x_1 + \cdots + \beta_p x_p}$
- Estimation of $\Lambda_0^{(k)}(t)$ and $\lambda_0^{(k)}(t)$ after a Cox model fit

Can all be performed in the usual way, using (Y, δ) data, where

$$\begin{aligned} Y &= \min(T, \text{time of LFU}) \\ \delta &= \begin{cases} 1 & c = k \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

For observed, cause-specific failure time $t_{(j)}$:

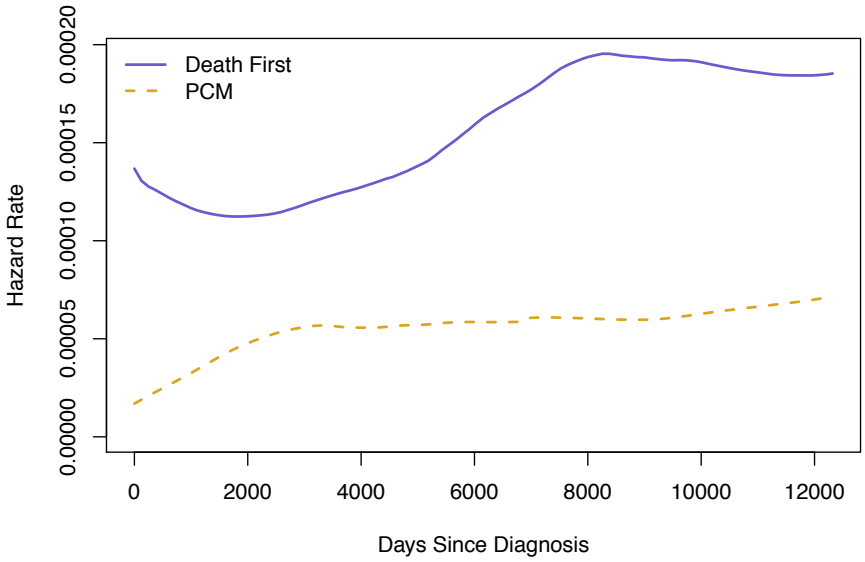
$\frac{D_{(j)}}{N_{(j)}}$ estimates $\Delta t \lambda^{(k)}(t_{(j)})$.

MONOCLONAL GAMMOPATHY

- 241 Mayo Clinic Patients (Monoclonal Gammopathy of Undetermined Significance)
- 20-35 years of follow-up
- Some developed plasma cell malignancy, some died without it.
- PCM and death without PCM are competing risks

R Kyle, Benign monoclonal gammopathy – after 20 to 35 years of follow-up, Mayo Clinic Proc 1993; 68:26-36

MGUS DATA



INTERPRETATION SUBTLETY

Cannot estimate what the hazard $\lambda^{(k)}$ (or survival function $S^{(k)}(t)$) would be if competing causes of failure were removed.

Reason: $\lambda^{(k)}(t)$ = risk of type k failure at t among those still at risk at t.

$\neq$ risk of type k failure at t among those still at risk at t if other causes were removed.

For these to be equal, population at risk at T would need to have the same risk of event k whether or not other causes of failure were removed. This is a strong and unverifiable assumption (Tsiatis, 1975).

INTERPRETATION SUBTLETY

$\lambda^{(k)}(t)$ = risk of type k failure at t among those still at risk at t.

$\neq$ risk of type k failure at t among those still at risk at t if other causes were removed.

Example 1: Assumption is that risk of recurrence is the same among those who have avoided death without recurrence as it would be in the entire population if we could remove all other causes of death.

Not true if those who died from another cause before diagnosed recurrence would have been more or less likely to recur than those who survived.

INTERPRETATION SUBTLETY

$\lambda^{(k)}(t)$ = risk of type k failure at t among those still at risk at t.

$\neq$ risk of type k failure at t among those still at risk at t if other causes were removed.

Example 2: Assumption is that risk of MI is the same among those who have avoided stroke, PE, and other causes of death as it would be in the whole population if we could remove the risk of stroke, PE, and other causes of death.

Not true if those who succumb to stroke, PE or other causes of death would later have been at different risk of MI than others, had they survived CVD-free.

In general, assumption is likely not true, and it is impossible to tell from data without a means of removing the risk of the other types of failure in the population (Tsiatis, 1975; Prentice et al., 1978).

ESTIMATION

Cause-specific Hazard Functions:

- For estimation, we can treat failures from other causes as censored, and estimate the cause-specific hazard $\lambda^{(k)}(t)$ and cumulative cause-specific hazard $\Lambda^{(k)}(t)$ in the usual way.

Q: Why does this work?

A:

Q: How are failures from other causes conceptually different from the censoring we have talked about earlier in this course?

A:

CUMULATIVE FUNCTIONS

- When there are competing risks, functions that describe the distribution of the event-specific time T_k do not make sense:
  - If the subject fails of another cause before t, T_k is not defined.
- Some other cumulative functions do make sense, depending on context:
  - The probability that a subject is alive and event-of-interest-free at t.
    - This means re-defining the event of interest to be the original event of interest or death.
  - The probability that an event of type k has (or has not) occurred by time t.
    - It has not occurred if the subject dies before t.

CUMULATIVE FUNCTIONS

Event-free Survival:

Estimating the probability a subject is alive and event-of-interest-free at time t is easy:

1. Redefine the event of interest to be either the original event of interest or death

$$\delta_i = \begin{cases} 1 & \text{event of interest or death from any cause} \\ 0 & \text{censored} \end{cases}$$

$$T_i = \text{ time to event of interest, death or censoring}$$

2. Compute the KM estimate of $S(t)$ in the usual way with (T_i, δ_i) data.

EXAMPLE

Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, Shenkier T, Cella D, Davidson NE. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. New England Journal of Medicine. 2007;357(26):2666–2676.

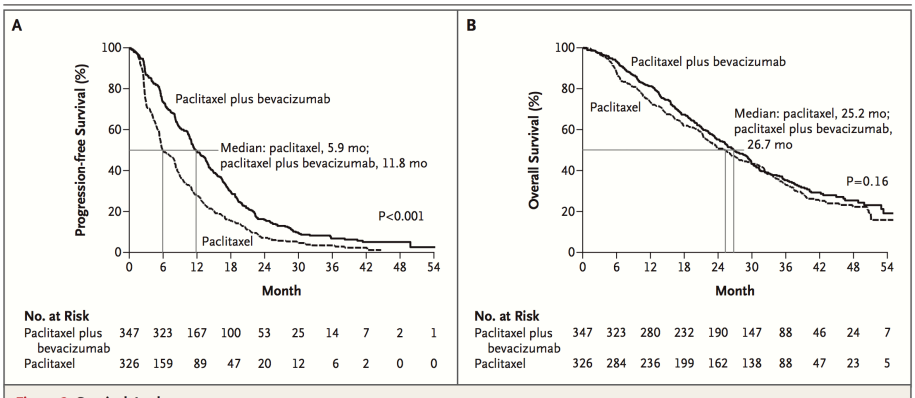
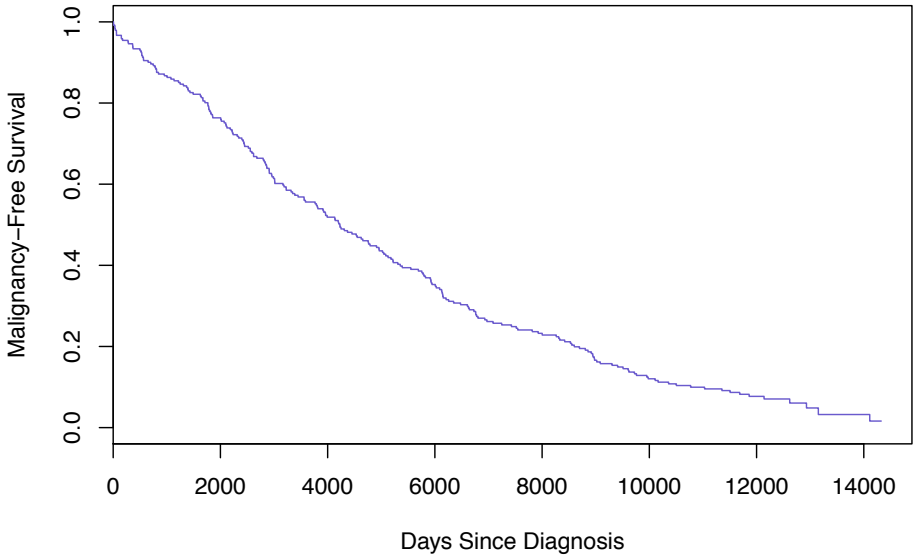


Figure 2. Survival Analyses.
Progression-free survival (Panel A) and overall survival (Panel B) in all eligible patients were analyzed with the use of the Kaplan–Meier method. Analyses including all patients assigned to treatment yielded similar results (data not shown).

MALIGNANCY-FREE SURVIVAL



CUMULATIVE FUNCTIONS

Cumulative Incidence:

The probability that an event of type k has occurred by time t.

- Makes sense without requiring that a time to the k^{th} type of event be defined for all subjects.
- Depends on who is at risk in the population at each time, so it will depend not only on the cause-specific hazard of the event of interest, but also the cause-specific hazards of all the other causes of failure.
- Is a population-specific quantity that depends on what other risks are operating in the population, and how they are related to the risk of the event of interest.

CUMULATIVE FUNCTIONS

- Define cause-specific cumulative incidence at time t as the event:

  [Failure due to cause k occurred by t] :

  Then the cause-specific cumulative incidence function is defined as

  $$F^{(k)}(t) = \Pr[\text{Failure due to cause } k \text{ occurred by } t]$$

- Could also look at $1 - F^{(k)}(t)$, but usually interest centers on $F^{(k)}(t)$.

ESTIMATION

Cumulative Incidence:

- We can treat failures from other causes as censored, and estimate the cause-specific hazard $\lambda^{(k)}(t)$ and cumulative cause-specific hazard $\Lambda^{(k)}(t)$ in the usual way.

- If we do the same thing for the cumulative incidence, the Kaplan-Meier $1 - \hat{S}^{(k)}(t)$ is a biased estimate of $F^{(k)}(t)$.

Q: Why?

A:

CUMULATIVE FUNCTIONS

In a study following a cohort of subjects with high cholesterol levels, for MI or death:

Q: What are some circumstances where you would be more interested in the cumulative incidence function than the event-free survival function?

A:

Q: What are some circumstances where you would be more interested in the event-free survival function than the cumulative incidence function?

A:

ESTIMATING CUMULATIVE INCIDENCE

- We can write

$$1 - \hat{S}^{(k)}(t) = \sum_{j:t_{(j)} \le t} \frac{D_{(j)}^{(k)}}{N_{(j)}} \hat{S}^{(k)}(t_{(j-1)})$$

- At the second failure time of type k,

$$1 - \hat{S}^{(k)}(t_{(2)}) = 1 - \frac{N_{(1)} - D_{(1)}^{(k)}}{N_{(1)}} \cdot \frac{N_{(2)} - D_{(2)}^{(k)}}{N_{(2)}} = \frac{D_{(1)}^{(k)}}{N_{(1)}} + \frac{D_{(2)}^{(k)}}{N_{(2)}} \cdot \frac{N_{(1)} - D_{(1)}^{(k)}}{N_{(1)}}$$

- If any failures of another type have occurred between $t_{(1)}$ and $t_{(2)}$, *the* $\frac{N_{(1)} - D_{(1)}^{(k)}}{N_{(1)}}$ term is too big.
- This bias will accumulate and get larger, as we move to larger and larger $t_{(j)}$.

ESTIMATING CUMULATIVE INCIDENCE

- Letting $D_{(j)}^{(\bar{k})}$ = the number of failures of types other than k at $t_{(j)}$, an unbiased estimate of $F^{(k)}(t)$ is given by

$$\sum_{j:t_{(j)}\le t} \frac{D_{(j)}^{(k)}}{N_{(j)}} \prod_{i=1}^{j-1} \frac{N_{(i)} - D_{(i)}^{(k)} - D_{(i)}^{(\bar{k})}}{N_{(i)}} = \sum_{j:t_{(j)}\le t} \frac{D_{(j)}^{(k)}}{N_{(j)}} \prod_{i=1}^{j-1} \frac{N_{(i)} - D_{(i)}^{(k)}}{N_{(i)}} \cdot \frac{N_{(i)} - D_{(i)}^{(\bar{k})}}{N_{(i)}}$$

↑

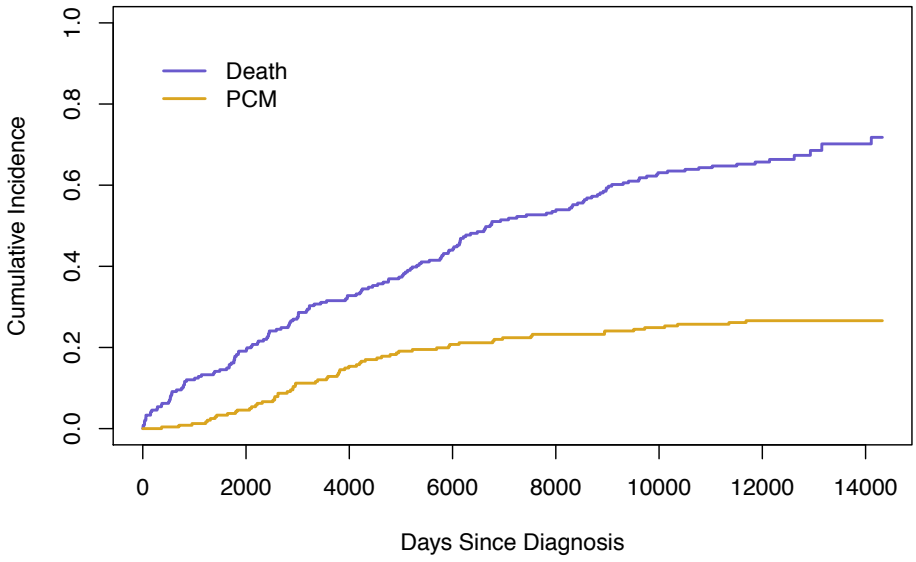
no ties between failures of different types

- Compare to biased upward

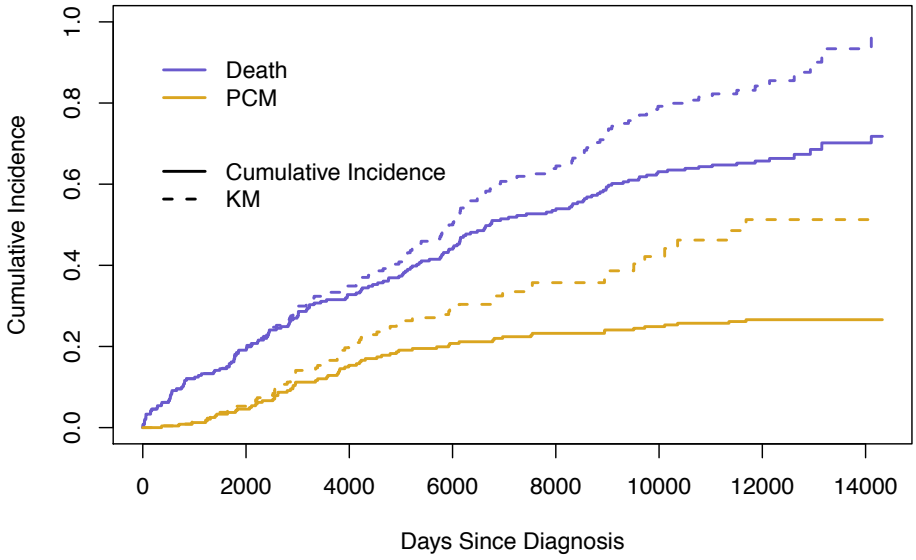
$$1 - \hat{S}^{(k)}(t) = \sum_{j:t_{(j)}\le t} \frac{D_{(j)}^{(k)}}{N_{(j)}} \hat{S}^{(k)}(t_{(j-1)}) = \sum_{j:t_{(j)}\le t} = \frac{D_{(j)}^{(k)}}{N_{(j)}} \prod_{i=1}^{j-1} \frac{N_{(i)} - D_{(i)}^{(k)}}{N_{(i)}}$$

.

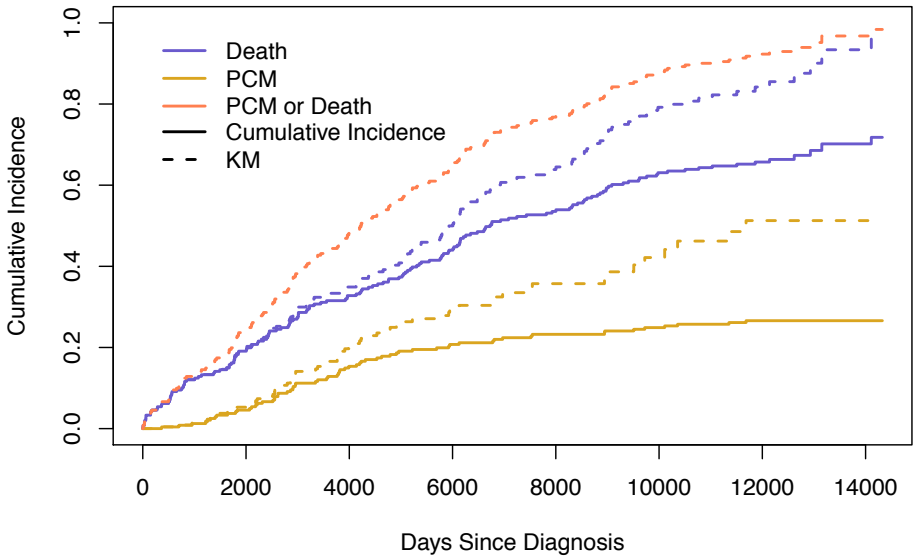
CUMULATIVE INCIDENCE



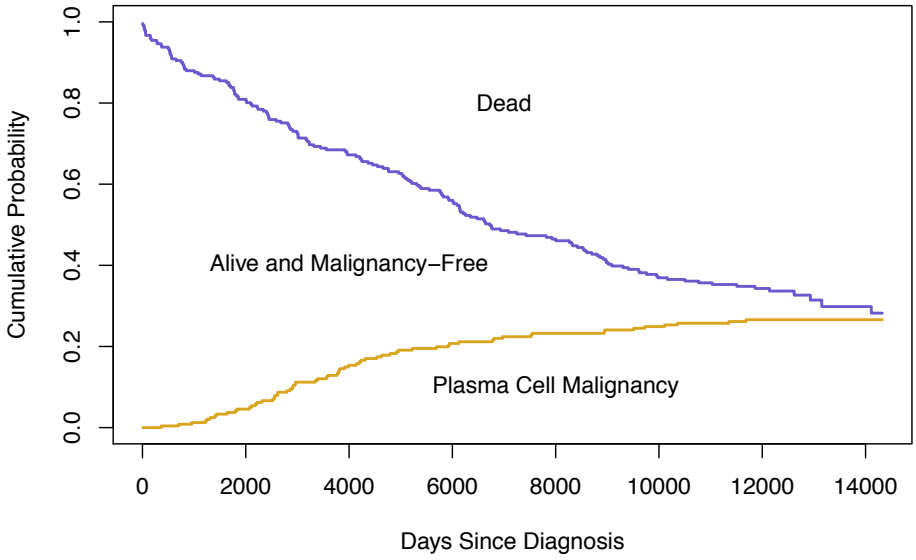
CUMULATIVE INCIDENCE



CUMULATIVE INCIDENCE



TOGETHER



CHOICE OF OUTCOME EVENT

- May not want cause-specific event as primary outcome.
- Interpretation cloudy, particularly for survival curves.
- More on this in Module 12

TO WATCH OUT FOR

- Interpretation in the presence of competing risks can be subtle and requires care.
  - S(t) defined in terms of the probability distribution of T_k does not make sense
  - Cannot interpret functions of the cause-specific hazard as applying in a population without competing risks present.
  - 1 – KM estimator can give upward biased estimate of cumulative incidence.
  - Cannot interpret cumulative incidence as applying in a population without competing risks present

In R

Load packages.

```
library(survival)
library(ggplot2)
library(muhaz)
```

Get data.

```
df <- flchain[flchain$futime > 7,]
Y <- with(df, Surv(futime, death))
```

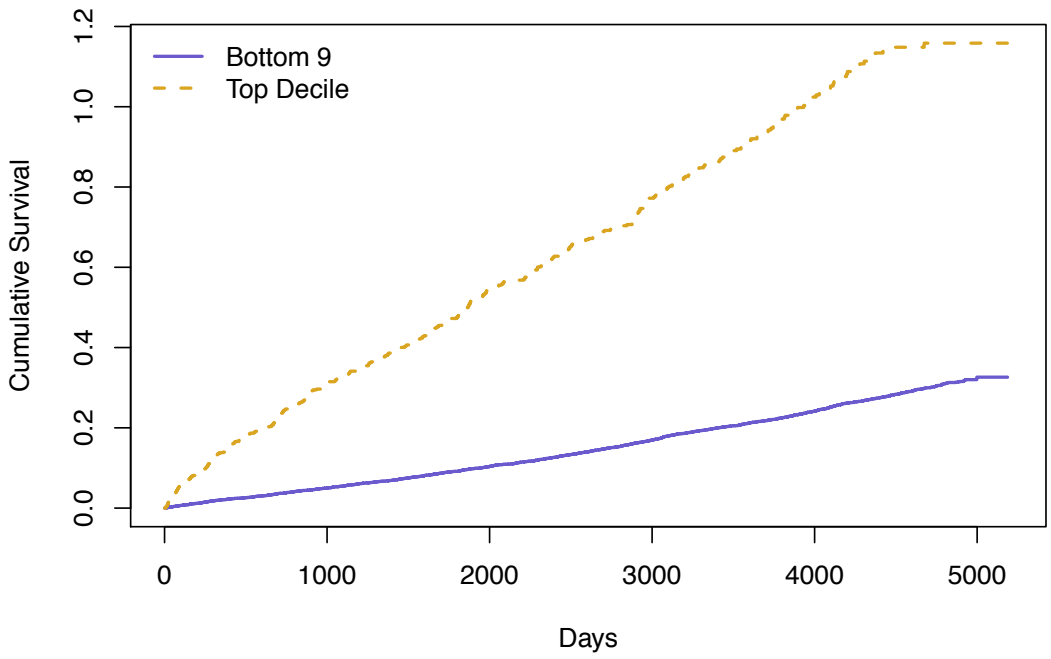
Make binary exposure

```
df$topdecile <- df$flc.grp > 9
```

Cumulative Hazard

```
colors <- c("slateblue", "goldenrod")
plot(survfit(Y ~ topdecile, data = df), xlab = "Days",
       ylab = "Cumulative Survival", lty = 1:2,
     col = colors, lwd = 2, fun = "cumhaz")
legend("topleft", legend = c("Bottom 9","Top Decile" ),
       lty = 1:3, col = colors, lwd = 2, bty = "n")
```

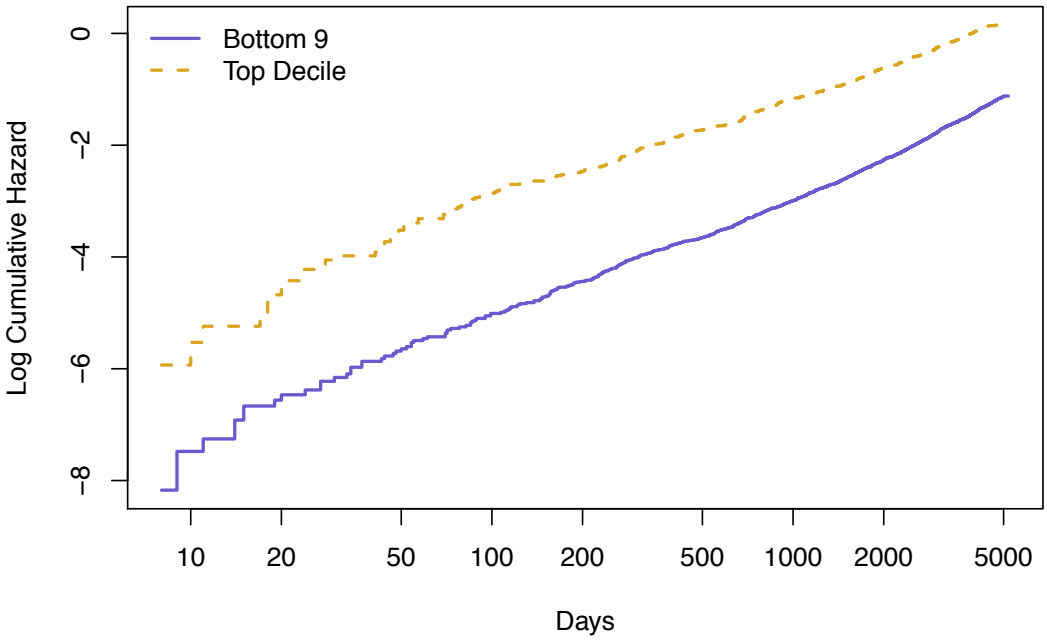
Cumulative Hazard



Log Cumulative Hazard

```
colors <- c("slateblue", "goldenrod")
plot(survfit(Y ~ topdecile, data = df), xlab = "Days",
       ylab = "Log Cumulative Hazard", lty = 1:2,
     col = colors, lwd = 2, fun = "cloglog")
legend("topleft", legend = c("Bottom 9","Top Decile" ),
       lty = 1:3, col = colors, lwd = 2, bty = "n")
```

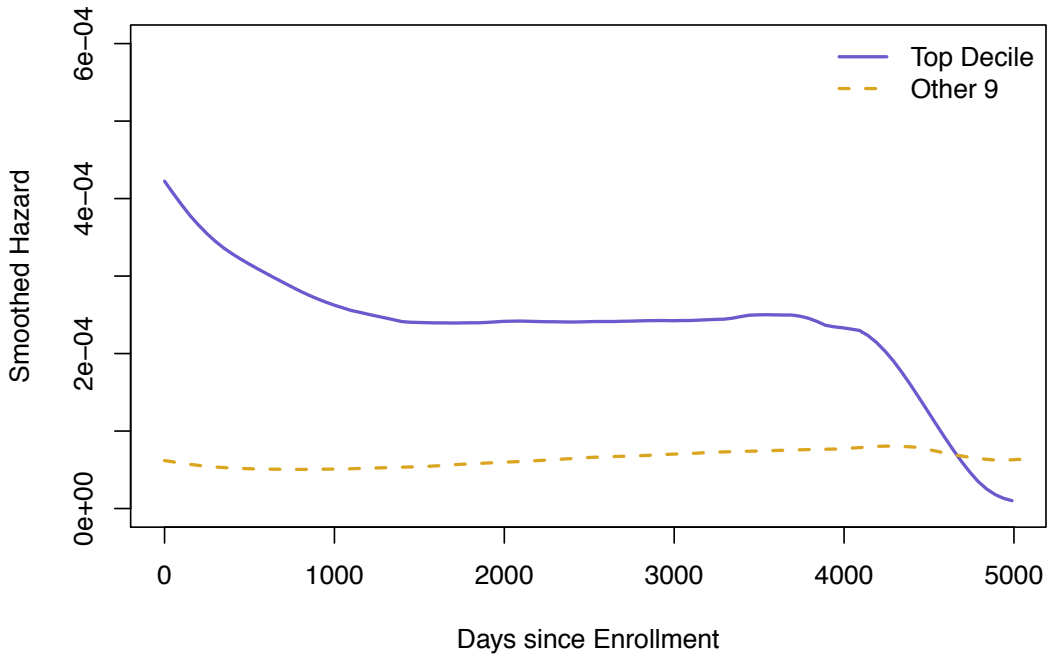
Log Cumulative Hazard



Kernel Hazard

```
ld1 <- with(df, muhaz(futime, death, topdecile))
ld2 <- with(df, muhaz(futime, death, !topdecile))
plot(ld1, type = "l", col = colors[1], lwd = 2,
     xlab = "Days since Enrollment", ylab = "Smoothed Hazard",
     ylim = c(0, 6e-04))
lines(ld2, lwd = 2, col = colors[2], lty = 2)
legend("topright", legend = c("Top Decile", "Other 9"),
       lty= c(1:2), col = colors, bty = "n", lwd = 2)
```

Kernel Hazard



Monoclonal Gammopathy Data

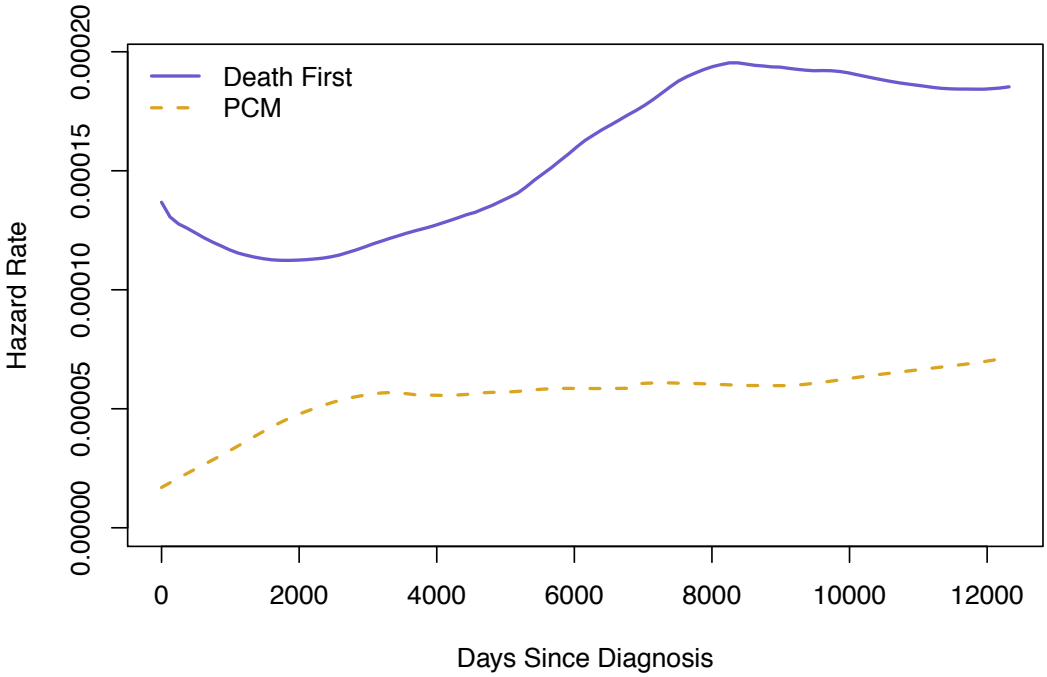
- 241 Mayo Clinic Patients (Monoclonal Gammopathy of Undetermined Significance)
- 20-35 years of follow-up
- Some developed plasma cell malignancy, some died without it.
- PCM and death without PCM are competing risks

R Kyle, Benign monoclonal gammopathy – after 20 to 35 years of follow-up, Mayo Clinic Proc 1993; 68:26-36

Cause-specific Hazards

```
diefirst <- mgus$death
diefirst[!is.na(mgus$pctime)] <- 0
pcm <- !is.na(mgus$pctime)
t <- mgus$futime
t[!is.na(mgus$pctime)] <- mgus$pctime[!is.na(mgus$pctime)]
deathhaz <- muhaz(t, diefirst)
pcmhaz <- muhaz(t,pcm)
colors <- c("slateblue", "goldenrod")
plot(deathhaz, lwd = 2, col = colors[1], xlab = "Days Since Diagnosis")
lines(pcmhaz, lwd = 2, col = colors[2], lty  = 2)
legend("topleft", col = colors, legend = c("Death First", "PCM"),
       lty = c(1,2), lwd = 2, bty = "n")
```

Cause-specific Hazards



Survival Curves Based on Cause-specific Hazards

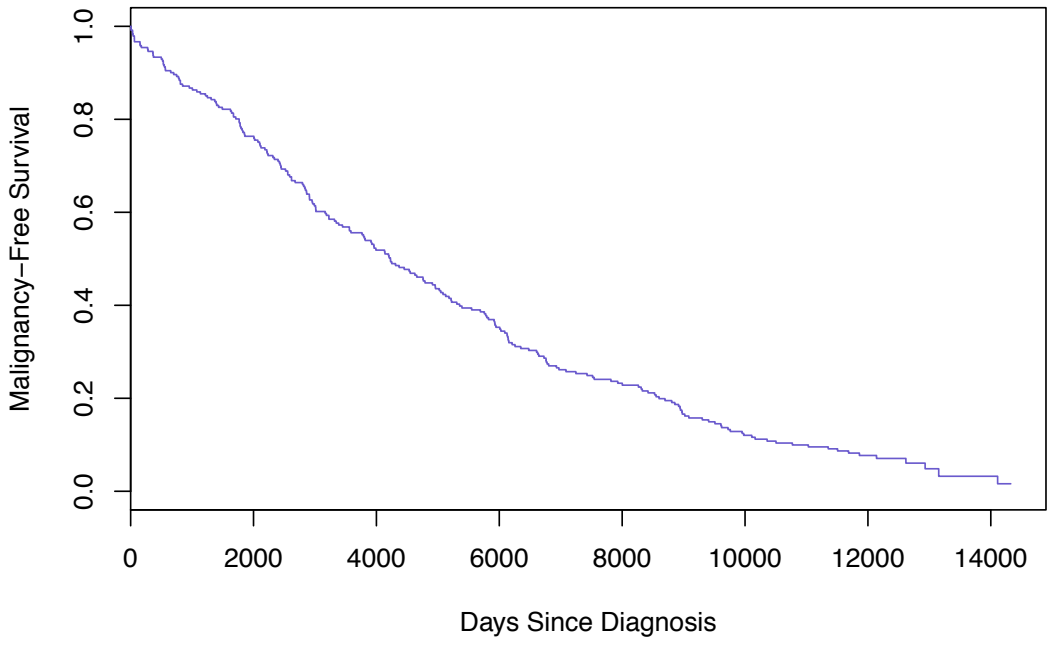
```
Yd <- Surv(t, diefirst)
fitd <- survfit(Yd ~ 1)

Ypcm <- Surv(t, pcm)
fitpcm <- survfit(Ypcm ~ 1)
```

Malignancy-Free survival

```
del <- with(mgus, pmax(diefirst, pcm))
Yany <- Surv(t, del)
anyfit <- survfit(Yany ~ 1)
plot(anyfit, col = colors, conf.int = FALSE,
     xlab = "Days Since Diagnosis", ylab = "Malignancy-Free Survival")
```

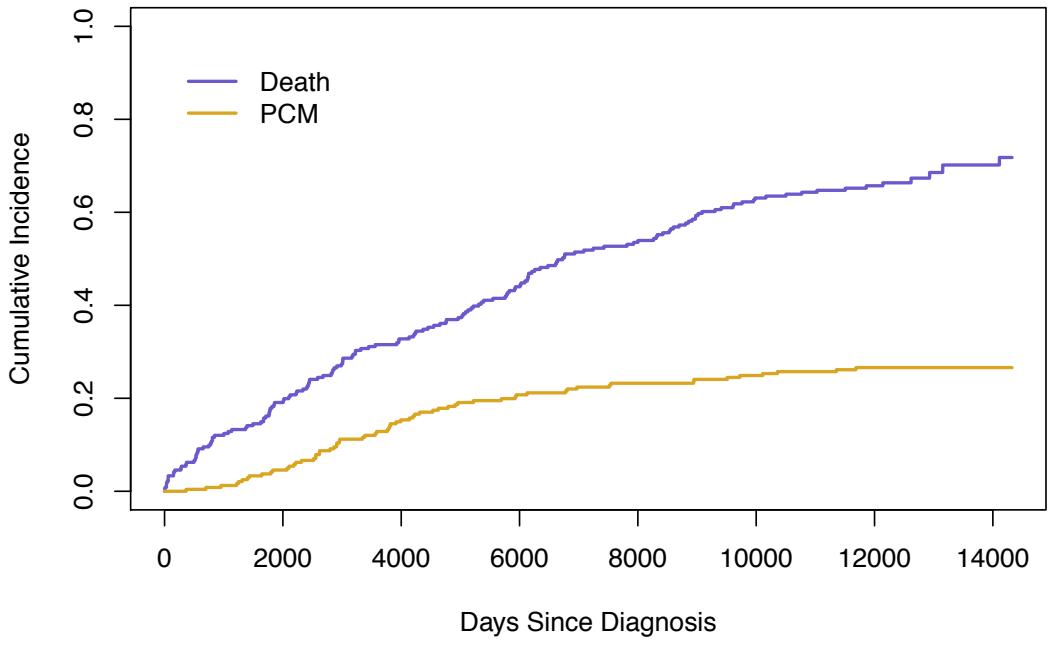
Malignancy-Free survival



Cumulative Incidence

```
etype <- mgus$death
etype[!is.na(mgus$pctime)] <- 2
Ym <- Surv(t, etype, type = "mstate")
cuminc <- survfit(Ym ~ 1)
plot(cuminc, fun = "event", lwd = 2, col = colors, ylim = c(0, 1),
     xlab = "Days Since Diagnosis", ylab = "Cumulative Incidence")
legend(0,.95, legend = c("Death", "PCM"), col = colors, bty = "n",
       lwd = 2)
```

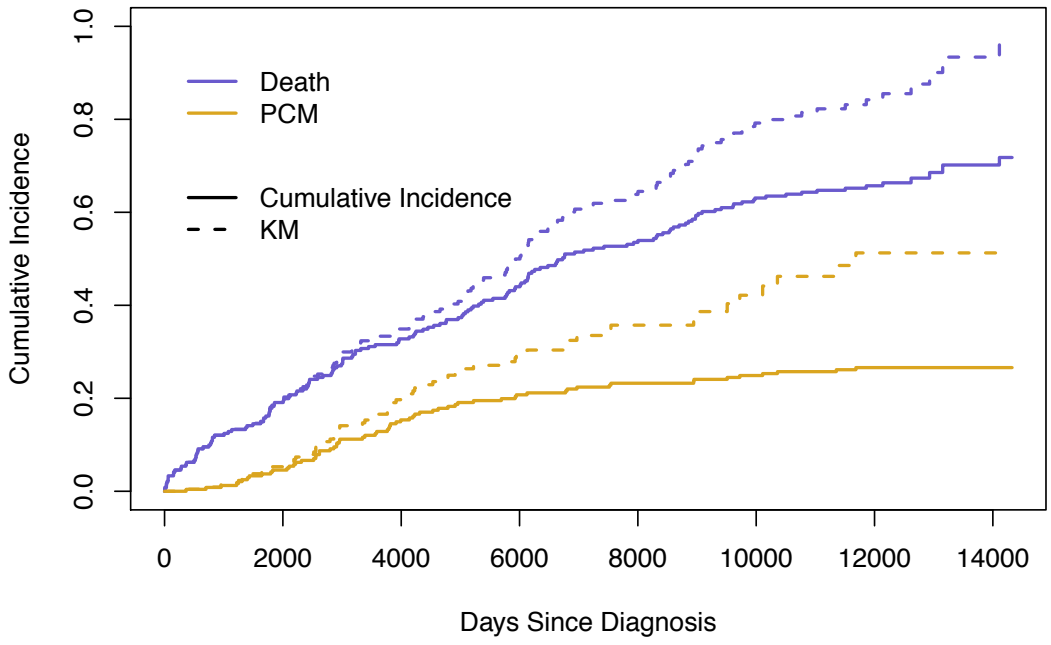
Cumulative Incidence



Compare Incidence Functions

```
plot(cuminc, fun = "event", lwd = 2, col = colors, ylim = c(0, 1),
     xlab = "Days Since Diagnosis", ylab = "Cumulative Incidence")
legend(0, .95, legend =  c("Death", "PCM"), col = colors, bty = "n",
       lwd = 2)
lines(fitd, fun = "event", conf.int = FALSE, lwd = 2, col = colors[1],
      lty = 2)
lines(fitpcm,fun = "event", conf.int = FALSE, lwd = 2, col = colors[2],
      lty = 2 )
legend(0, .7, legend = c("Cumulative Incidence", "KM"), lty = c(1:2),
       lwd = 2, bty = "n")
```

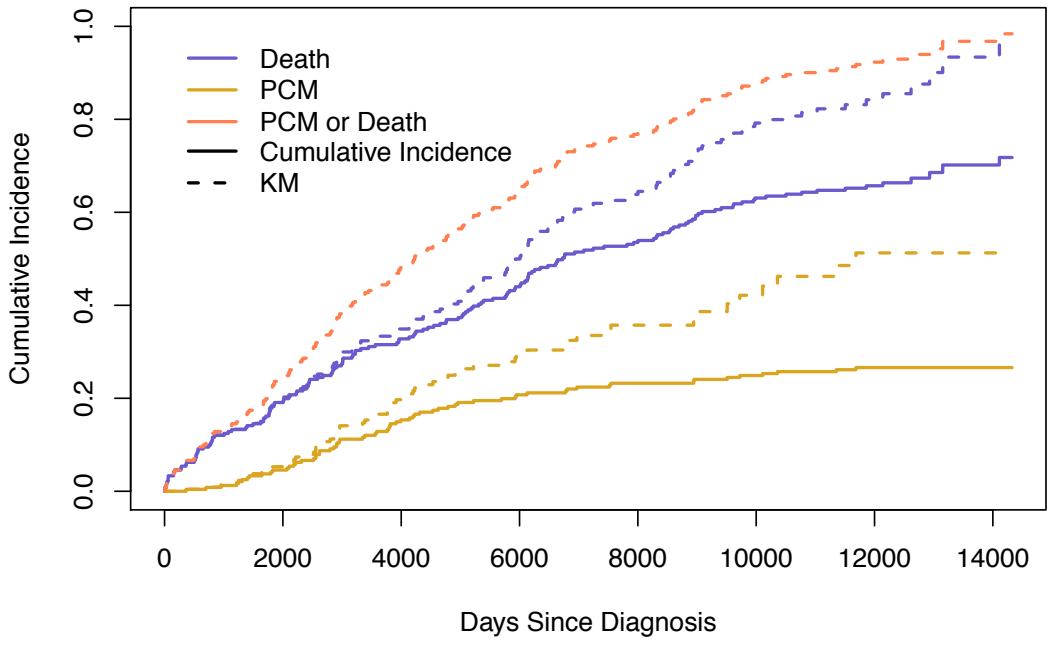
Compare Incidence Functions



Add Malignancy-free Survival Incidence

```
colors <- c("slateblue", "goldenrod", "coral")
plot(cuminc, fun = "event", lwd = 2, col = colors, ylim = c(0, 1),
     xlab = "Days Since Diagnosis", ylab = "Cumulative Incidence")
legend(0, 1, legend  = c("Death", "PCM", "PCM or Death"), col = colors,
       bty = "n", lwd = 2)
lines(fitd, fun = "event", conf.int = FALSE, lwd = 2, col = colors[1],
      lty = 2)
lines(fitpcm,fun = "event", conf.int = FALSE, lwd = 2, col = colors[2],
      lty = 2 )
legend(0, .8, legend = c("Cumulative Incidence", "KM"), lty = c(1:2),
       lwd = 2, bty = "n")
lines(anyfit, fun = "event", conf.int = FALSE, col = "coral", lwd = 2,
      lty = 2)
```

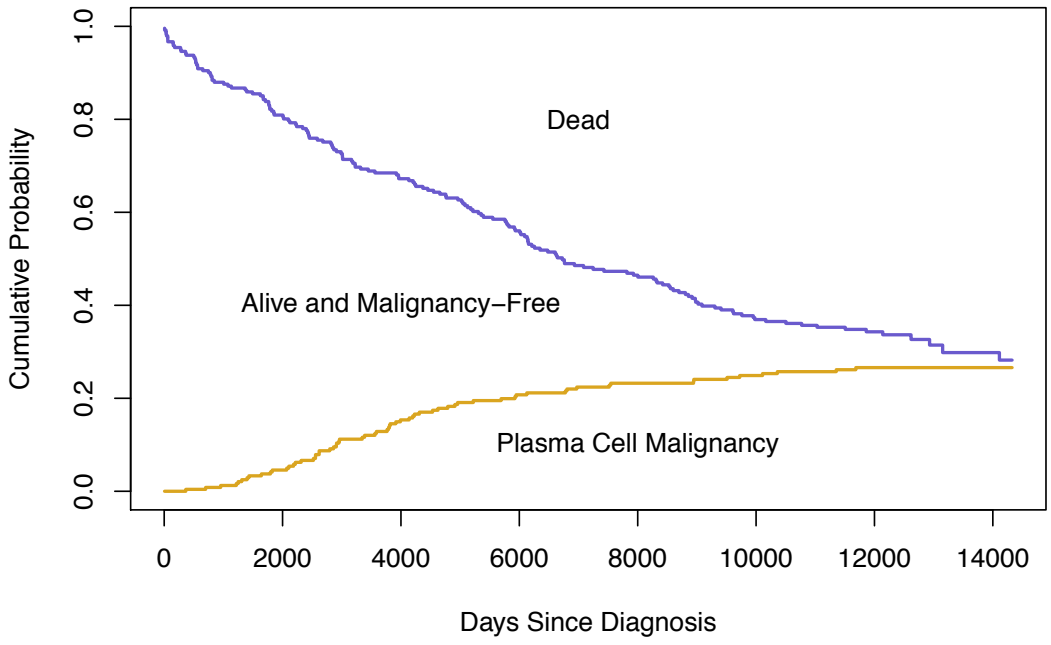
Add Malignancy-free Survival Incidence



PCM or Death

```
plot(cuminc$time, 1 - cuminc$prev[,1], type = "s", lwd = 2,
     col = colors, ylim = c(0,1), ylab = "Cumulative Probability",
     xlab = "Days Since Diagnosis")
lines(cuminc$time, cuminc$prev[,2], type = "s", lwd = 2, col = colors[2])
text(7000, .8, "Dead"); text(8000, .1, "Plasma Cell Malignancy")
text(4000, .4, "Alive and Malignancy-Free")
```

PCM or Death



Your turn

Using the monoclonal gammopathy data in mgus:

1. Compute and plot cause-specific hazard functions separately for those for whom the monoclonal protein spike at diagnosis (mspike) is greater than 2 and those for whom it is less than or equal to 2.
2. Fit Cox models relating the cause specific hazard functions for PCM and death without prior PCM to diagnosis mspike > 2.

SESSION 3: CHOICE OF THE TIME SCALE AND INTERACTIONS WITH TIME

Module 8: Survival Analysis for Observational Data

Summer Institute in Statistics for Clinical Research
University of Washington
July, 2016

Barbara McKnight, Ph.D.

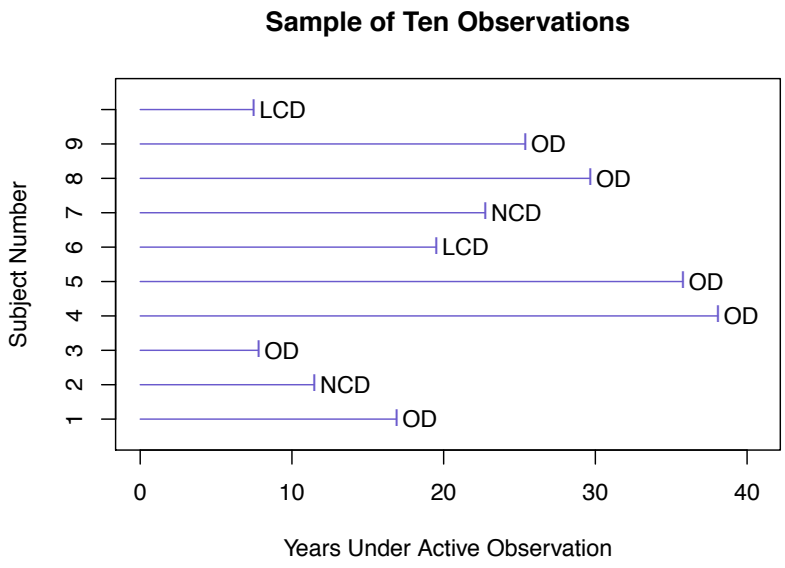
OUTLINE

- Choice of the time scale for analysis
- Left entry into observation (left truncation)
- Cox models including interaction with time variables
- Cox models with time-dependent coefficients

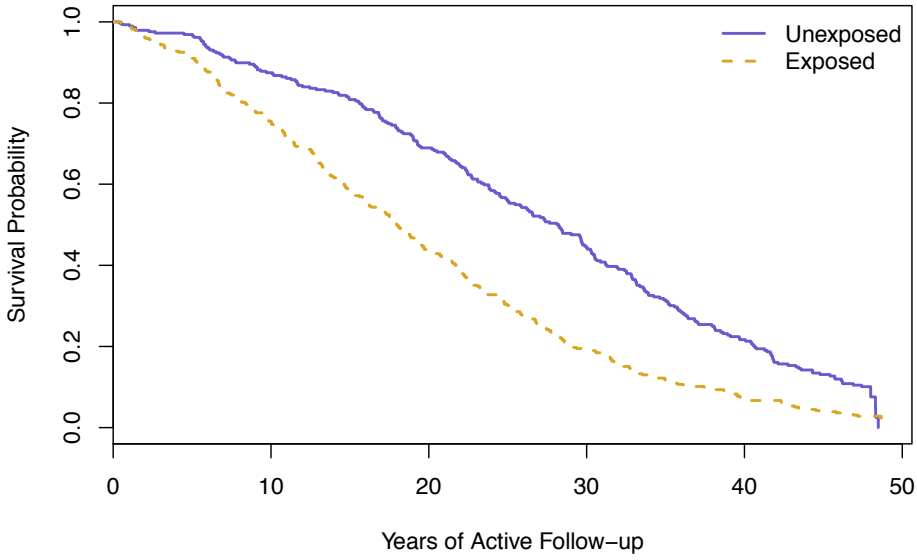
WELSH NICKEL REFINERS STUDY

- 679 nickel refinery workers identified twice on paysheets April 1929, 1934, 1939, 1944, 1949
- Follow-up until 1981
- Refinery cleaned up by various means 1922-1932, so all important exposure occurred before beginning of follow-up
- Interest in whether duration of employment in high-exposure areas, and age at first exposure, were related to lung and nasal sinus cancer mortality risk.

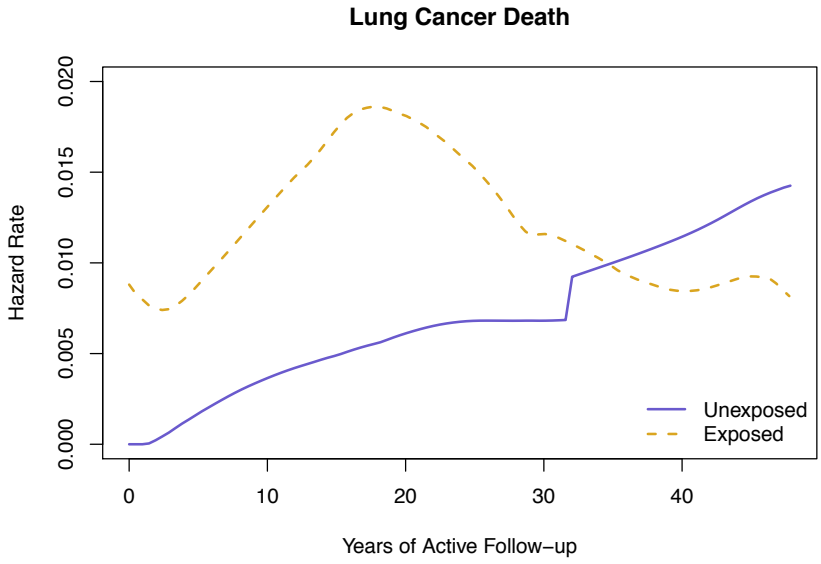
WELSH NICKEL REFINERS



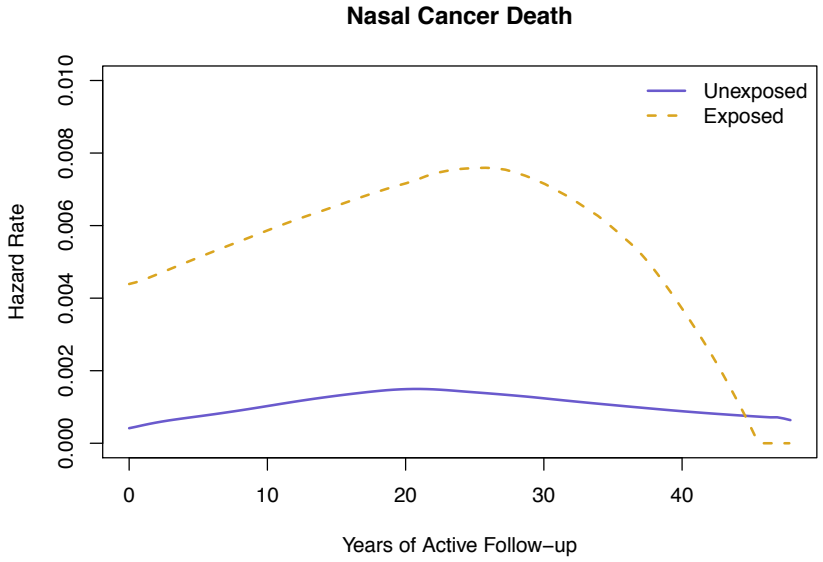
ALL-CAUSE MORTALITY



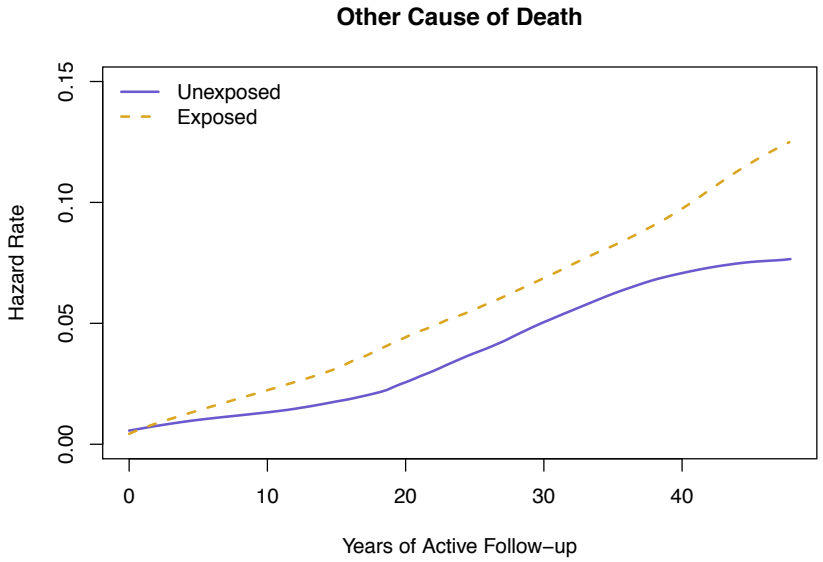
WELSH NICKEL REFINERS



WELSH NICKEL REFINERS



WELSH NICKEL REFINERS



LUNG CANCER FU TIME

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	0.9200182	2.509336	0.1869493	4.921217	9e-07

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.6030012	1.8275955	0.2121299	2.8426041	0.0044747
exp4.5 - 8.0	1.0862839	2.9632419	0.2828485	3.8405146	0.0001228
exp8.5-12.0	1.2772969	3.5869307	0.3742268	3.4131628	0.0006421
exp12.5+	1.4873597	4.4253955	0.4798472	3.0996524	0.0019375
afe20-27.5	0.8103938	2.2487934	0.3079688	2.6314149	0.0085030
afe27.5 - 35	0.9149895	2.4967489	0.3291081	2.7802097	0.0054324
afe35+	0.8068991	2.2409482	0.4237839	1.9040342	0.0569057
yfe1910-1914	0.3342204	1.3968510	0.2695145	1.2400835	0.2149445
yfe1915-1919	-0.1340505	0.8745459	0.3749097	-0.3575540	0.7206771
yfe1920-1925	0.0744977	1.0773429	0.2966621	0.2511197	0.8017216

NASAL CANCER FU TIME

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	1.614074	5.023236	0.3516507	4.589994	4.4e-06

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.8356274	2.3062606	0.4032111	2.072432	0.0382252
exp4.5 - 8.0	1.1366437	3.1162916	0.4706657	2.414970	0.0157365
exp8.5-12.0	2.2945326	9.9197981	0.5117936	4.483316	0.0000073
exp12.5+	2.8713357	17.6605917	0.5697217	5.039892	0.0000005
afe20-27.5	1.4686105	4.3431963	0.7518514	1.953326	0.0507810
afe27.5 - 35	2.1598639	8.6699580	0.7588726	2.846148	0.0044252
afe35+	3.4767227	32.3535148	0.7843101	4.432842	0.0000093
yfe1910-1914	0.7130093	2.0401213	0.3728470	1.912338	0.0558329
yfe1915-1919	0.5040978	1.6554913	0.5034466	1.001294	0.3166849
yfe1920-1925	-0.9304088	0.3943924	0.5152666	-1.805684	0.0709677

OTHER CAUSES FU TIME

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	0.3962896	1.4863	0.0972056	4.076818	4.57e-05

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.1318081	1.1408894	0.1105672	1.1921083	0.2332188
exp4.5 - 8.0	0.1308735	1.1398236	0.1603797	0.8160231	0.4144869
exp8.5-12.0	0.0324914	1.0330250	0.2563862	0.1267282	0.8991555
exp12.5+	-0.0774111	0.9255093	0.3964677	-0.1952520	0.8451957
afe20-27.5	0.5275548	1.6947832	0.1539622	3.4265217	0.0006114
afe27.5 - 35	1.1070376	3.0253827	0.1653356	6.6956992	0.0000000
afe35+	1.9740626	7.1998671	0.1942464	10.1626701	0.0000000
yfe1910-1914	-0.2148112	0.8066937	0.1515491	-1.4174361	0.1563555
yfe1915-1919	-0.5297679	0.5887416	0.1766843	-2.9983870	0.0027141
yfe1920-1925	-1.1456390	0.3180206	0.1502442	-7.6251795	0.0000000

COX REGRESSION MODEL

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

Interpretation of e^{β_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".

$\lambda(t)$ for $x_1 + 1$: $\lambda_0(t)e^{\beta_1(x_1+1)+\cdots+\beta_k x_k}$

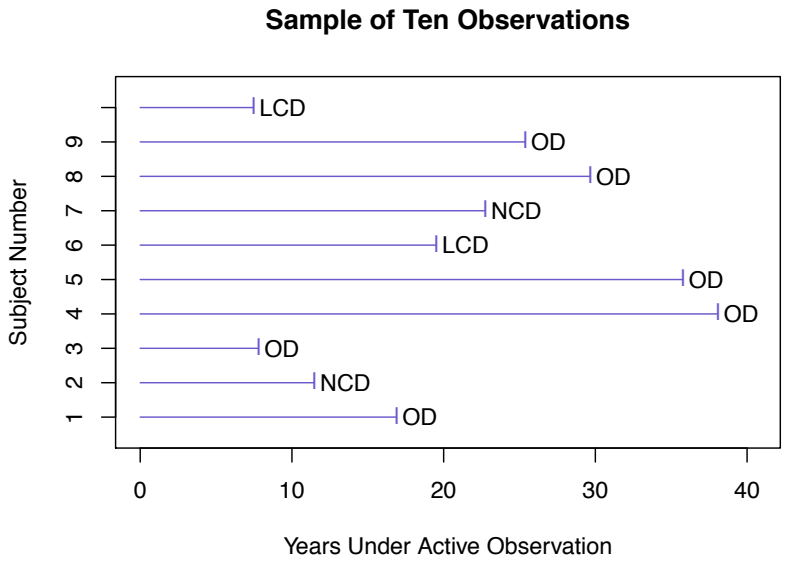
$\lambda(t)$ for x_1: $\lambda_0(t)e^{\beta_1 x_1+\cdots+\beta_k x_k}$

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

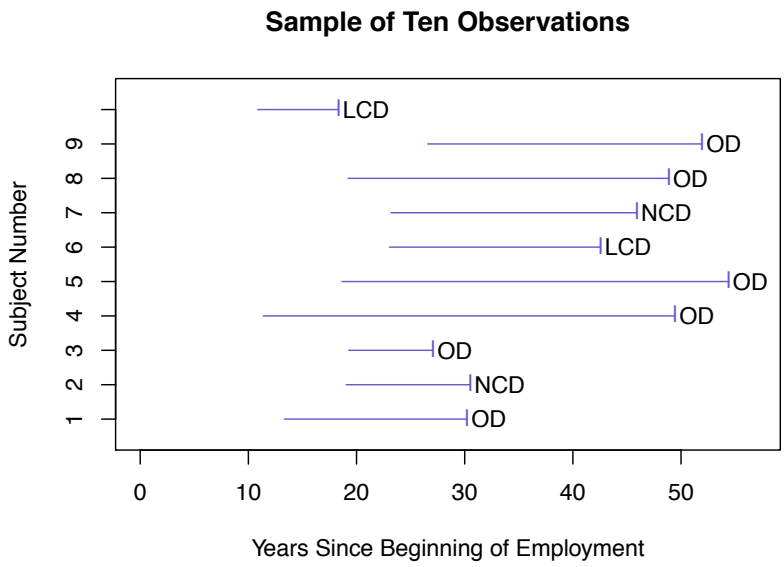
COX REGRESSION MODEL

- $e^{\beta 1}$ is the RR associated with a one-unit difference of $x_{1,}$ holding other x's and t constant.
- Some functional form is required for how the hazard function at each t depends on $x_2 \ldots x_k$.
- No functional form is required for how the hazard at each $x_2 \ldots x_k$ depends on t, since $\lambda_0(t)$ can be any function.
- The time scale for t is the variable that is adjusted for the most finely/thoroughly.

WELSH NICKEL REFINERS



WELSH NICKEL REFINERS



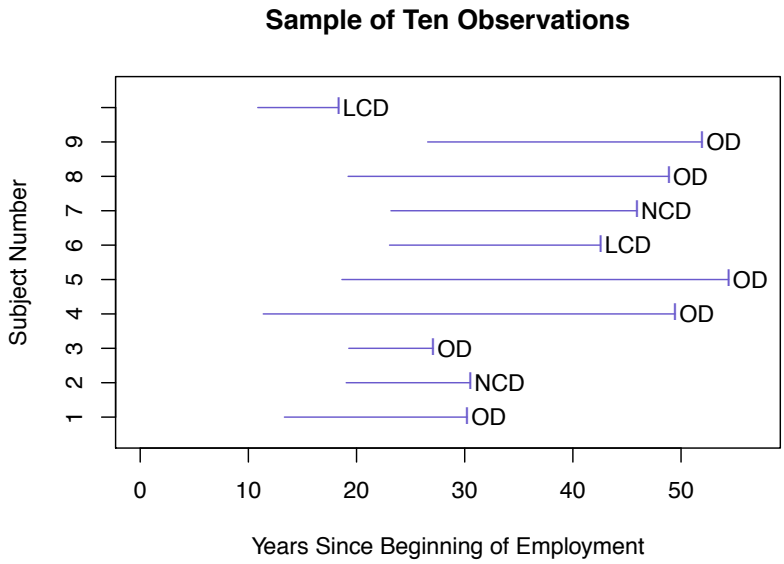
OBSERVATION STARTING LATE

- Should not include subjects in risk sets before they are under observation:
  - Other subjects "just like" them who died before their entry time are not observed
  - Falsely inflates the numbers at risk in early risk sets
  - Biases cause-specific hazard estimation
  - Can bias Cox model estimation

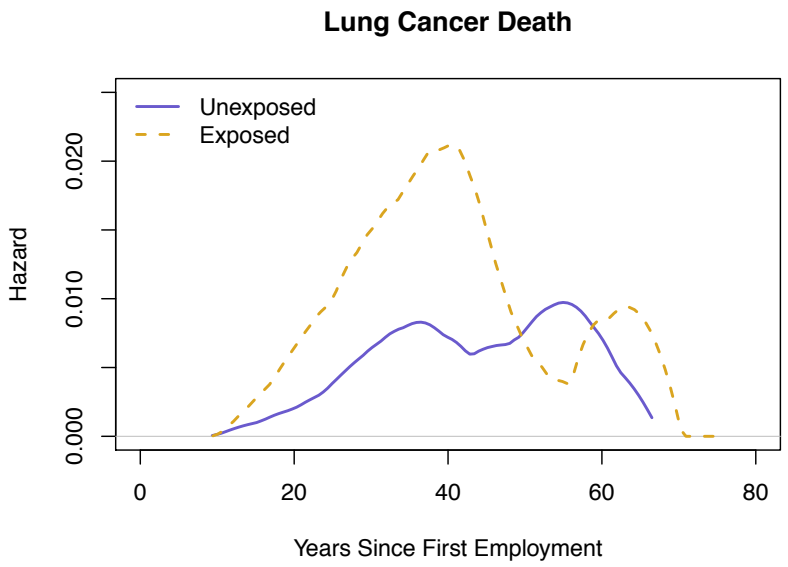
OBSERVATION STARTING LATE

- Solution: "Left enter" subjects at time when active follow-up starts
  - Subjects only contribute to risk sets where their event could have been observed
  - They are only in the denominator if we could have seen them in the numerator

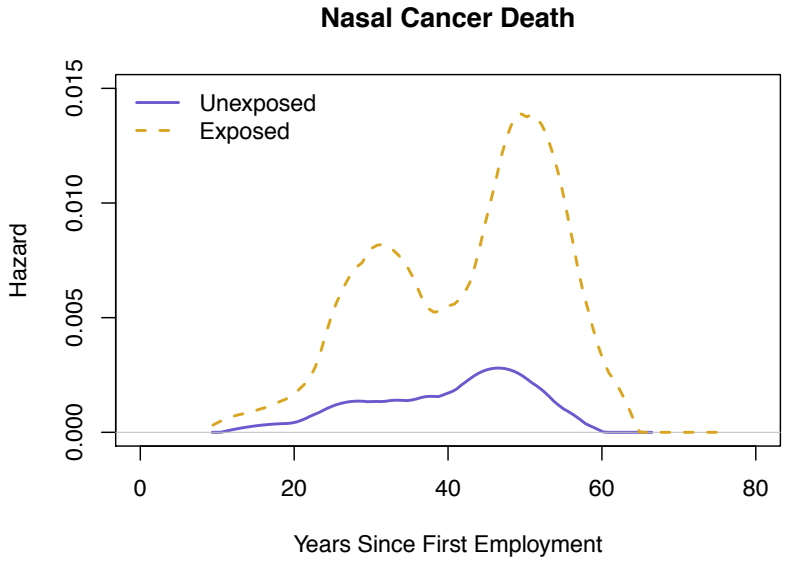
WELSH NICKEL REFINERS



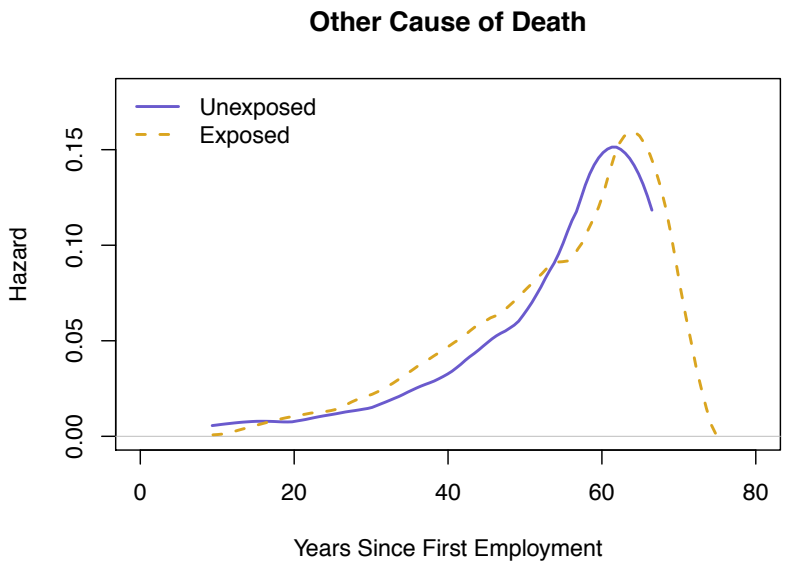
LUNG CANCER



NASAL CANCER



OTHER CAUSES



LUNG CANCER

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	0.8000334	2.225615	0.1860041	4.301159	1.7e-05

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.6111674	1.842581	0.2123734	2.877796	0.0040046
exp4.5 - 8.0	1.0952795	2.990018	0.2838639	3.858467	0.0001141
exp8.5-12.0	1.2880174	3.625591	0.3739070	3.444754	0.0005716
exp12.5+	1.4327121	4.190048	0.4791166	2.990321	0.0027868
afe20-27.5	0.7604881	2.139320	0.3081636	2.467806	0.0135944
afe27.5 - 35	0.8670846	2.379962	0.3281099	2.642665	0.0082256
afe35+	0.7982183	2.221579	0.4224336	1.889571	0.0588154
yfe1910-1914	0.4358460	1.546271	0.2724801	1.599552	0.1096981
yfe1915-1919	0.1753274	1.191636	0.3775109	0.464430	0.6423397
yfe1920-1925	0.6547157	1.924595	0.2991155	2.188839	0.0286086

NASAL CANCER

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	1.540408	4.666495	0.3503185	4.397165	1.1e-05

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.8958359	2.449382	0.4044464	2.2149680	0.0267623
exp4.5 - 8.0	1.1991717	3.317368	0.4727052	2.5368277	0.0111862
exp8.5-12.0	2.3214816	10.190761	0.5173928	4.4868842	0.0000072
exp12.5+	2.8655920	17.559445	0.5727364	5.0033346	0.0000006
afe20-27.5	1.4721869	4.358757	0.7527320	1.9557917	0.0504897
afe27.5 - 35	2.1770312	8.820082	0.7601145	2.8640834	0.0041822
afe35+	3.6025888	36.693104	0.7886401	4.5681026	0.0000049
yfe1910-1914	1.0373701	2.821786	0.3798834	2.7307593	0.0063189
yfe1915-1919	1.1291520	3.093033	0.5130845	2.2007137	0.0277563
yfe1920-1925	0.0166965	1.016837	0.5257787	0.0317558	0.9746668

OTHER CAUSE OF DEATH

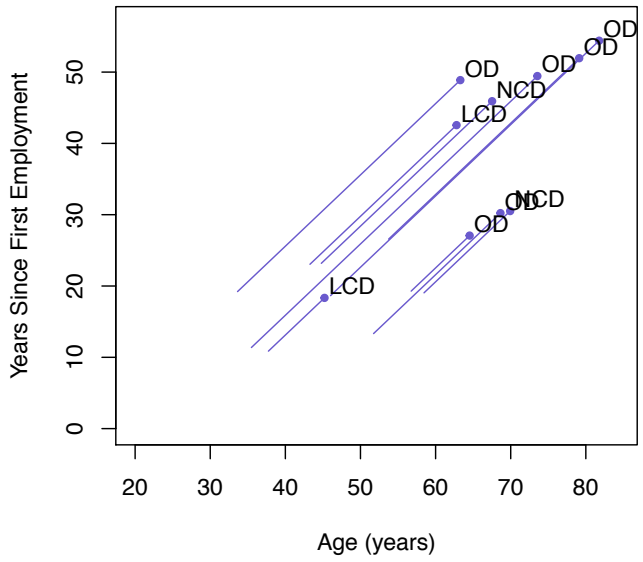
	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	0.2164895	1.24171	0.0966131	2.240788	0.0250398

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exp0.5 - 4.0	0.1685250	1.183558	0.1106070	1.5236376	0.1275993
exp4.5 - 8.0	0.2360561	1.266245	0.1602288	1.4732445	0.1406851
exp8.5-12.0	0.0585201	1.060266	0.2564181	0.2282213	0.8194742
exp12.5+	0.0245456	1.024849	0.3964995	0.0619059	0.9506378
afe20-27.5	0.5704774	1.769111	0.1545876	3.6903186	0.0002240
afe27.5 - 35	1.1656136	3.207891	0.1665088	7.0003136	0.0000000
afe35+	2.0835886	8.033245	0.1957375	10.6448086	0.0000000
yfe1910-1914	0.2087081	1.232085	0.1540413	1.3548842	0.1754544
yfe1915-1919	0.2329453	1.262312	0.1788233	1.3026563	0.1926921
yfe1920-1925	0.1024386	1.107869	0.1529133	0.6699127	0.5029135

CHOOSING A TIME SCALE

- What time scale makes the most sense for the Welsh Nickel Refiners study?

TWO TIME SCALES



CHOOSING A TIME SCALE

- Cardiovascular Health Study
  - NHLBI cohort of older Americans (65+)
  - Many baseline demographic and health measures.
  - Follow-up for more than 20 years for a large number of health conditions.
- What is the best time scale: age or time since baseline?

TIME INTERACTIONS

- So far, all our Cox models have assumed that the hazard ratio is constant over time
- It's possible to incorporate interaction terms with functions of time to allow the HR to depend on time.

TIME INTERACTIONS

One way for the hazard ratio to depend on time: interaction with a function of time t.

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x + \beta_2 f(t)}$$

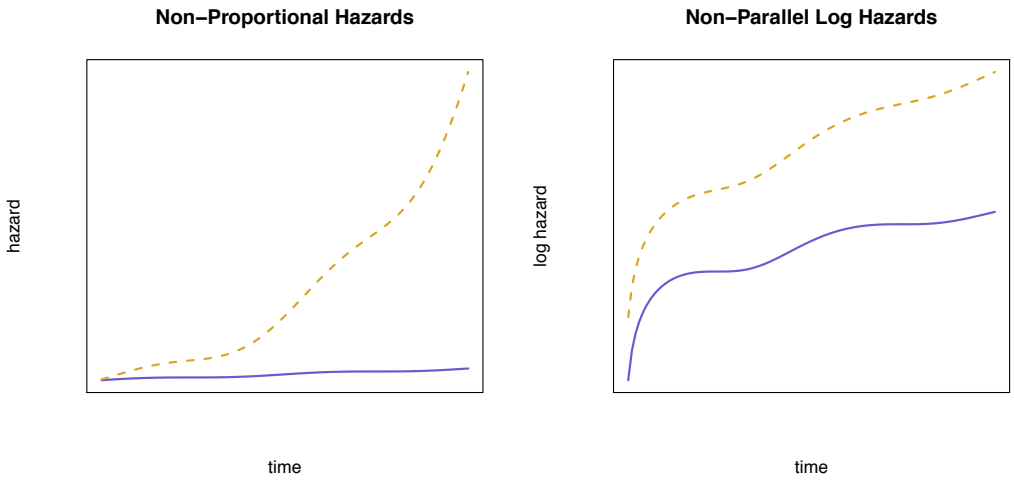
Here the hazard ratio depends on time through the interaction term:

$$\begin{aligned} \lambda(t|x+1) &= \lambda_0(t)e^{\beta_1(x+1)+\beta_2(x+1)\cdot f(t)} \\ \lambda(t|x) &= \lambda_0(t)e^{\beta_1 x+\beta_2 x\cdot f(t)} \\ \text{hazard ratio} &= e^{\beta_1(x+1)+\beta_2(x+1)\cdot f(t)-\beta_1 x-\beta_2 x\cdot f(t)} = e^{\beta_1+\beta_2 f(t)} \end{aligned}$$

Requires a hypothesized functional form for the dependence of the hazard ratio at time t on t.

Commonly-used functions of t are $f(t) = t$ and $f(t) = \log(t)$.

TIME INTERACTIONS

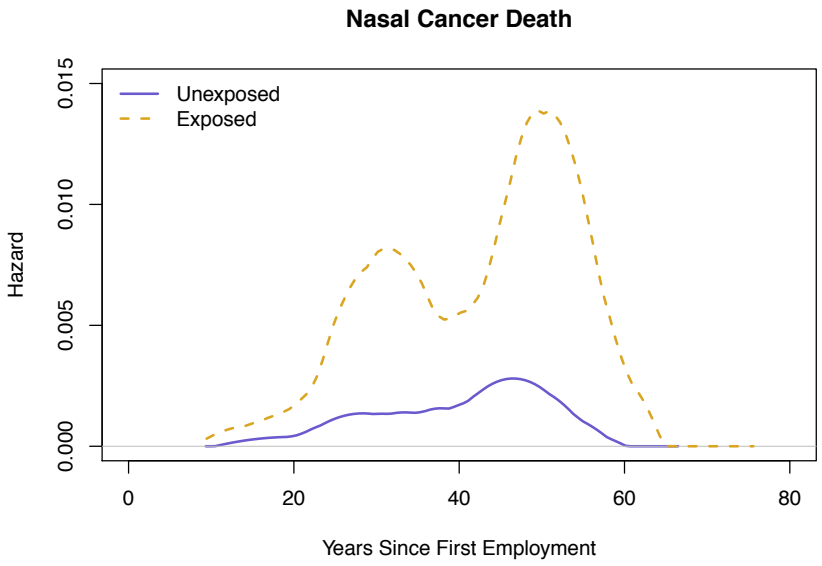


NASAL CANCER TIME INTERACTION

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	1.540408	4.666495	0.3503185	4.397165	1.1e-05

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
exposedTRUE	1.0613334	2.890222	5.161871	0.2056102	0.8370954
tt(exposed)	0.1290554	1.137753	1.388229	0.0929641	0.9259321

NASAL CANCER



ESTIMATING THE HR AS A FUNCTION OF TIME

- In exploratory analyses, may be of interest to estimate how the hazard ratio varies over time
- Estimate based on ratio of kernel-smoothed hazard estimates can be very variable
- Better choice is based on smoothed Schoenfeld residuals
- Can be thought of as an estimate of a time-dependent coefficient of a fixed variable

ESTIMATING THE HR AS A FUNCTION OF TIME

Another way for the hazard ratio to depend on time: time-dependent coefficients.

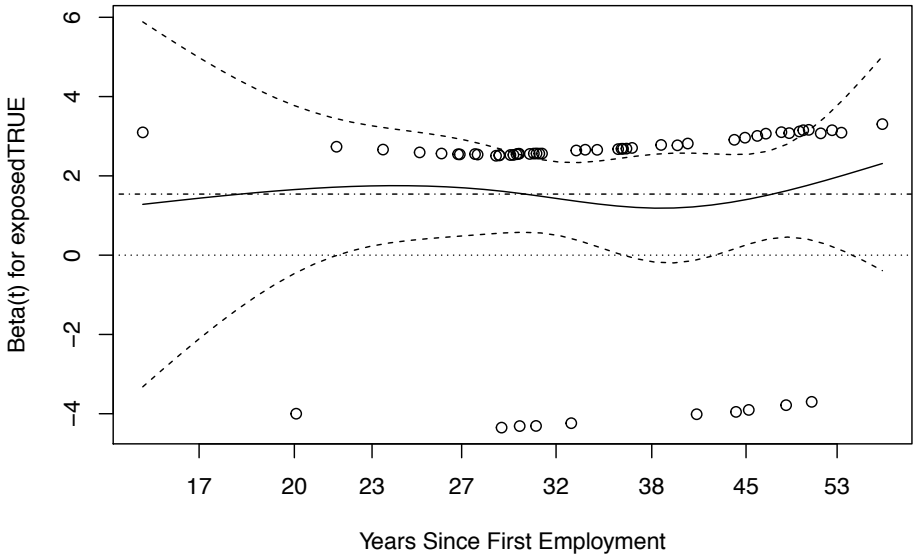
$$\lambda(t) = \lambda_0(t)e^{\beta(t)x}$$

Here the hazard ratio depends on time through the time-dependent coefficient $\beta(t)$

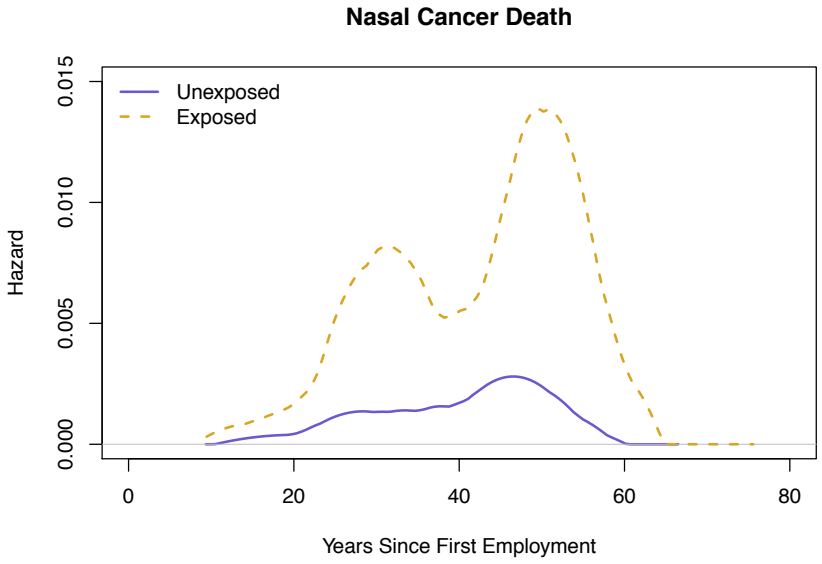
$$\begin{aligned} \lambda(t|x+1) &= \lambda_0(t)e^{\beta(t)(x+1)} \\ \lambda(t|x) &= \lambda_0(t)e^{\beta(t)(x)} \\ \text{hazard ratio} &= e^{\beta(t)(x+1)-\beta(t)x} = e^{\beta(t)} \end{aligned}$$

Estimated hazard ratio can be an arbitrary function of time $e^{\beta(t)}$.

NASAL HR ESTIMATE



NASAL CANCER



In R

Load packages.

```
library(survival)
library(Epi)
library(muhaz)
library(foreign)
library(knitr)
```

Get data.

```
data(nickel)
df <- nickel
head(nickel)
```

```
##   id icd exposure      dob  age1st   agein  ageout
## 1  3   0        5 1889.019 17.4808 45.2273 92.9808
## 2  4 162        5 1885.978 23.1864 48.2684 63.2712
## 3  6 163       10 1881.255 25.2452 52.9917 54.1644
## 4  8 527        9 1886.340 24.7206 47.9067 69.6794
## 5  9 150        0 1879.500 29.9575 54.7465 76.8442
## 6 10 163        2 1889.915 21.2877 44.3314 62.5413
```

Create variables

```
df$dead <- df$icd > 0
df$lungca <- with(df, icd == 162 | icd == 163)
df$nasalca <- with(df, icd == 160)
df$other <- with(df, icd > 0 & icd !=160 & icd !=162 & icd != 163)
df$etype <- with(df, factor((icd > 0) + lungca + 2*nasalca,
                 labels = c("Alive", "Dead Other", "Dead Lung Ca",
                            "Dead Nasal Ca")))
table(df$etype)
```

```
##
##         Alive    Dead Other  Dead Lung Ca Dead Nasal Ca
##            47           439           137            56
```

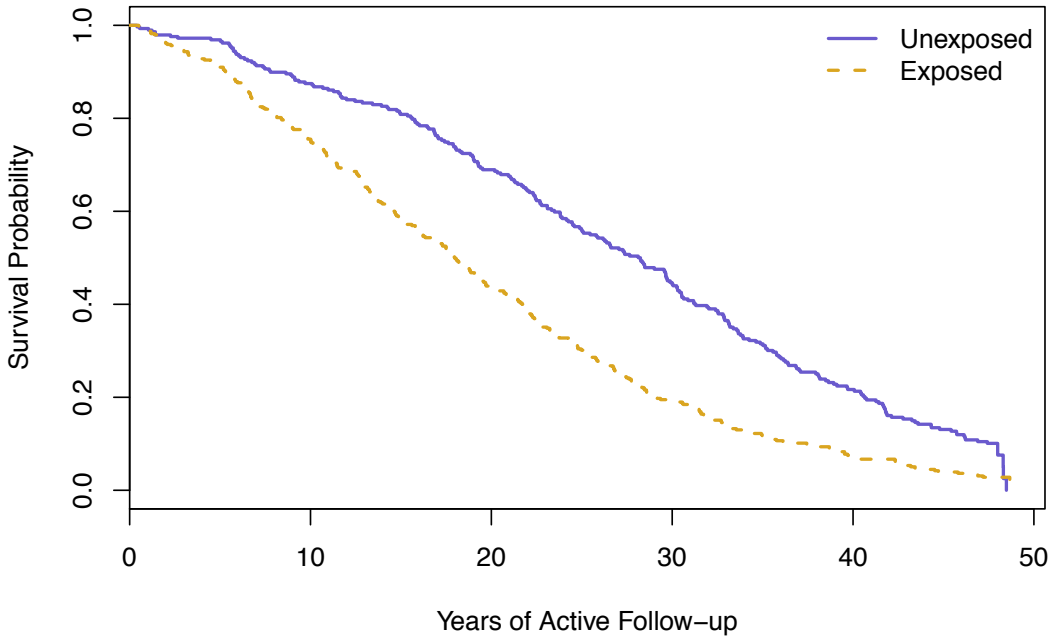
Create Variables

```
tempyr <- (df$dob + df$age1st)
df$yfe <- (tempyr > 1909.999) + (tempyr > 1914.999) +
  (tempyr > 1919.999)
df$yfe <- factor(df$yfe, labels = c("1900-1909","1910-1914",
                                    "1915-1919", "1920-1925"))
df$afe <- (df$age1st > 19.999) + (df$age1st > 27.499) +
  (df$age1st > 34.999)
df$afe <- factor(df$afe, labels = c("15-19", "20-27.5",
                                    "27.5 - 35", "35+"))
df$time <- with(df, ageout-agein)
df$exp <- (df$exposure > 0) + (df$exposure > 4.0) +
  (df$exposure > 8.0) + (df$exposure > 12.4)
df$exp <- factor(df$exp, labels = c("0", "0.5 - 4.0",
                "4.5 - 8.0","8.5-12.0", "12.5+"))
df$exposed <- df$exposure > 0
df$yfec <- (df$age1st + df$dob - 1915)/10
df$yfec2 <- df$yfec^2
df$logexp <- log(df$exposure + 1)
df$logage <- log(df$age1st - 10)
Y <- with(df, Surv(time, dead))
```

First KM All-Cause Mortality

```
colors <- c("slateblue", "goldenrod")
survtime <- survfit(Y ~ exposed, data  = df)
plot(survtime, xlab  = "Years of Active Follow-up",
               ylab = "Survival Probability", col = colors, lwd = 2,
     lty = c(1,2) )
legend("topright", col = colors, lty = c(1,2), lwd = 2,
       legend = c("Unexposed", "Exposed"), bty = "n")
```

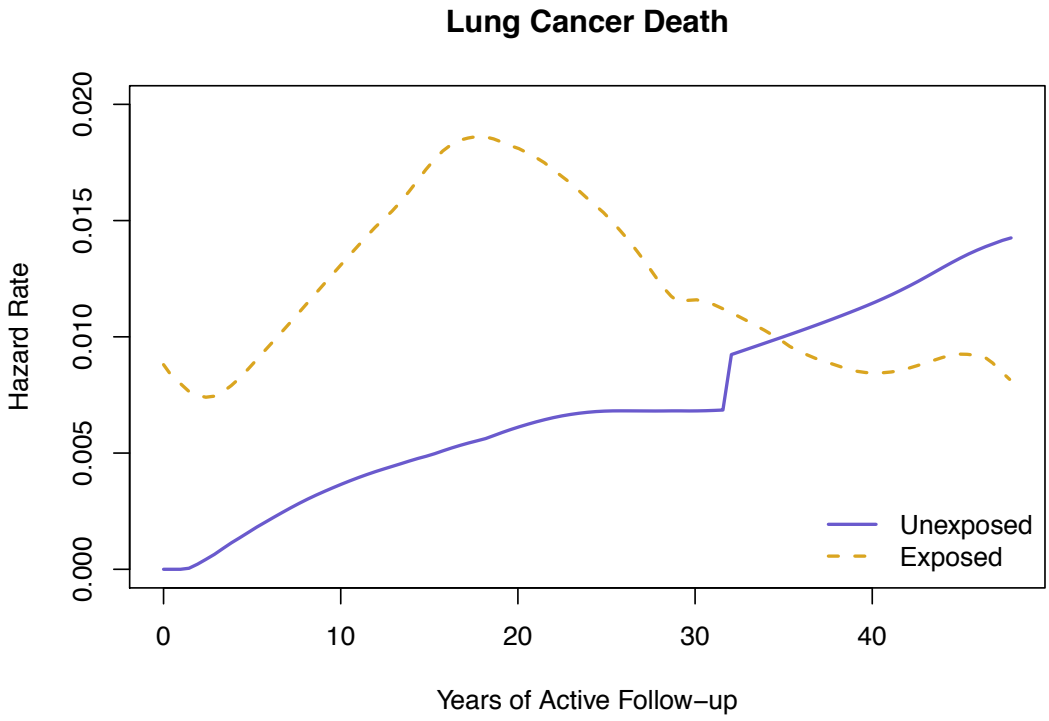
First KM All-Cause Mortality



Cause-specific Hazards

```
lungcahaz1 <- with(df, muhaz(time, lungca, subset = exposed))
lungcahaz0 <- with(df, muhaz(time, lungca, subset = !exposed))
plot(lungcahaz0, col = colors[1], lwd = 2, ylim = c(0,.02),
     xlab = "Years of Active Follow-up")
lines(lungcahaz1, col = colors[2], lwd = 2, lty = 2)
legend("bottomright", col = colors, lty = c(1,2), lwd = 2,
       legend = c("Unexposed", "Exposed"), bty = "n")
title(main = "Lung Cancer Death")
```

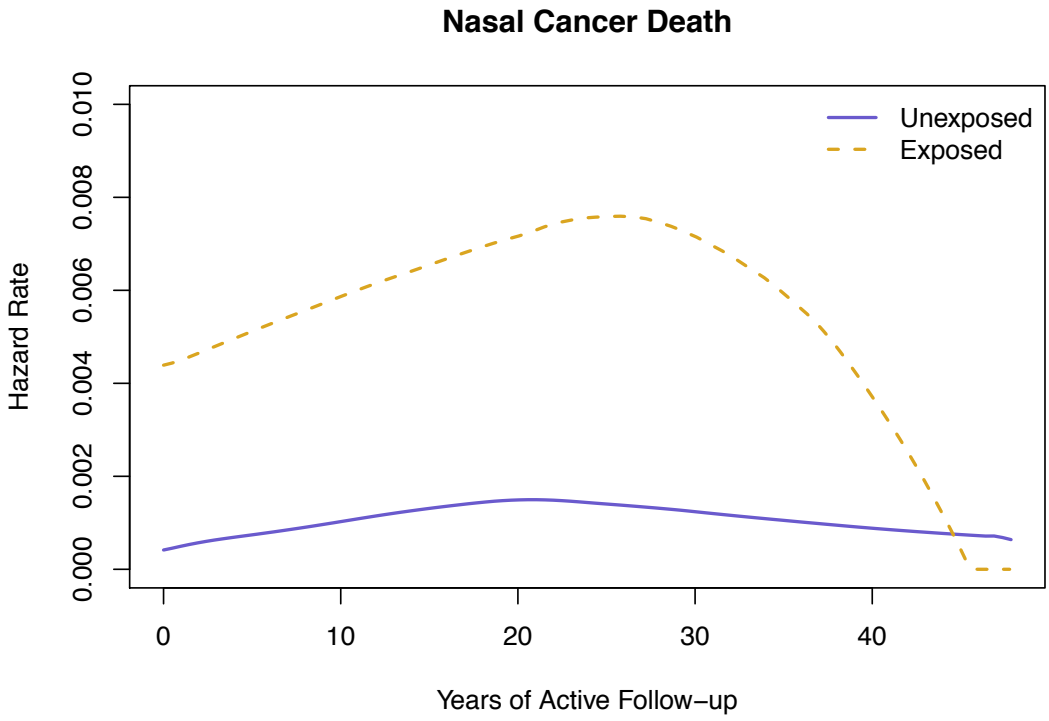
Cause-specific Hazards



Cause-specific Hazards

```
nasalhaz1 <- with(df, muhaz(time, nasalca, subset = exposed))
nasalhaz0 <- with(df, muhaz(time, nasalca, subset = !exposed))
otherhaz <- with(df, muhaz(time, other))
plot(nasalhaz0, col = colors[1], lwd = 2, ylim = c(0,.01),
     xlab = "Years of Active Follow-up")
lines(nasalhaz1, col = colors[2], lwd = 2, lty = 2)
legend("topright", col = colors, lty = c(1,2), lwd = 2, legend = c(
  "Unexposed", "Exposed"), bty = "n")
title(main = "Nasal Cancer Death")
```

Cause-specific Hazards

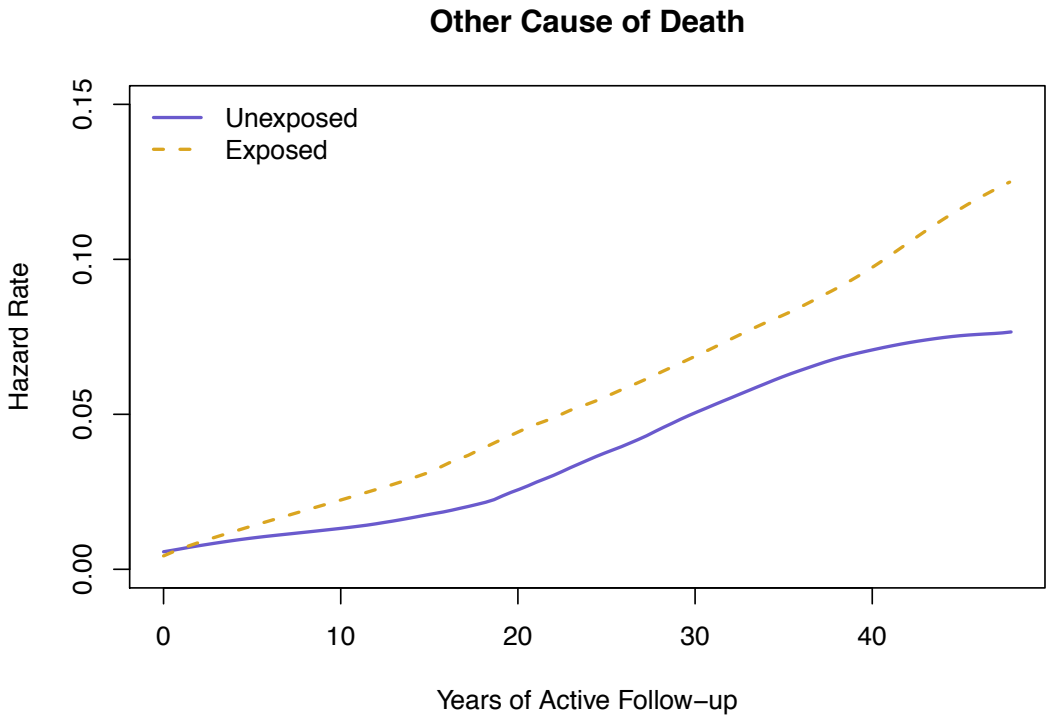


Cause-specific Hazards

```
otherhaz1 <- with(df, muhaz(time, other, subset = exposed))
otherhaz0 <- with(df, muhaz(time, other, subset = !exposed))

plot(otherhaz0, col = colors[1], lwd = 2, ylim = c(0,.15),
     xlab = "Years of Active Follow-up")
lines(otherhaz1, col = colors[2], lwd = 2, lty = 2)
legend("topleft", col = colors, lty = c(1,2), lwd = 2, legend = c(
  "Unexposed", "Exposed"), bty = "n")
title(main = "Other Cause of Death")
```

Cause-specific Hazards



Cox models lung

```
lungobs <- with(df, Surv(ageout - agein, lungca))

crudelungobs <- coxph(lungobs ~ exposed, data = df)
adjlungobs <- coxph(lungobs ~ exp + afe + yfe, data = df)
coef(summary(crudelungobs))
```

```
##                   coef exp(coef)  se(coef)        z     Pr(>|z|)
## exposedTRUE 0.9200182  2.509336 0.1869493 4.921217 8.600789e-07
```

```
coef(summary(adjlungobs))
```

```
##                     coef exp(coef)  se(coef)          z     Pr(>|z|)
## exp0.5 - 4.0   0.6030012 1.8275955 0.2121299  2.8426041 0.0044746619
## exp4.5 - 8.0   1.0862839 2.9632419 0.2828485  3.8405146 0.0001227767
## exp8.5-12.0    1.2772969 3.5869307 0.3742268  3.4131628 0.0006421356
## exp12.5+       1.4873597 4.4253955 0.4798472  3.0996524 0.0019374785
## afe20-27.5     0.8103938 2.2487934 0.3079688  2.6314149 0.0085030159
## afe27.5 - 35   0.9149895 2.4967489 0.3291081  2.7802097 0.0054323806
## afe35+         0.8068991 2.2409482 0.4237839  1.9040342 0.0569057338
## yfe1910-1914   0.3342204 1.3968510 0.2695145  1.2400835 0.2149444997
## yfe1915-1919 -0.1340505 0.8745459 0.3749097 -0.3575540 0.7206771345
## yfe1920-1925   0.0744977 1.0773429 0.2966621  0.2511197 0.8017215522
```

Cox models nasal

```
nasalobs <- with(df, Surv(ageout - agein, nasalca))

crudenasalobs <- coxph(nasalobs ~ exposed, data = df)
adjnasalobs <- coxph(nasalobs ~ exp + afe + yfe, data = df)
coef(summary(crudenasalobs))
```

```
##                 coef exp(coef)  se(coef)        z     Pr(>|z|)
## exposedTRUE 1.614074  5.023236 0.3516507 4.589994 4.432595e-06
```

```
coef(summary(adjnasalobs))
```

```
##                     coef  exp(coef)  se(coef)         z     Pr(>|z|)
## exp0.5 - 4.0   0.8356274  2.3062606 0.4032111  2.072432 3.822520e-02
## exp4.5 - 8.0   1.1366437  3.1162916 0.4706657  2.414970 1.573650e-02
## exp8.5-12.0    2.2945326  9.9197981 0.5117936  4.483316 7.349197e-06
## exp12.5+       2.8713357 17.6605917 0.5697217  5.039892 4.657943e-07
## afe20-27.5     1.4686105  4.3431963 0.7518514  1.953326 5.078102e-02
## afe27.5 - 35   2.1598639  8.6699580 0.7588726  2.846148 4.425157e-03
## afe35+         3.4767227 32.3535148 0.7843101  4.432842 9.299915e-06
## yfe1910-1914   0.7130093  2.0401213 0.3728470  1.912337 5.583292e-02
## yfe1915-1919   0.5040978  1.6554913 0.5034466  1.001293 3.166849e-01
## yfe1920-1925  -0.9304088  0.3943924 0.5152666 -1.805684 7.096766e-02
```

Cox model other

```
otherobs <- with(df, Surv(ageout - agein, other))
crudeotherobs <- coxph(otherobs ~ exposed, data = df)
adjotherobs <- coxph(otherobs ~ exp + afe + yfe, data = df)
coef(summary(crudeotherobs))
```

```
##                  coef exp(coef)   se(coef)        z     Pr(>|z|)
## exposedTRUE 0.3962896    1.4863 0.09720561 4.076818 4.565619e-05
```

```
coef(summary(adjotherobs))
```

```
##                     coef exp(coef)  se(coef)          z     Pr(>|z|)
## exp0.5 - 4.0   0.13180810 1.1408894 0.1105672  1.1921083 2.332188e-01
## exp4.5 - 8.0   0.13087353 1.1398236 0.1603797  0.8160231 4.144869e-01
## exp8.5-12.0    0.03249138 1.0330250 0.2563862  0.1267282 8.991555e-01
## exp12.5+      -0.07741111 0.9255093 0.3964677 -0.1952520 8.451957e-01
## afe20-27.5     0.52755484 1.6947832 0.1539622  3.4265217 6.113649e-04
## afe27.5 - 35   1.10703760 3.0253827 0.1653356  6.6956992 2.146427e-11
## afe35+         1.97406257 7.1998671 0.1942464 10.1626701 0.000000e+00
## yfe1910-1914  -0.21481121 0.8066937 0.1515491 -1.4174361 1.563555e-01
## yfe1915-1919  -0.52976792 0.5887416 0.1766843 -2.9983870 2.714128e-03
## yfe1920-1925  -1.14563904 0.3180206 0.1502442 -7.6251795 2.442491e-14
```

Time since First Employment

```
tfelung <- with(df, Surv(agein - age1st, ageout-age1st,lungca))
tfenasal <- with(df, Surv(agein - age1st, ageout-age1st, nasalca))
tfeother <- with(df, Surv(agein - age1st, ageout-age1st, other))
```

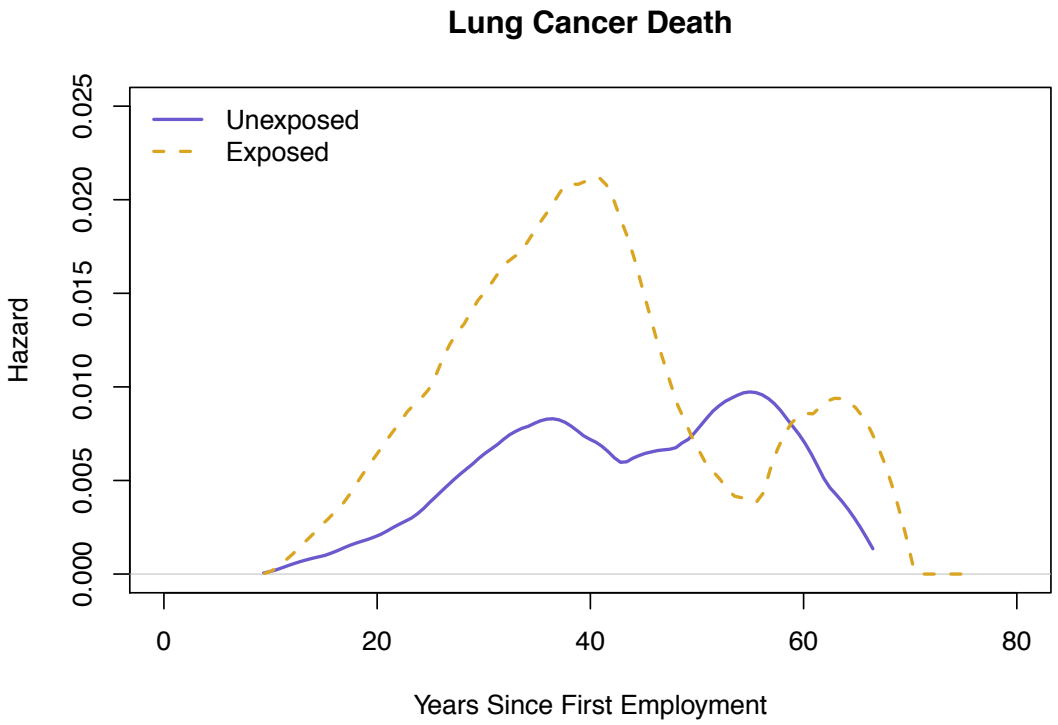
My kernel-smoothed hazard function

```
myhaz <- function(survfit.obj, numt = 100){
  x <- survfit.obj
  ok <- x$n.risk > 0
  u <- x$time[ok]
  w <- x$n.event[ok]/x$n.risk[ok]
  hazard <- density(u, weight = w, kernel = "epanechnikov", n = numt,
                    from = min(x$time), to = max(x$time))
  }
```

Cause-specific Hazards TFE

```
fitlung <- survfit(tfelung ~ exposed, data = df)
lunghaz1 <- myhaz(fitlung[1])
lunghaz0 <- myhaz(fitlung[2])
plot(lunghaz1, xlab = "Years Since First Employment", ylab = "Hazard",
     main = "Lung Cancer Death", col = colors[1], lwd = 2,
     xlim = c(0, 80), ylim = c(0, .025) )
lines(lunghaz0, lwd = 2, col = colors[2], lty= 2)
legend("topleft", col = colors, lty = c(1,2), lwd = 2,
       legend = c("Unexposed", "Exposed"), bty = "n")
```

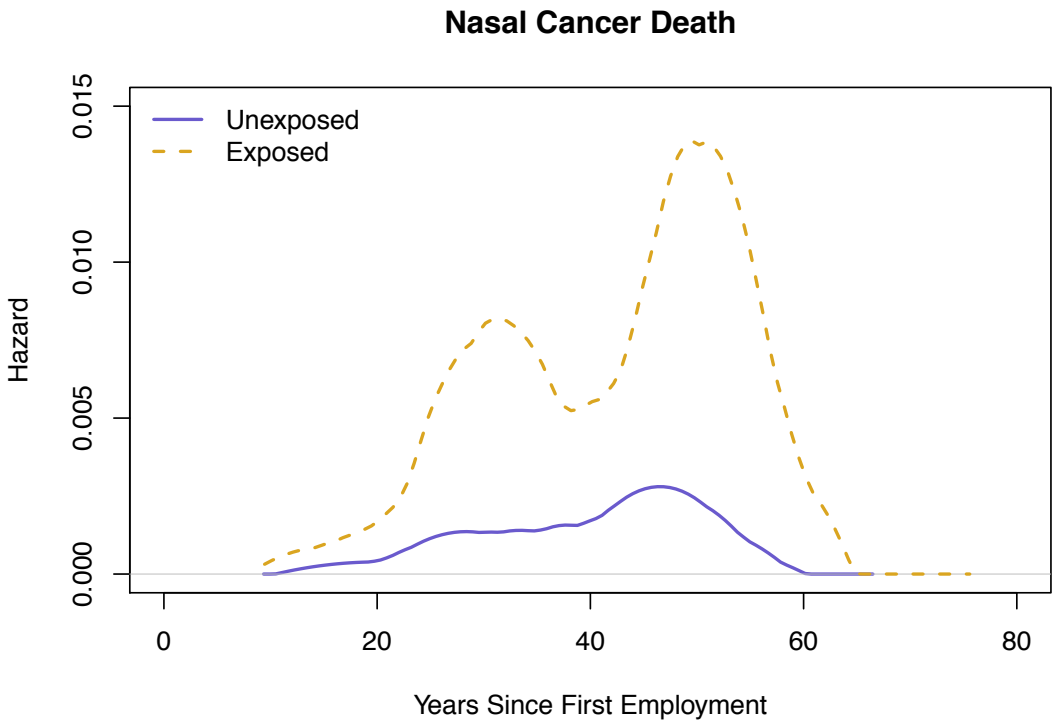
Cause-specific Hazards TFE



Cause-specific Hazards TFE

```
fitnasal <- survfit(tfenasal ~ exposed, data = df)
nasalhaz1 <- myhaz(fitnasal[1])
nasalhaz0 <- myhaz(fitnasal[2])
plot(nasalhaz1, xlab = "Years Since First Employment", ylab = "Hazard",
     main = "Nasal Cancer Death", col = colors[1], lwd = 2,
     xlim = c(0, 80), ylim = c(0, .015) )
lines(nasalhaz0, lwd = 2, col = colors[2], lty= 2)
legend("topleft", col = colors, lty = c(1,2), lwd = 2,
       legend = c("Unexposed", "Exposed"), bty = "n")
```

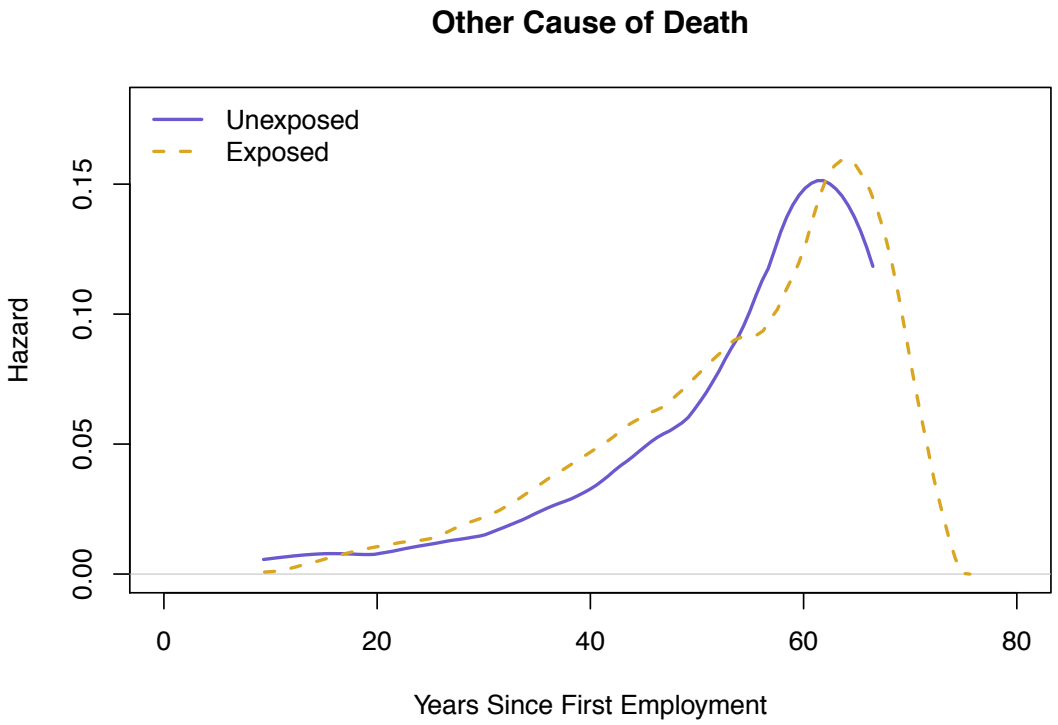
Cause-specific Hazards TFE



Cause-specific Hazards TFE

```
fitother <- survfit(tfeother ~ exposed, data = df)
otherhaz1 <- myhaz(fitother[1])
otherhaz0 <- myhaz(fitother[2])
plot(otherhaz1, xlab = "Years Since First Employment", ylab = "Hazard",
     main = "Other Cause of Death", col = colors[1], lwd = 2,
     xlim = c(0,80), ylim = c(0, .18) )
lines(otherhaz0, lwd = 2, col = colors[2], lty= 2)
legend("topleft", col = colors, lty = c(1,2), lwd = 2,
       legend = c("Unexposed", "Exposed"), bty = "n")
```

Cause-specific Hazards TFE



Cox models lung TFE

```
crudelungtfe <- coxph(tfelung ~ exposed, data = df)
adjlungtfe <- coxph(tfelung ~ exp + afe + yfe, data = df)
coef(summary(crudelungtfe))
```

```
##                  coef exp(coef)  se(coef)        z    Pr(>|z|)
## exposedTRUE 0.8000334  2.225615 0.1860041 4.301159 1.69907e-05
```

```
coef(summary(adjlungtfe))
```

```
##                    coef exp(coef)  se(coef)        z     Pr(>|z|)
## exp0.5 - 4.0  0.6111674  1.842581 0.2123734 2.877796 0.0040046375
## exp4.5 - 8.0  1.0952795  2.990018 0.2838639 3.858467 0.0001141003
## exp8.5-12.0   1.2880174  3.625591 0.3739070 3.444754 0.0005715804
## exp12.5+      1.4327121  4.190048 0.4791166 2.990320 0.0027868489
## afe20-27.5    0.7604881  2.139320 0.3081636 2.467806 0.0135943973
## afe27.5 - 35  0.8670846  2.379962 0.3281099 2.642665 0.0082256384
## afe35+        0.7982183  2.221579 0.4224336 1.889571 0.0588153762
## yfe1910-1914  0.4358460  1.546271 0.2724801 1.599552 0.1096981002
## yfe1915-1919  0.1753274  1.191636 0.3775109 0.464430 0.6423396667
## yfe1920-1925  0.6547157  1.924595 0.2991155 2.188839 0.0286085617
```

Cox models nasal TFE

```
crudenasaltfe <- coxph(tfenasal ~ exposed, data = df)
adjnasaltfe <- coxph(tfenasal ~ exp + afe + yfe, data = df)
coef(summary(crudenasaltfe))
```

```
##                  coef exp(coef)  se(coef)        z     Pr(>|z|)
## exposedTRUE 1.540408  4.666495 0.3503185 4.397165 1.096741e-05
```

```
coef(summary(adjnasaltfe))
```

```
##                    coef exp(coef)  se(coef)          z     Pr(>|z|)
## exp0.5 - 4.0  0.89583588  2.449382 0.4044464 2.21496802 2.676226e-02
## exp4.5 - 8.0  1.19917175  3.317368 0.4727052 2.53682770 1.118620e-02
## exp8.5-12.0   2.32148155 10.190761 0.5173928 4.48688417 7.227234e-06
## exp12.5+      2.86559196 17.559445 0.5727364 5.00333459 5.634702e-07
## afe20-27.5    1.47218695  4.358757 0.7527320 1.95579169 5.048970e-02
## afe27.5 - 35  2.17703118  8.820082 0.7601145 2.86408340 4.182179e-03
## afe35+        3.60258884 36.693104 0.7886401 4.56810258 4.921592e-06
## yfe1910-1914  1.03737012  2.821786 0.3798834 2.73075925 6.318861e-03
## yfe1915-1919  1.12915201  3.093033 0.5130845 2.20071374 2.775630e-02
## yfe1920-1925  0.01669652  1.016837 0.5257787 0.03175579 9.746668e-01
```

Cox models other TFE

```
crudeothertfe <- coxph(tfeother ~ exposed, data = df)
adjothertfe <- coxph(tfeother ~ exp + afe + yfe, data = df)
coef(summary(crudeothertfe))
```

```
##                  coef exp(coef)   se(coef)        z   Pr(>|z|)
## exposedTRUE 0.2164895   1.24171 0.09661312 2.240788 0.02503981
```

```
coef(summary(adjothertfe))
```

```
##                     coef exp(coef)  se(coef)           z     Pr(>|z|)
## exp0.5 - 4.0  0.16852497  1.183558 0.1106070  1.52363760 1.275993e-01
## exp4.5 - 8.0  0.23605613  1.266245 0.1602288  1.47324454 1.406851e-01
## exp8.5-12.0   0.05852009  1.060266 0.2564181  0.22822131 8.194742e-01
## exp12.5+      0.02454565  1.024849 0.3964995  0.06190588 9.506378e-01
## afe20-27.5    0.57047737  1.769111 0.1545876  3.69031860 2.239734e-04
## afe27.5 - 35  1.16561360  3.207891 0.1665088  7.00031359 2.553957e-12
## afe35+        2.08358859  8.033245 0.1957375 10.64480862 0.000000e+00
## yfe1910-1914  0.20870807  1.232085 0.1540413  1.35488425 1.754544e-01
## yfe1915-1919  0.23294527  1.262312 0.1788233  1.30265630 1.926921e-01
## yfe1920-1925  0.10243855  1.107869 0.1529133  0.66991265 5.029135e-01
```

Interactions with time

```
binnasaltfe <- coxph(tfenasal ~ exposed, data = df)
binnasaltfeint <- coxph(tfenasal ~ exposed + tt(exposed), data = df,
                        tt = function(x, t, ...) x * log(t+5))
#adjnasaltfe <- coxph(tfenasal ~ exposure  + afe + yfe, data = df)
coef(summary(binnasaltfe))
```

```
##                  coef exp(coef)  se(coef)        z     Pr(>|z|)
## exposedTRUE 1.540408  4.666495 0.3503185 4.397165 1.096741e-05
```

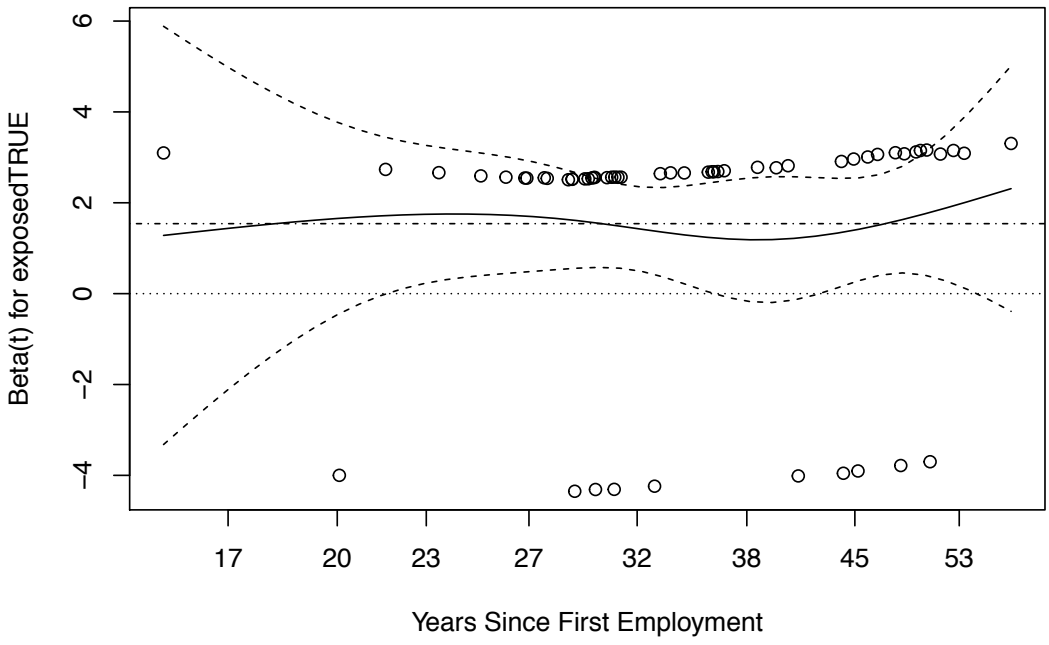
```
coef(summary(binnasaltfeint))
```

```
##                    coef exp(coef) se(coef)          z  Pr(>|z|)
## exposedTRUE 1.0613334  2.890222 5.161871 0.20561022 0.8370954
## tt(exposed) 0.1290554  1.137753 1.388229 0.09296406 0.9259321
```

Estimating HR as a function of time

```
temp <- cox.zph(binnasaltfe, transform = function(t){log(t)})
plot(temp, xlab = "Years Since First Employment")
abline(h = 0, lty = 3)
abline(h = coef(binnasaltfe), lty = 4)
```

Estimating HR as a function of time



Your turn

Using the "nickel" data in the Epi package in R:

1. Plot cause-specific hazard functions as a function of age
2. Fit Cox models as a function of age
3. Look for interactions with time = age in Cox models
4. Estimate the coefficient of exposed (or exposure) as a function of time = age

SESSION 4:IMMORTAL TIME BIAS AND TIME-DEPENDENT COVARIATES

Module 8: Survival Analysis for Observational Data

Summer Institute in Statistics for Clinical Research
University of Washington
July, 2016

Barbara McKnight, Ph.D.

OUTLINE

- Immortal-time bias
- Correction using time-dependent covariates
- Examples:
  - Stanford Heart Transplant data
- More complicated time-dependent covariates

IMMORTAL TIME BIAS

- Suissa S. Immortal time bias in observational studies of drug effects. Pharmacoepidem Drug Safe. 2007 Mar 1;16(3):241–249.
- When exposed time is counted incorrectly as an exposed person or not counted as at risk, while surviving until exposure occurs.
  - Diabetics, use of statins and outcome of starting insulin therapy
  - Heart-failure hospital patients, prescription for beta-blockers, and outcome of readmission to hospital

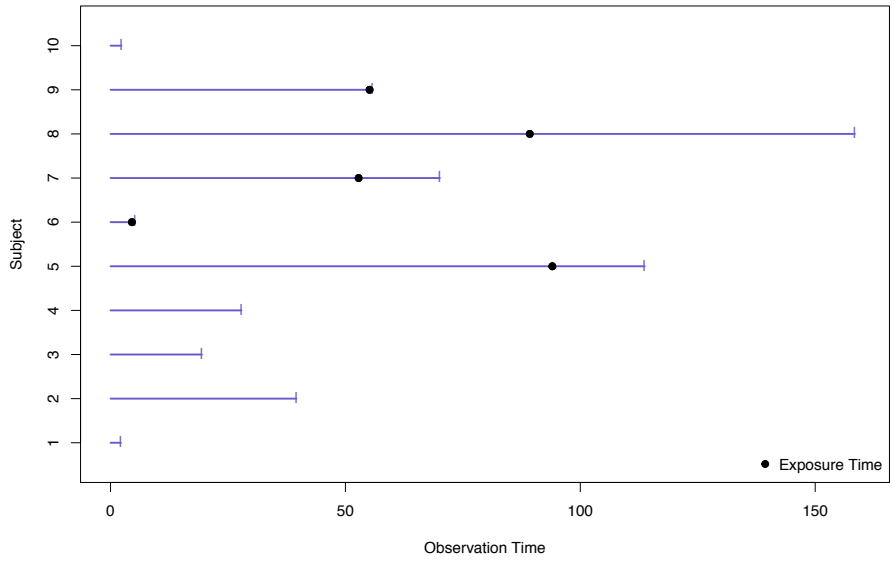
OLDER EXAMPLES

- Survival of “responders” vs “non-responders” in Cancer clinical trials.
- Hormone use in cohort with Benign Breast Disease and Breast cancer risk
- Effectiveness of Heart Transplant in prolonging survival

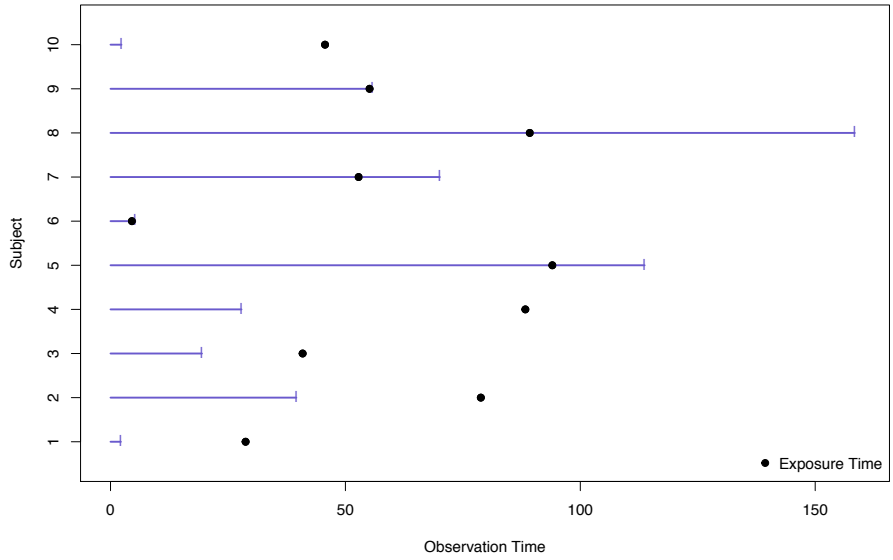
PROBLEM

- Subject spends some time under observation for outcome before “exposure” occurs
- Subject is not given credit for survival as a non-exposed person until exposure occurs
  - In some bad analyses, the time prior to exposure is omitted (left entry at exposure time)
  - In others, the subject is counted as exposed before exposure occurs
- In both cases, bias is toward making exposure appear to be associated with longer survival

PICTURE



PICTURE



BIAS

- Exposure times and survival times were generated independently.
- Mean survival time for those who were exposed before death:
- Mean survival time for those who were not exposed before death:
- REASON: Those who lived long enough to be exposed, lived <u>longer</u>

SIMULATION

- Previous plot was a subset of one of the simulated data sets
- No association between exposure and survival
- 1000 replications of sample size 100
- Compare two analysis strategies
  - Ordinary Cox model counting any subject exposed before death as exposed
  - Cox model left entering exposed subjects when they are exposed.

SIMULATION

- Ordinary Cox model counting any subject exposed before death as exposed:
  - All coefficients negative, indicating protective effect of exposure.
- Cox model with left entry at exposure time for exposed observations:
  - All coefficients negative.

	mean coefficient	power
ordinary	-1.8036237	1.00
left-enter	-0.9475986	0.94

SOLUTION

- Time-dependent exposure variable!
- Let subject be categorized as not exposed at times before exposure occurs, and let exposure status change when exposure has occurred

SOLUTION

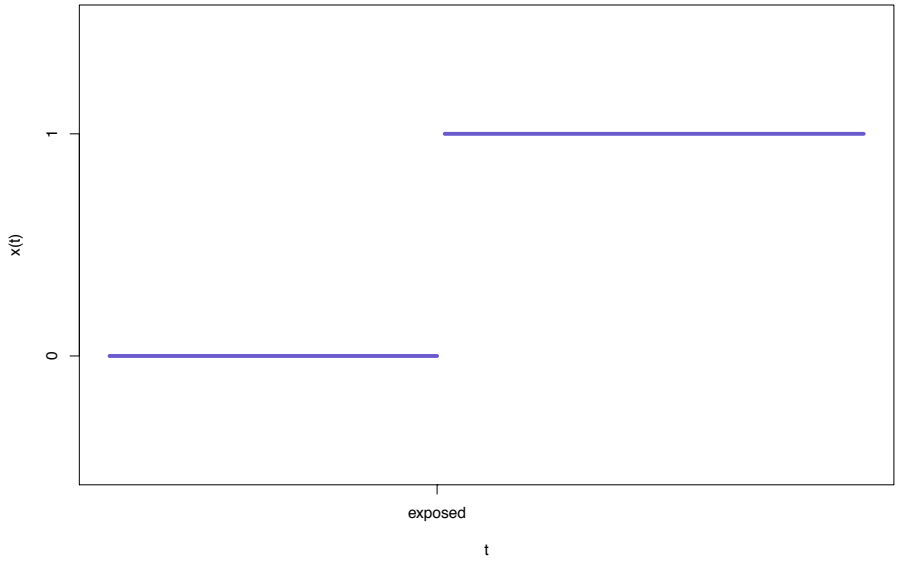
Let the time-dependent binary prior exposure variable be:

$$x(t) = \begin{cases} 1 & \text{exposed prior to time } t \\ 0 & \text{Otherwise} \end{cases}.$$

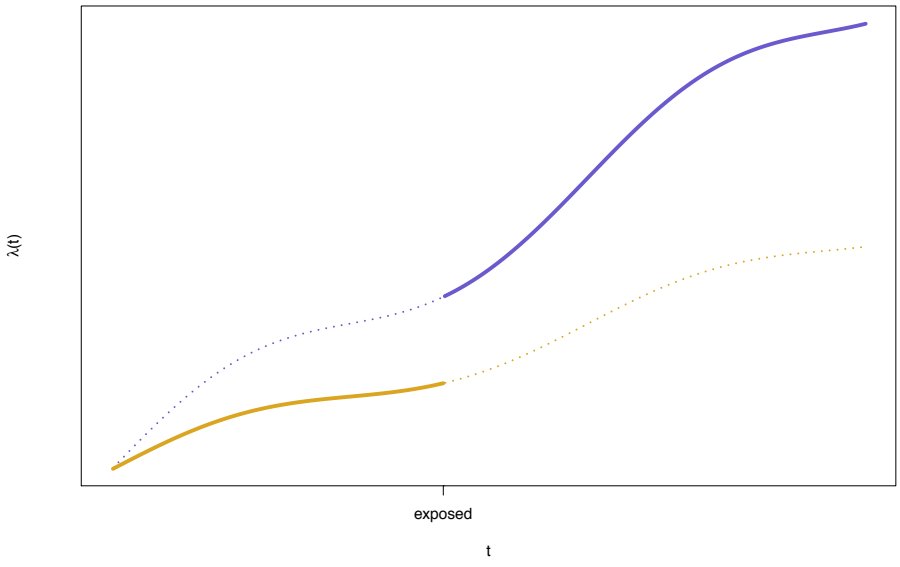
Then the model is

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

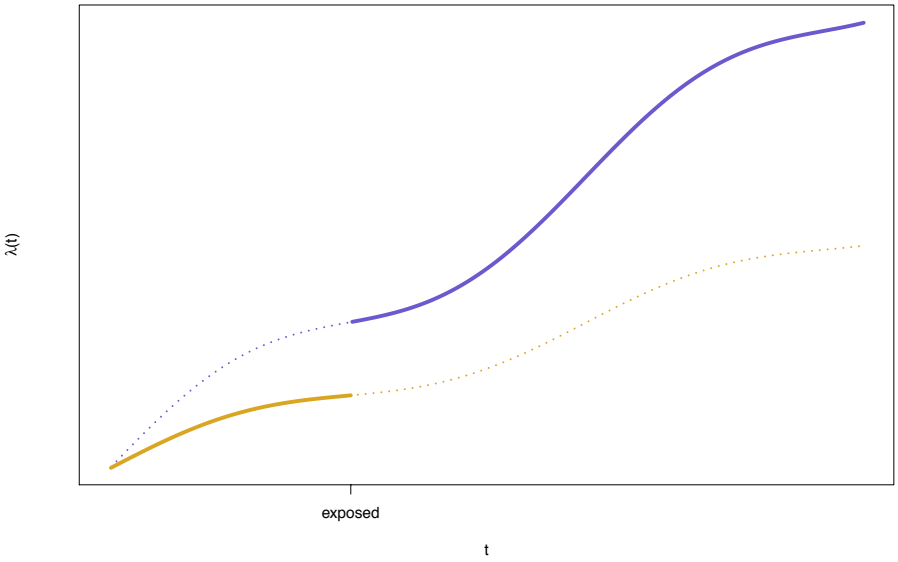
TIME-DEPENDENT EXPOSURE



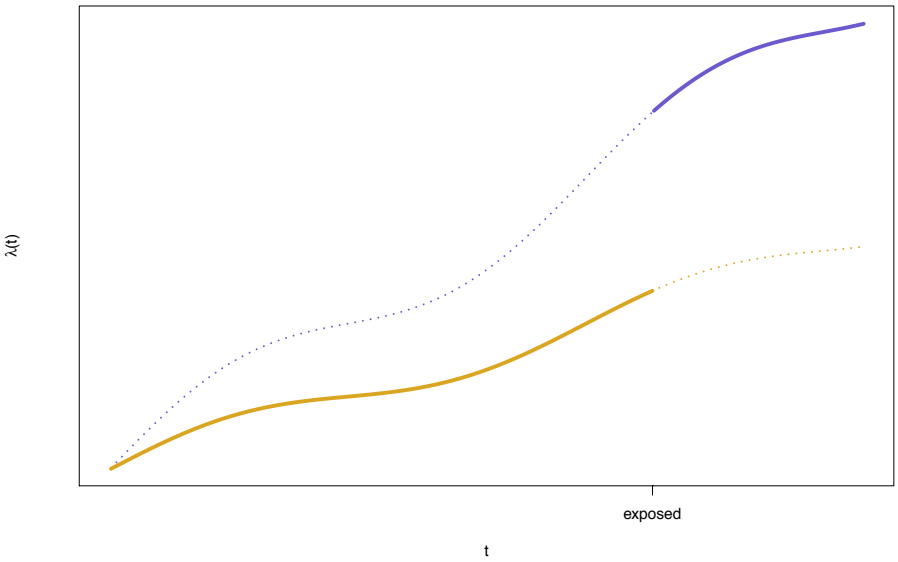
TIME-DEPENDENT EXPOSURE



EARLIER



LATER



SOLUTION

- Exposed subject contributes survival to risk sets before s/he is exposed
- Exposed subject contributes survival to risk sets after s/he is exposed until censoring or death
- Exposed subject contributes death to risk set when s/he dies

SIMULATION

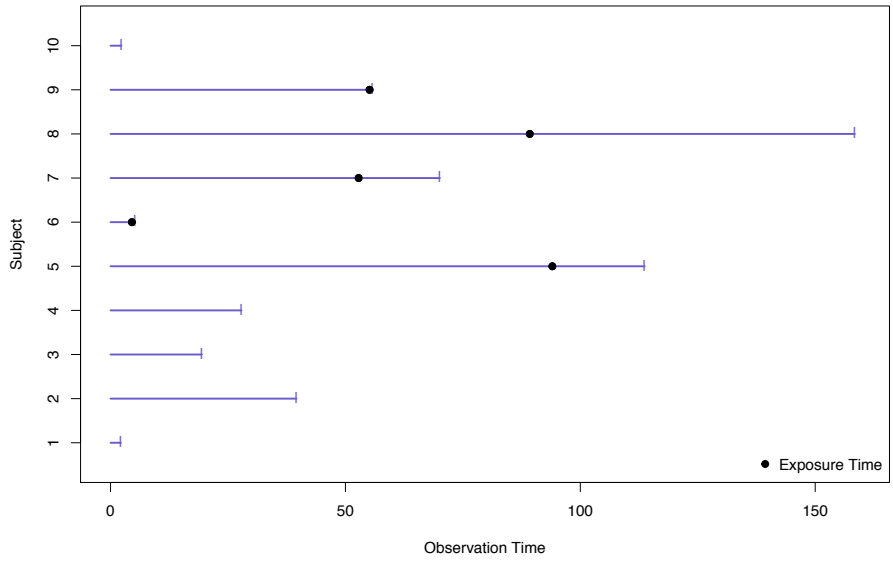
Compare to correct time-dependent exposure model:

	mean coefficient	power
ordinary	-1.8036237	1.000
left-enter	-0.9475986	0.940
correct	-0.0061170	0.048

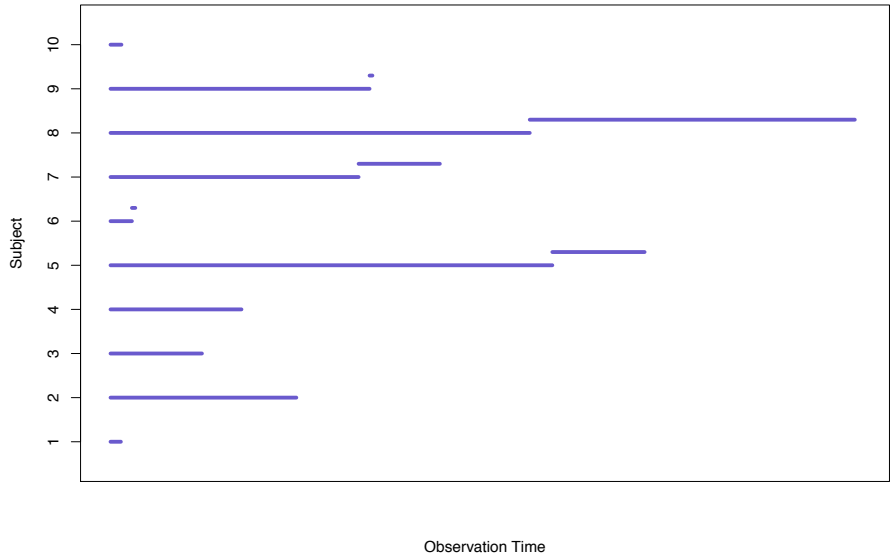
HOW TO DO IT

- Divide exposed subjects' information into two records:
- The first record starts at time zero (or entry into observation), has exposure coded as unexposed, and removes the subject from risk sets (as if censored) at the time of exposure.
- The second record starts at the time of exposure, has exposure coded as exposed, and follows subjects until they die.

PICTURE



PICTURE



EXAMPLE

- Early days of Stanford Heart Transplant program
  - Subjects admitted to program when heart condition was sufficiently severe
  - Donor heart was sought
  - Some patients received heart
  - Some died before a suitable heart could be found
- Question: did heart transplant prolong survival?

STANFORD

- Without covariables
- Naïve model examines survival as a function of whether subject received a heart transplant
- Subjects who lived long enough to receive a transplant lived longer:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
transplant	-1.323445	0.2662166	0.2438026	-5.428345	1e-07

STANFORD

- With correct model for time-dependent transplant status:

	coef	exp(coef)	se(coef)	Z	P-value
Time-dependent transplant	0.127	1.14	0.30	0.422	0.673

- No evidence of influence of prior transplant

OTHER POSSIBILITIES

More than one change in status:

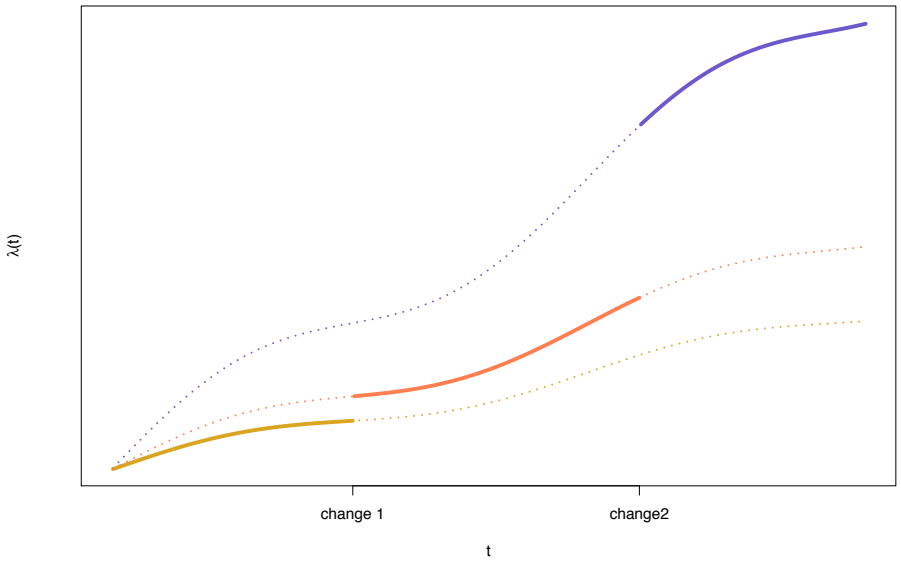
Let $\lambda(t)$ be the hazard for stroke:

$$x_{AF1}(t) = \begin{cases} 1 & \text{First Episode Atrial Fibrillation by } t \\ 0 & \text{Otherwise} \end{cases}$$

$$x_{AF2}(t) = \begin{cases} 1 & \text{Second Episode Atrial Fibrillation by } t \\ 0 & \text{Otherwise} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_{AF1}(t) + \beta_2 x_{AF2}(t)}$$

TWO CHANGES



OTHER POSSIBILITIES

A change in numerical value of a continuous variable.

Examples:

$x(t)$ = most recently recorded value of fasting insulin at time t.

$x(t)$ = cumulative recorded exposure to radon at time t.

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

PRIMARY BILIARY CIRRHOSIS

- 312 patients in RCT of d-penacillamine
- Some biomarkers were measured repeatedly over time
- Compare influence of baseline measures on survival (non-time-dependent model) to influence of most recent measure (time-dependent model) on survival.

PRIMARY BILIARY CIRRHOSIS

x = bilirubin (mg/dl) measured at baseline

$x(t)$ = most recently measured bilirubin (mg/dl) at day t.

Baseline model:

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

Time-dependent model:

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

PRIMARY BILIARY CIRRHOSIS

Baseline model:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
log(bili)	0.9890831	2.688768	0.0783597	12.62235	0

Time-dependent model:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
log(bili)	1.370255	3.936355	0.0949917	14.425	0

OTHER POSSIBILITIES

- Time-interaction with time-dependent exposure variable like prior heart transplant

Stanford heart transplant data

```
data(heart)
head(heart)
```

```
##   start stop event        age      year surgery transplant id
## 1     0   50     1 -17.155373 0.1232033       0          0  1
## 2     0    6     1   3.835729 0.2546201       0          0  2
## 3     0    1     0   6.297057 0.2655715       0          0  3
## 4     1   16     1   6.297057 0.2655715       0          1  3
## 5     0   36     0  -7.737166 0.4900753       0          0  4
## 6    36   39     1  -7.737166 0.4900753       0          1  4
```

Stanford heart transplant data

```
correct <- coxph(Surv(start, stop, event) ~ transplant, data  = heart)
wrong <- coxph(Surv(futime, fustat) ~ transplant, data = jasa)

coef(summary(wrong))
```

```
##                 coef exp(coef)  se(coef)         z     Pr(>|z|)
## transplant -1.323445 0.2662166 0.2438026 -5.428345 5.687894e-08
```

```
coef(summary(correct))
```

```
##                  coef exp(coef)  se(coef)         z  Pr(>|z|)
## transplant1 0.1271411  1.135577 0.3011411 0.4221978 0.6728806
```

PBC data

Select RCT participants and some variables

```
temp <- subset(pbc, id <= 312, select=c(id:sex, stage))
head(temp)
```

```
##   id time status trt      age sex stage
## 1  1  400      2   1 58.76523   f     4
## 2  2 4500      0   1 56.44627   f     3
## 3  3 1012      2   1 70.07255   m     4
## 4  4 1925      2   1 54.74059   f     4
## 5  5 1504      1   2 38.10541   f     3
## 6  6 2503      2   2 66.25873   f     3
```

Start creating tdc data frame by setting range of possible times for variables to change value.

```
pbctdc <- tmerge(temp, temp, id=id, death = event(time, status))
head(pbctdc)
```

```
##   id time status trt      age sex stage tstart tstop death
## 1  1  400      2   1 58.76523   f     4      0   400     2
## 2  2 4500      0   1 56.44627   f     3      0  4500     0
## 3  3 1012      2   1 70.07255   m     4      0  1012     2
## 4  4 1925      2   1 54.74059   f     4      0  1925     2
## 5  5 1504      1   2 38.10541   f     3      0  1504     1
## 6  6 2503      2   2 66.25873   f     3      0  2503     2
```

PBC data

Look at data frame containing time-dependent data.

```
head(pbcseq)
```

```
##   id futime status trt      age sex day ascites hepato spiders edema bili
## 1  1    400      2   1 58.76523   f   0       1      1       1     1 14.5
## 2  1    400      2   1 58.76523   f 192       1      1       1     1 21.3
## 3  2   5169      0   1 56.44627   f   0       0      1       1     0  1.1
## 4  2   5169      0   1 56.44627   f 182       0      1       1     0  0.8
## 5  2   5169      0   1 56.44627   f 365       0      1       1     0  1.0
## 6  2   5169      0   1 56.44627   f 768       0      1       1     0  1.9
##   chol albumin alk.phos   ast platelet protime stage
## 1  261    2.60     1718 138.0      190    12.2     4
## 2   NA    2.94     1612   6.2      183    11.2     4
## 3  302    4.14     7395 113.5      221    10.6     3
## 4   NA    3.60     2107 139.5      188    11.0     3
## 5   NA    3.55     1711 144.2      161    11.6     3
## 6   NA    3.92     1365 144.2      122    10.6     3
```

PBC data

Revise data frame to include multiple lines per subject with differing bilirubin values over different time intervals.

```
pbctdc <- tmerge(pbctdc, pbcseq, id=id, bili = tdc(day, bili))
head(pbctdc)
```

```
##   id time status trt      age sex stage tstart tstop death bili
## 1  1  400      2   1 58.76523   f     4      0   192     0 14.5
## 2  1  400      2   1 58.76523   f     4    192   400     2 21.3
## 3  2 4500      0   1 56.44627   f     3      0   182     0  1.1
## 4  2 4500      0   1 56.44627   f     3    182   365     0  0.8
## 5  2 4500      0   1 56.44627   f     3    365   768     0  1.0
## 6  2 4500      0   1 56.44627   f     3    768  1790     0  1.9
```

PBC data

```
model1 <- coxph(Surv(time, status==2) ~ log(bili), pbc)
model2 <- coxph(Surv(tstart, tstop, death ==2) ~ log(bili), pbctdc)
coef(summary(model1))
```

```
##                coef exp(coef)   se(coef)        z Pr(>|z|)
## log(bili) 0.9890831  2.688768 0.07835969 12.62235        0
```

```
coef(summary(model2))
```

```
##                coef exp(coef)   se(coef)      z Pr(>|z|)
## log(bili) 1.370255  3.936355 0.09499167 14.425        0
```

Your turn

Using the pbc and pbcseq data,

1. Make time-dependent versions of prothrombin time (protime), ascites and alkaline phosphaotase (alk.phos), and fit a model that includes them with time-dependent log(bilirubin).
2. Compare to the same model with baseline values.
3. Draw pictures of hazards or log hazards displaying subject trajectories for this model.
4. Create formulas and plot hazards or log hazards displaying subject trajectories when there is an interaction between a time-dependent binary variable (like heart-transplant status) and time or log time.