5. Estimating Measures of Association: Proportional Hazards Modeling

Patrick T. Bradshaw, Ph.D.

PUBHLTH 250C: Advanced Epidemiological Methods
School of Public Health
University of California, Berkeley

Table of Contents

Suggested References

Required/recommended: (see bCourses folder for previous lecture)

1. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. British J Cancer. 2003; 89:232-238.
2. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis–an introduction to concepts and methods. British J Cancer. 2003; 89:431-436.
3. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. British J Cancer. 2003; 89:605-611.
4. Vittinghoff E, Stephen S, and Glidden DV, McCulloch CE. *Regression methods in biostatistics, 2nd ed.* 2012. **(Chapters 3 and 6)**
   Book website:
   `http://www.epibiostat.ucsf.edu/biostat/vgsm/`

Suggested References

For more depth:

- Hosmer Jr, David W., Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time to event data.* Vol. 618. John Wiley & Sons, 2011.
  Examples:
  `http://www.ats.ucla.edu/stat/examples/`

Proportional Hazards Modeling

Motivation for regression modeling:

- Recall the relationships between the various functions describing event-time data:

$$
\begin{aligned}
h(t) &= F'(t)/S(t) = f(t)/S(t) \text{ (hazard function)} \\
H(t) &= \int_0^t h(u)du \text{ (cumulative hazard function)} \\
S(t) &= 1 - F(t) \text{ (survival function)} \\
&= \exp\left[-H(t)\right]
\end{aligned}
$$

- So modeling one of these functions is enough to describe the others.

Proportional Hazards Modeling

- Modeling the hazards (rate of occurrence of events) tends to be convenient.
  - Models look familiar (like logistic and linear regression).
- We assume that the hazards between 2 groups (e.g. a unit change in a covariate) are *proportional* to each other.
  - Let $h_2(t)$ be the hazard among those obese at baseline
  - Let $h_1(t)$ be the hazard among those normal weight at baseline.

Proportional Hazards Modeling

- Mathematically:

$$h_2(t)/h_1(t) = HR$$

- Describes how many *times* faster obese subject die compared to those of normal weight.
- This implies that the hazard ratios do not vary over time.
  - Will see later how to relax this assumption.

Proportional Hazards Models

- As a logistic regression model relates the log-odds of an outcome to a combination of covariates ($\mathbf{x}$), the proportional hazards model relates the log-*hazard* (rate of an event) to covariates:

$$\ln[h(t|\mathbf{x})] = \ln[h_0(t)] + x_1\beta_1 + \cdots + x_p\beta_p = \ln[h_0(t)] + \mathbf{x}\beta \quad (1)$$

  with $h_0(t)$ representing the *baseline hazard* (the rate of event at time t with all $\mathbf{x} = 0$).
  - $\ln[h_0(t)]$ is analogous to the intercept term in a generalized linear model.
  - "Baseline" hazard can be a confusing term–invokes thoughts of $t = 0$ (probably better referred to as "referent" hazard).

Proportional Hazards Models

- The variation in the (log-)hazard with respect to time is entirely accounted for by the baseline hazard.
  - *Proportional hazards assumption* (will discuss how to evaluate this later)
- Proportional hazards models are *multiplicative* models (like logistic regression).
- Changes in the covariates multiply the underlying (baseline) rate of the event occurring:

$$h(t|\mathbf{x}) = h_0(t) \times \exp(\mathbf{x}\beta)$$

- There are also *additive* hazards models. (not today)

Cox Proportional Hazards Models

- We've been somewhat vague about $h_0(t)$.
- Two options to deal with it:
  - Assume it has a specific shape and model it (e.g. Weibull, exponential, others…).
  - Work around it (Cox model).
  - We aren't often concerned with $h_0(t)$ (rarely interested in making inferences about baseline hazard)

Cox Proportional Hazards Models

- Note: Cox models are a specific *type* of proportional hazards model.
  - Refers to the model that doesn't specify the baseline hazard.
- If you have a good idea of the shape of $h_0(t)$, then sometimes better to use a parametric model for this.
  - There are options to flexibly model the baseline hazard in a parametric framework.*
  - Will cover some basic options later.

*Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statist. Med.* 2002; **21**:2175-2197.

Cox Proportional Hazards Models

- One problem with the Cox model: tied failure times problematic.
  - Occur because of coarse measurement or common event.
- Solution: use Efron method for ties.

Cox Proportional Hazards Models: Example

Back to our example:

- Want to estimate the association between BMI at baseline (study enrollment) with death.
- Proportional hazards model:

$$\ln\left[h(t|\mathbf{x})\right] = \ln[h_0(t)] + \texttt{underwt} \times \beta_1 + \texttt{overwt} \times \beta_2 + \texttt{obese} \times \beta_3$$

- Will use Cox model (not specifying $h_0(t)$).

Cox Proportional Hazards Models: Example

In R:

```
survmod.cox_unadj <- coxph(Surv(time_yrs,timedth_yrs, death)
                           ~ as.factor(bmi_cat), data=frmgham_recoded,
                           method="efron")
summary(survmod.cox_unadj)
```

The first argument to Surv function is optional if everyone starts at same time. (as here)

Cox Proportional Hazards Models: Example

```
  n= 4415, number of events= 1537
   (19 observations deleted due to missingness)

                        coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(bmi_cat)1 0.08643   1.09028  0.23939 0.361    0.718
as.factor(bmi_cat)3 0.27890   1.32168  0.05660 4.928 8.32e-07 ***
as.factor(bmi_cat)4 0.52156   1.68466  0.07494 6.960 3.41e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
as.factor(bmi_cat)1     1.090     0.9172     0.682     1.743
as.factor(bmi_cat)3     1.322     0.7566     1.183     1.477
as.factor(bmi_cat)4     1.685     0.5936     1.455     1.951

Concordance= 0.547  (se = 0.007 )
Rsquare= 0.012   (max possible= 0.997 )
Likelihood ratio test= 52.94  on 3 df,   p=1.892e-11
Wald test            = 53.81  on 3 df,   p=1.234e-11
Score (logrank) test = 54.58  on 3 df,   p=8.458e-12
```

- P-value for this crude model equivalent to logrank test.
- HR associated with obesity (compared to normal weight): 1.68.

Cox Proportional Hazards Models: Example

Adjusted for age (continuous) and sex (binary):

$$\ln\left[h(t|\mathbf{x})\right] = \ln\left[h_0(t)\right] + \texttt{underwt} \times \beta_1 + \texttt{overwt} \times \beta_2 + \texttt{obese} \times \beta_3 + \texttt{age} \times \beta_4 + \texttt{male} \times \beta_5$$

```
survmod.cox_adj <- coxph(Surv(time_yrs,timedth_yrs, death)
                           ~ as.factor(bmi_cat) + age + male,
                         data=frmgham_recoded, method="efron")
summary(survmod.cox_adj)
```

Cox Proportional Hazards Models: Example

(Partial output:)

```
                    exp(coef) exp(-coef) lower .95 upper .95
as.factor(bmi_cat)1      1.27      0.788     0.794      2.03
as.factor(bmi_cat)3      1.01      0.992     0.902      1.13
as.factor(bmi_cat)4      1.43      0.701     1.231      1.65
age                      1.10      0.912     1.090      1.10
male                     1.91      0.522     1.729      2.12

Concordance= 0.72  (se = 0.007 )
Rsquare= 0.207   (max possible= 0.997 )
Likelihood ratio test= 1025  on 5 df,   p=0
Wald test            = 956  on 5 df,   p=0
Score (logrank) test = 1043  on 5 df,   p=0
```

Cox Proportional Hazards Models: Example

Interpretations?

- *HR* obese vs. normal weight: 1.43?
- *HR* for age: 1.10?
- *HR* for male: 1.91?

Proportional Hazards Models: Assessing PH Assumption

- We've assumed that these effects are constant over the entire follow-up.
- What if the *effect* of baseline obesity changes over time?
  - $\beta = \beta(t)$
  - The parameter we have assumed so far doesn't provide full picture. (Interpreted as time-averaged effect)

Proportional Hazards Models: Assessing PH Assumption

How can you assess if this is a problem?

- Graphically (plot of $\ln\left[-\ln(S(t))\right]$ vs. $\ln(t)$).
- Analytically with an interaction between follow-up time and covariate.
- Analytically with an analysis of residuals. [Won't cover here–see references (and `cox.zph` command in `survival` package).]

Should technically assess this for **every** covariate in model (often folks focus on exposures only, but not great idea).

Graphical Assessment of PH Assumption

Log-log survival plots:

- Plot estimate of $\ln\left[-\ln(S(t))\right]$ vs. $\ln(t)$
- If proportional hazards assumption met, the $\ln\left[-\ln(S(t))\right]$ for different groups will be (fairly) parallel over $\ln(t)$.
  - Only applicable to categorical covariates (cumbersome, limited utility).
  - Visual inspection of plots often inconclusive unless gross violation.
- Sometimes presented as **negative** of the log-log survival (e.g. $-\ln\left[-\ln(S(t))\right]$ vs. $\ln(t)$) (same idea, just changes direction)
  - Stata does this.

Graphical Assessment of PH Assumption

Some intuition (maybe):
We know that

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\left[-H(t|\mathbf{x})\right] = \exp\left[-\int_0^t h(u|\mathbf{x})du\right] \\ &= \exp\left[-\exp(\mathbf{x}\beta)\int_0^t h_0(u)du\right] \end{aligned}$$

Taking the $\ln\left[-\ln(\cdot)\right]$ transformation of both sides:

$$\ln\left[-\ln(S(t|\mathbf{x}))\right] = \mathbf{x}\beta + \ln(S_0(t))$$

So, if the PH assumption holds, the difference between any 2 $\ln(-\ln(S(t|\mathbf{x})))$ curves is constant over time:

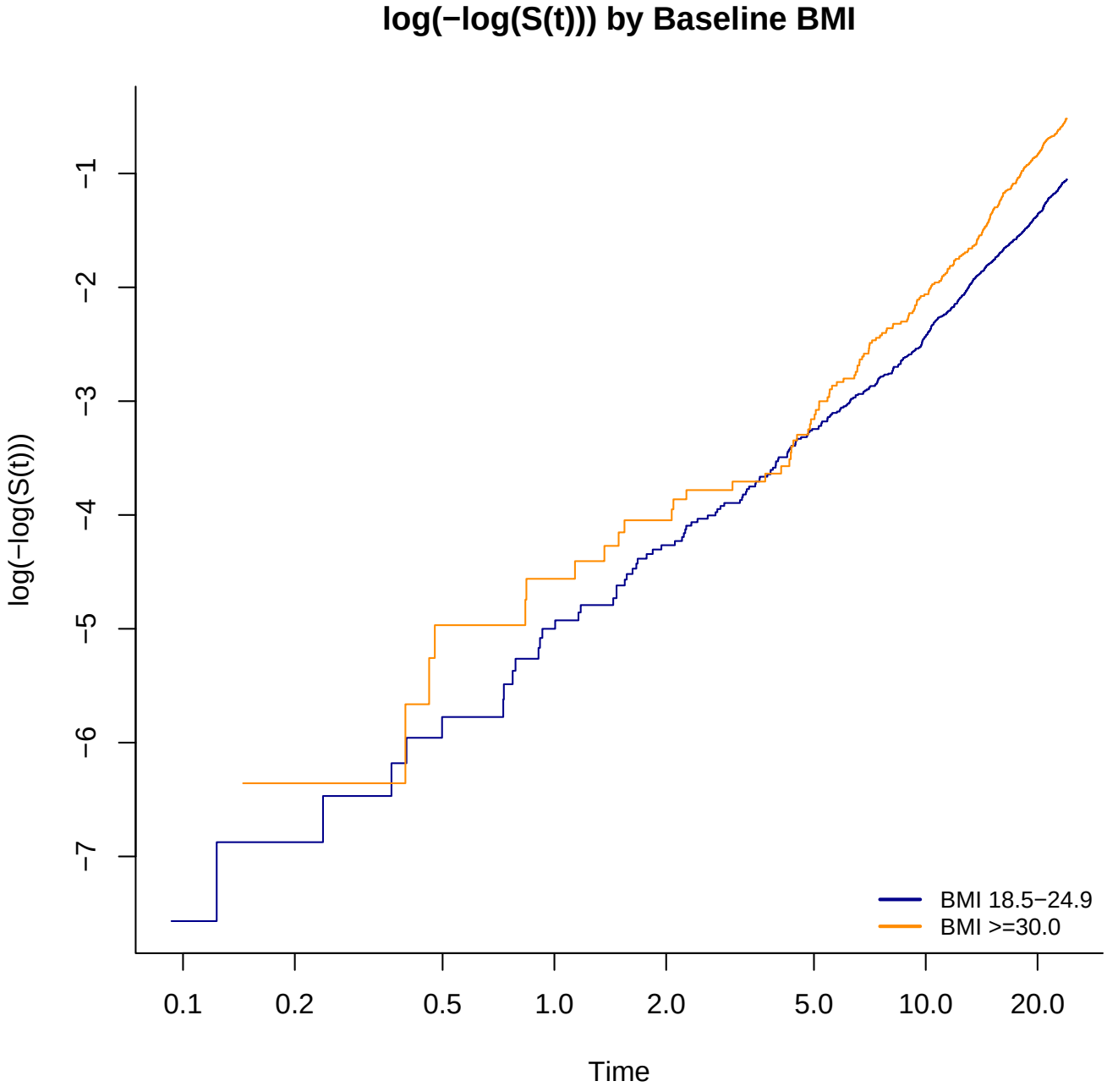
$$\ln\left[-\ln(S(t|\mathbf{x}=1))\right] - \ln\left[-\ln(S(t|\mathbf{x}=0))\right] = (1-0)\beta = \beta$$

Graphical Assessment of PH Assumption: Example

R code:

```
sf.cox_unadj <- survfit(Surv(time_yrs,timedth_yrs, death) ~
     as.factor(bmi_cat),
     data=frmgham_recoded, subset= bmi_cat==2 | bmi_cat==4)

plot(sf.cox_unadj, main="log(-log(S(t))) by Baseline BMI",
     xlab="Time", ylab="log(-log(S(t)))",
     fun=cloglog<-function(x) log(-log(x)), log="x",
     bty="l", col=c("darkblue","darkorange"))
legend("bottomright",legend=c("BMI 18.5-24.9","BMI >=30.0"),
       col=c("darkblue","darkorange"),
       cex=0.8, lty=1, lwd=2, bty="n")
```


log(−log(S(t))) by Baseline BMI
log(−log(S(t)))
−1
−2
−3
−4
−5
−6
−7
0.1
0.2
0.5
1.0
2.0
5.0
10.0
20.0
Time
BMI 18.5−24.9
BMI >=30.0

Graphical Assessment of PH Assumption: Example

Conclusion:

- Not parallel, but hard to tell how strong this deviation is.
- Red flag should go off if you see strong patterns:
  - Curves that start out wide apart, then converge.
  - Curves that cross each other (can happen multiple times).
  - Curves that start out near each other, then diverge.
- Covariate adjustment?

Analytical Assessment of PH Assumption: Time IXN

- Intuitive method to assess proportionality is to simply interact some function of follow-up time with each covariate.

$$\ln[h(t|\mathbf{x})] = \ln[h_0(t)] + \mathbf{x}\beta + (\mathbf{x} \times g(t))\gamma$$

  Note: No main effect for $g(t)$–already accounted for by $h_0(t)$.
- If $\gamma \neq 0$ then the effect of the corresponding covariate isn't constant over time (PH violation).

Analytical Assessment of PH Assumption: Time IXN

- For our adjusted model, need to include time interactions with BMI indicators, and (continuous) age, and male sex indicator.
- Common to assess interactions with linear time, or $\ln(\text{time})$.
  - Here will use linear time ($g(t) \equiv t$).
  - Good idea to check both (will just show linear today).
- Can use other transformations (categorized time–implies proportionality within regions).

Analytical Assessment of PH Assumption: Time IXN

- If using *single observation format* you must create the interactions *within* the `coxph` command in R (*DO NOT* create a new variable in the data set).
  - Stata: `tvc` option in `stcox` command. (similar to R)
  - SAS: create covariate-time interaction within the `proc phreg` step (not `data` step).

Analytical Assessment of PH Assumption: Time IXN

R requires you to use the `tt` ("time transform") option:

- Tells it which variables to interact with a function of time.
- Requires specification of function of time–very flexible.
  - Interaction with linear time:
    `tt=function(x,t,...) x*t`
  - Interaction with log time:
    `tt=function(x,t,...) x*log(t)`
  - Interaction with indicator of time $>$ 5 years:
    `tt=function(x,t,...) x*as.integer(t>5)`
- I don't think this works with `factor` (categorical) variables: need to code indicators. (Continuous variables OK.)

Analytical Assessment of PH Assumption: Time IXN

R code* (interaction with linear time):

```
survmod.cox_ixn <- coxph(Surv(time_yrs,timedth_yrs, death) ~
     underwt + overwt + obese + age + male +
          tt(underwt) + tt(overwt)+ tt(obese)+ tt(age) +tt(male),
     data=frmgham_recoded, method="efron",
          tt=function(x,t,...) x*t)
summary(survmod.cox_ixn)
```

*This can take a while. Go get a beverage, walk the dog, etc…

Analytical Assessment of PH Assumption: Time IXN

Results (part 1):

```
  n= 4415, number of events= 1537
   (19 observations deleted due to missingness)

                  coef  exp(coef)   se(coef)      z Pr(>|z|)
underwt       0.1215414  1.1292362  0.5667746  0.214 0.830201
overwt       -0.1651615  0.8477568  0.1376583 -1.200 0.230220
obese         0.0118410  1.0119114  0.1825159  0.065 0.948272
age           0.0812003  1.0845881  0.0076865 10.564  < 2e-16 ***
male          0.4644012  1.5910611  0.1258740  3.689 0.000225 ***
tt(underwt)   0.0087473  1.0087856  0.0376544  0.232 0.816302
tt(overwt)    0.0120500  1.0121229  0.0088620  1.360 0.173913
tt(obese)     0.0247721  1.0250815  0.0117705  2.105 0.035326 *
tt(age)       0.0007992  1.0007995  0.0004995  1.600 0.109589
tt(male)      0.0133733  1.0134631  0.0081093  1.649 0.099119 .
---
Signif. codes:  0 '***' 0.001 '**?'0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant interaction of obesity with time. (PH violated)

Analytical Assessment of PH Assumption: Time IXN

Results (part 2):

```
             exp(coef) exp(-coef) lower .95 upper .95
underwt         1.1292     0.8856    0.3718     3.429
overwt          0.8478     1.1796    0.6473     1.110
obese           1.0119     0.9882    0.7076     1.447
age             1.0846     0.9220    1.0684     1.101
male            1.5911     0.6285    1.2432     2.036
tt(underwt)     1.0088     0.9913    0.9370     1.086
tt(overwt)      1.0121     0.9880    0.9947     1.030
tt(obese)       1.0251     0.9755    1.0017     1.049
tt(age)         1.0008     0.9992    0.9998     1.002
tt(male)        1.0135     0.9867    0.9975     1.030

Concordance= 0.72  (se = 0.275 )
Rsquare= 0.209   (max possible= 0.997 )
Likelihood ratio test= 1035  on 10 df,   p=0
Wald test            = 964.6  on 10 df,   p=0
Score (logrank) test = 1054  on 10 df,   p=0
```

Analytical Assessment of PH Assumption: Time IXN

Time interactions very flexible way to assess PH violation.

- Easy to evaluate.
- Flexible (can accomodate continuous covariates, unlike $\ln(-\ln[S(t)])$ plots).

Cons:

- Sometimes not easy to identify relevant transformation of time.

Proportional Hazards Models: Relaxing PH Assumption

You have a violation of the PH assumption–what do you do now?

- If the violation is in your **exposure** of interest, use model with time interactions and report time-specific effects.*
- If the violation is in a **confounder**:
  - Include time interactions with confounder in the model (regardless of any time interactions with exposure).
  - If the confounder is categorical (or can be readily categorized), you could estimate a **stratified** Cox proportional hazards model.

*For more on this (and a slightly different approach), see Hernan "The hazards of hazard ratios." *Epidemiology*. 2010.

Proportional Hazards Models: Relaxing PH Assumption

- The **stratified Cox proportional hazards model** allows the baseline hazard (which you're not estimating) to vary according to groups defined by a covariate (say Z, with J categories).

$$\ln\left[h(t|\mathbf{x}, z)\right] = \ln[h_{0,j}(t)] + \mathbf{x}\beta \text{ for } z = j$$

where $j = 1, \ldots, J$.

Proportional Hazards Models: Relaxing PH Assumption

Notes on the stratified Cox model:

- *Stratifying* the baseline hazard function on a covariate precludes evaluation of its association with the outcome. (Only useful for a categorical confounder.)
- **Assumption:** The effect of the covariates (**x**) on the hazard are the same within strata defined by *z*.

Proportional Hazards Models: Relaxing PH Assumption

- Can be combined with time interactions for exposure variables (if you have violations in both).
- Let's assume we had a violation with regard to sex (not really indicated, but close).
- Our model then becomes:

$$\ln\left[h(t|\mathbf{x})\right] = \begin{cases} \ln[h_{0,0}(t)] + \mathbf{x}\beta + \mathbf{x} \times t\gamma & \text{if } \texttt{male} = 0 \\ \ln[h_{0,1}(t)] + \mathbf{x}\beta + \mathbf{x} \times t\gamma & \text{if } \texttt{male} = 1 \end{cases}$$

Different baseline hazard functions allow heterogeneity in sex-effect over time.

- $h_{0,0}(t)$ represents the baseline (referent) hazard among females.
- $h_{0,1}(t)$ represents the baseline (referent) hazard among males.

Proportional Hazards Models: Relaxing PH Assumption

$$\ln\left[h(t|\mathbf{x})\right] = \begin{cases} \ln[h_{0,0}(t)] + \mathbf{x}\beta + \mathbf{x} \times t\gamma & \text{if } \texttt{male} = 0 \\ \ln[h_{0,1}(t)] + \mathbf{x}\beta + \mathbf{x} \times t\gamma & \text{if } \texttt{male} = 1 \end{cases}$$

- **This model** assumes hazard ratio for 1-unit change in x at time t:

  $$HR(t) = \exp\left\{\beta + \gamma \times t\right\}$$

  common HR in all strata (effect of covariate doesn't differ by males/females).

Proportional Hazards Models: Relaxing PH Assumption

```
survmod.cox_ixn_male <- coxph(Surv(time_yrs,timedth_yrs, death) ~
                                    underwt + overwt + obese + age +
                                    strata(male) +
                                    tt(underwt) + tt(overwt)+ tt(obese),
                                data=frmgham_recoded, method="efron",
                                tt=function(x,t,...) x*t)
summary(survmod.cox_ixn_male)
```

Stratified by categorical sex variable (`male`).

Proportional Hazards Models: Relaxing PH Assumption

Results (Part 1): Baseline hazard stratified by sex.

```
  n= 4415, number of events= 1537
   (19 observations deleted due to missingness)

                coef exp(coef)  se(coef)      z Pr(>|z|)
underwt      0.123494  1.131443  0.567923  0.217   0.8279
overwt      -0.179521  0.835670  0.137228 -1.308   0.1908
obese       -0.006486  0.993535  0.182638 -0.036   0.9717
age          0.092450  1.096859  0.003229 28.632   <2e-16 ***
tt(underwt)  0.008180  1.008213  0.037745  0.217   0.8284
tt(overwt)   0.013153  1.013240  0.008828  1.490   0.1363
tt(obese)    0.025758  1.026092  0.011786  2.186   0.0289 *
---
Signif. codes:  0 '***' 0.001 '**?'0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Proportional Hazards Models: Relaxing PH Assumption

Results (Part 2): Baseline hazard stratified by sex.

```
             exp(coef) exp(-coef) lower .95 upper .95
underwt         1.1314     0.8838    0.3717     3.444
overwt          0.8357     1.1966    0.6386     1.094
obese           0.9935     1.0065    0.6946     1.421
age             1.0969     0.9117    1.0899     1.104
tt(underwt)     1.0082     0.9919    0.9363     1.086
tt(overwt)      1.0132     0.9869    0.9959     1.031
tt(obese)       1.0261     0.9746    1.0027     1.050

Concordance= 0.71  (se = 0.282 )
Rsquare= 0.189   (max possible= 0.994 )
Likelihood ratio test= 924.1  on 7 df,   p=0
Wald test            = 861.7  on 7 df,   p=0
Score (logrank) test = 951.7  on 7 df,   p=0
```

Proportional Hazards Models: Relaxing PH Assumption

- We've modeled BMI-time interactions $\Rightarrow$ the BMI-mortality association depends on time.
- Have to report BMI-mortality association at *specific timepoints* now.

Proportional Hazards Models: Relaxing PH Assumption

Association of obesity (vs. normal weight) at time t (holding $\mathbf{x}_{-\texttt{BMI}} = \textit{age}, \text{sex}$ constant):

$$HR_{\texttt{obese}}(t) = \frac{h(t|\texttt{obese}, \mathbf{x}_{-\texttt{BMI}})}{h(t|\texttt{normal weight}, \mathbf{x}_{-\texttt{BMI}})}$$

Log-transform:

$$\begin{aligned}\log[HR_{\texttt{obese}}(t)] = &\log\left[h(t|\texttt{obese}, \mathbf{x}_{-\texttt{BMI}})\right] - \\ &\log\left[h(t|\texttt{normal weight}, \mathbf{x}_{-\texttt{BMI}})\right]\end{aligned}$$

Proportional Hazards Models: Relaxing PH Assumption

Model:

$$\begin{aligned}\ln\left[h(t|\mathbf{x})\right] = \ln[h_{0,j}(t)] &+ \texttt{underweight} \times \beta_1 + \texttt{overweight} \times \beta_2 \\ &+ \texttt{obese} \times \beta_3 + \texttt{age} \times \beta_4 \\ &+ \texttt{underweight} \times t \times \gamma_1 + \texttt{overweight} \times t \times \gamma_2 \\ &+ \texttt{obese} \times t \times \gamma_3\end{aligned}$$

Proportional Hazards Models: Relaxing PH Assumption

Thus

$$\begin{aligned}\log\left[h(t|\texttt{obese}, \mathbf{x}_{-\texttt{BMI}})\right] = &\ln[h_{0,j}(t)] + 0 \times \beta_1 + 0 \times \beta_2 \\ &+ 1 \times \beta_3 + age \times \beta_4 \\ &+ 0 \times t \times \gamma_1 + 0 \times t \times \gamma_2 \\ &+ 1 \times t \times \gamma_3\end{aligned} \tag{2}$$

and

$$\begin{aligned}\log\left[h(t|\texttt{normal weight}, \mathbf{x}_{-\texttt{BMI}})\right] = &\ln[h_{0,j}(t)] + 0 \times \beta_1 + 0 \times \beta_2 \\ &+ 0 \times \beta_3 + age \times \beta_4 \\ &+ 0 \times t \times \gamma_1 + 0 \times t \times \gamma_2 \\ &+ 0 \times t \times \gamma_3\end{aligned} \tag{3}$$

Proportional Hazards Models: Relaxing PH Assumption

Subtracting equation (3) from (2) yields:

$$\begin{aligned}\log\left[HR_{\text{obese}}(t)\right] =& 0 \times \beta_1 + 0 \times \beta_2 + 1 \times \beta_3 + 0 \times \beta_4 \\ &+ 0 \times t \times \gamma_1 + 0 \times t \times \gamma_2 \\ &+ 1 \times t \times \gamma_3\end{aligned}$$

So we can get the $\log[HR_{\text{obese}}(t)]$ by multiplying each element of β by the corresponding elements of the vector:

$$\mathbf{k} = (0, 0, 1, 0, 0, 0, t)$$

and summing.

- Take anti-log of this quantity to get $HR_{\text{obese}}(t)$.

Proportional Hazards Models: Relaxing PH Assumption

- So when you multiply **k** and $(\beta' \ \gamma')'$ you get the linear combination:

$$\mathbf{k} \times \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \beta_3 + \gamma_3 \times t$$

R Code (example of association of obesity at t=5; requires `glht` function from `multcomp` package):

```
# Create vector to indicate linear contrast:
# Obese at t=5 (main effect + 5*obese(tt))
k5 <- matrix(c(0,0,1,0,0,0,5),1)

# Effect of obese at t=5
exp(confint(glht(survmod.cox_ixn_male,linfct=k5))$confint)[,1:3]
```

Proportional Hazards Models: Relaxing PH Assumption

Edited output:

```
# Association of obese at t=0:
> exp(confint(glht(survmod.cox_ixn_male,linfct=k0))$confint)[,1:3]
 Estimate       lwr       upr
0.9935351 0.6945786 1.4211668

# Association of obese at t=2:
> exp(confint(glht(survmod.cox_ixn_male,linfct=k2))$confint)[,1:3]
 Estimate       lwr       upr
1.0460584 0.7623093 1.4354253

# Association of obese at t=5:
> exp(confint(glht(survmod.cox_ixn_male,linfct=k5))$confint)[,1:3]
 Estimate       lwr       upr
1.1300950 0.8738601 1.4614635

# Association of obese at t=10:
> exp(confint(glht(survmod.cox_ixn_male,linfct=k10))$confint)[,1:3]
 Estimate      lwr      upr
1.285425 1.078585 1.531930
```

HR at year 0 (baseline) approximately 0.99 (null effect). HR at year 10 approximately 1.29 (significant effect). Baseline BMI takes a while to affect mortality.

Proportional Hazards Models: Relaxing PH Assumption

Alternative to continuous time: could consider effects proportional within regions of time:

- Separate effects for $t > 5$ years and $t <= 5$ years:

$$\ln\left[h(t|\mathbf{x})\right] = \ln[h_0(t)] + \mathbf{x}\beta + \mathbf{x} \times I_{(5,\infty)}(t)\gamma$$

  where $I_{(5,\infty)}(t) = 1$ if $t > 5$; 0 otherwise.
  - HR for a 1-unit change in $\mathbf{x}$ within the 1^{st} 5 years: $HR_{t<=5} = \exp(\beta)$.
  - HR for a 1-unit change in $\mathbf{x}$ after 5 years is $HR_{t>5} = \exp(\beta + \gamma)$.
- Assumes constant effect within each of these 2 regions.
  - Could re-assess interactions with continuous time within each.

Proportional Hazards Models: Relaxing PH Assumption

R code:

```
survmod.cox_ixn_t5_male <- coxph(Surv(time_yrs,timedth_yrs, death) ~
                                      underwt + overwt + obese + age
                                      + strata(male)
                                      + tt(underwt) + tt(overwt)+ tt(obese),
                                 data=frmgham_recoded, method="efron",
                                 tt=function(x,t,...) x*as.integer(t>5))
```

Proportional Hazards Models: Relaxing PH Assumption

Model output (last part):

```
             exp(coef) exp(-coef) lower .95 upper .95
underwt         1.0144     0.9858    0.2489     4.135
overwt          0.8069     1.2393    0.5830     1.117
obese           0.8825     1.1332    0.5564     1.400
age             1.0967     0.9118    1.0898     1.104
tt(underwt)     1.2917     0.7741    0.2909     5.736
tt(overwt)      1.2854     0.7780    0.9093     1.817
tt(obese)       1.7152     0.5830    1.0543     2.790

Concordance= 0.71  (se = 0.282 )
Rsquare= 0.189   (max possible= 0.994 )
Likelihood ratio test= 924.2  on 7 df,   p=0
Wald test            = 862.9  on 7 df,   p=0
Score (logrank) test = 952.3  on 7 df,   p=0
```

Proportional Hazards Models: Relaxing PH Assumption

Verify on your own the **k**-vector for obesity-mortality *HR* before year 5:

$$\mathbf{k}_{\leq 5\text{yrs}} = (0, 0, 1, 0, 0, 0, 0)$$

and after year 5:

$$\mathbf{k}_{>5\text{yrs}} = (0, 0, 1, 0, 0, 0, 1).$$

Q: What would it be for the effect of overweight before/after 5 years?

Proportional Hazards Models: Relaxing PH Assumption

R code:

```r
# Obese before 5 years (only main effect)
k.before5 <- matrix(c(0,0,1,0,0,0,0),1)

# Obese 5 years and after (main effect + 1*obese(tt))
k.after5 <- matrix(c(0,0,1,0,0,0,1),1)

# Effect of obese before 5 years
exp(confint(glht(
          survmod.cox_ixn_t6_male,linfct=k.before5))$confint)[,1:3]

# Effect of obese after 5 years
exp(confint(glht(
          survmod.cox_ixn_t6_male,linfct=k.after5))$confint)[,1:3]
```

Proportional Hazards Models: Relaxing PH Assumption

HR in years 1-5:

```
 Estimate       lwr       upr
0.8824530 0.5564278 1.3995047
```

HR in years 5+:

```
 Estimate      lwr      upr
1.513567 1.295240 1.768696
```

Proportional Hazards Models: Relaxing the PH Assumption

The specific form of the time interaction imposes an assumption on the model:

- Interaction with t assumes that the $HR(t) = e^{\beta} \times e^{\gamma t}$.
- Interaction with $\ln(t)$ assumes that the $HR(t) = e^{\beta} \times t^{\gamma}$.
- Interaction with categorized time assumes $HR(t)$ constant within regions of time.

Proportional Hazards Models: Time-varying Covariates

- So far, we have focused on covariates that are fixed over the follow-up period (e.g. BMI measured at baseline)
- Possible to extend the Cox model to allow for covariates to change over time.
    - BMI at each follow-up visit.
- We have actually already considered one time-varying covariate (follow-up time, interacted with exposure).

Proportional Hazards Models: Time-varying Covariates

- The time-varying covariate model expresses the hazard as a function of variables indexed by time:

$$\ln[h(t|\mathbf{x})] = \ln[h_0(t)] + x_1(t)\beta_1 + \cdots + x_p(t)\beta_p$$

  where the notation $x(t)$ emphasizes that the covariate values change over time.
- Some variables may not change over time:
  - For covariates that don't change over time, $x_i(t) \equiv x_i$ for all t

Time-varying Covariates: Example

randid	enter	exit	death2	bmi	age	sex
6238	0	2156	0	28.73	46	2
6238	2156	4344	0	29.43	52	2
6238	4344	8766	0	28.5	58	2
10552	0	1977	0	28.58	61	2
10552	1977	2956	1	30.18	67	2
11252	0	2072	0	23.1	46	2
⋮						

- Each observation (within subject) corresponds to a point at which the time-varying covariate is assessed. (Here, corresponding to follow-up visits).
- For this, data needs to be in multiple-record format.
- See posted R code for how to reshape Framingham data into this form. (much easier in Stata–see `stsplit`)

Proportional Hazards Model: Time-varying Covariates

R code:

```
survmod.cox_unadj2 <- coxph(Surv(enter,exit, death2) ~ as.factor(bmi0_cat),
                            data=frmgham_recoded, method="efron")
```

death2: period-specific death indicator.

Proportional Hazards Models: Relaxing PH Assumption

Results (edited):

```
  n= 11591, number of events= 1537
   (36 observations deleted due to missingness)

                     exp(coef) exp(-coef) lower .95 upper .95
as.factor(bmi0_cat)1     1.090     0.9172     0.682     1.743
as.factor(bmi0_cat)3     1.322     0.7566     1.183     1.477
as.factor(bmi0_cat)4     1.685     0.5936     1.455     1.951

Concordance= 0.547  (se = 0.007 )
Rsquare= 0.005   (max possible= 0.886 )
Likelihood ratio test= 52.94  on 3 df,   p=1.892e-11
Wald test            = 53.81  on 3 df,   p=1.234e-11
Score (logrank) test = 54.58  on 3 df,   p=8.458e-12
```

Same as single-observation model. (But you can tell this is different since it has more observations, but same number of subjects, failures, time at risk)

Proportional Hazards Models: Relaxing PH Assumption

Now using time-varying BMI and age, and fixed sex:

```
survmod.cox_unadj2 <-  coxph(Surv(enter,exit, death2) ~ as.factor(bmi_cat),
                             data=frmgham_recoded, method="efron")
```

Proportional Hazards Models: Relaxing PH Assumption

Results (edited):

```
                    exp(coef) exp(-coef) lower .95 upper .95
as.factor(bmi_cat)1     2.038     0.4906     1.469     2.828
as.factor(bmi_cat)3     1.081     0.9249     0.968     1.208
as.factor(bmi_cat)4     1.173     0.8527     1.009     1.364

Concordance= 0.519  (se = 0.007 )
Rsquare= 0.002   (max possible= 0.885 )
Likelihood ratio test= 17.74  on 3 df,   p=0.0004967
Wald test            = 20.57  on 3 df,   p=0.000129
Score (logrank) test = 21.18  on 3 df,   p=9.665e-05
```

Association changes!

Proportional Hazards Model: Time-varying Covariates

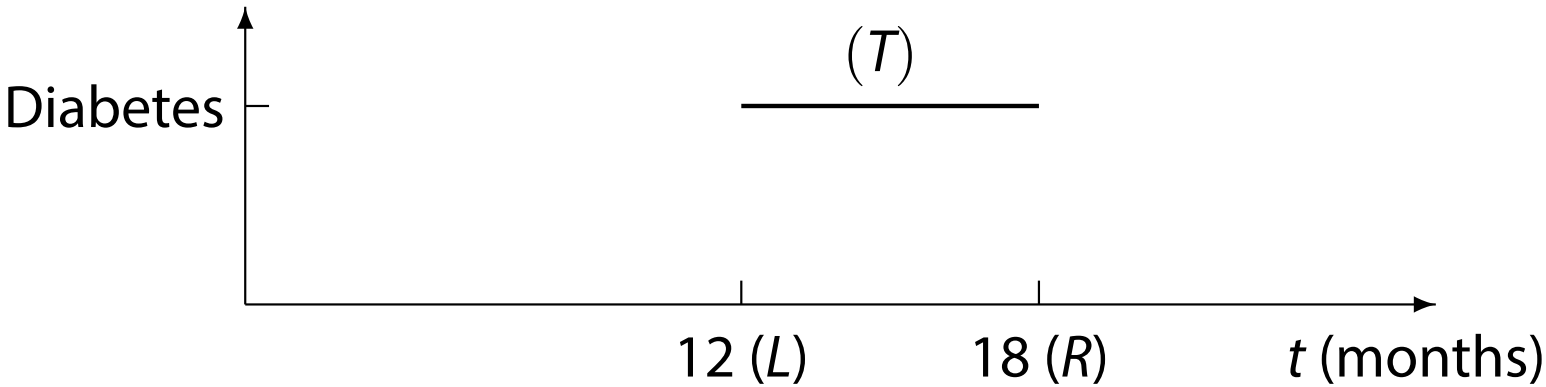
- Good idea to also verify no time interaction with a TVC.
  - Just because the model allows for the covariate to change doesn't mean you've correctly modeled the temporal relationships.
- Pay attention to time-varying confounders that are consequences of exposure.
  - Draw out your DAG.
  - If necessary, consider IP-weighting [Robins et al. (*Epidemiology* 2000), Hernan et al. (*Epidemiology* 2000)].
  - May need to use discrete time methods (historically, software doesn't handle weighting well in the Cox model with TVC).

Interval Censoring

- Another application of generalized linear models: estimating ratio measures of effect (hazard ratios) for time-to-event data.
- Often for a time-to-event analysis we know (T, D) where T is the event/censoring time, and D is an indicator for the event.
- Problem: we may not always know the exact time an event occurred.

Interval Censoring

- We might only know the *interval of time* during which the outcome occurred: $T \in [L, R]$.
  - For example, a subject developed diabetes between months 12 and 18 of our study.



- *Q: What scenarios would lead to this situation? Do you think these are common in epidemiologic research?*

Interval Censoring

- Common methods to analyze interval censored data:
  - Assume the event happened at the beginning of the interval $T = L$.
  - Assume the event happened at the end of the interval $T = R$.
  - Assume the event happened at the midpoint of the interval $T = \frac{L+R}{2}$.
- None of these are a great idea.

Types of Interval Censoring: Synchronous

This type of data is called *interval censored*.

- **Synchronous interval censoring**: all subjects measured at the same intervals.
  - E.g. Subjects are evaluated at baseline (0 months), 3 months, 6 months, 12 months and 18 months.
  - Don't need to be evenly spaced, just that measurement times are common to all individuals.
  - Also referred to as *grouped survival data*.
  - Tend to be more common for epidemiologic/clinical studies.
  - Data that doesn't strictly adhere to synchronous censoring scheme can often be treated as such (e.g. if actual exam dates are within a small interval).
  - More straightforward to handle.

Types of Interval Censoring: Asynchronous

- **Asynchronous interval censoring**: subjects measured at different intervals.
  - E.g. Subject 1 measured at 0, 3, 9 and 18 months; subject 2 measured at 0, 4, 6, 8 and 12 months.
  - This structure is more challenging to deal with.
  - Can occur in epidemiologic/clincal studies.
    - Even though follow-up visits are often *scheduled* for regular intervals, an exact schedule is rarely achieved.

Interval Censoring

- The objective is to model the probability of death, given that the individual survived up to the given period (j):

$$P(Y_j = 1 | \mathbf{y}_{j-1} = 0) = h_j$$

which we call the *hazard*.

Definition

The **discrete hazard** is the conditional probability of experiencing the event within interval *j given* that the subject did not experience it in any other period.

Complementary Log-Log Model

The *complementary log-log* model can be used to model synchronous interval censored data*.

- Divide your follow-up times into a series of J intervals.
  - E.g. If you have 10 years of follow-up, could divide into 5 intervals, each of 2 years in length.
- Model the complementary log-log transformation of hazard:

$$\log(-\log(1 - h(t_j)) = \alpha_j + \mathbf{x}\beta \quad j \in (1, \ldots, J)$$

α_j captures the effect of the baseline hazard h_{0j}. Note that the baseline hazard within each interval is constant ($h_{0.}$) only changes where 2 intervals meet).

*See Hosmer and Lemeshow, *Applied Survival Analysis* (Ch 7). Vittinghoff et al. 2012, Chapter 5.

Complementary Log-Log Model

- Essentially model the probability of an event happening within a given interval.*
  - Y $\sim$ Bernoulli (i.e. Binomial with $n = 1$).
  - $P[Y_j = 1 | \mathbf{y}_{j-1} = 0, \mathbf{x}] = h_j$
- $\exp(\beta)$ from this model is equivalent to hazard ratios (HR) from proportional hazards model.
- Benefit of these types of models allows some flexibility in modeling the baseline hazard (increase number of intervals, use splines, etc…).

*Hosmer and Lemeshow. *Applied Survival Analysis*, (Ch. 7)
SAS and Stata code available at http://www.ats.ucla.edu/stat/Stata/examples/asa/.

Complementary Log-Log Model

Model specification:

- Generalized linear model.
  - Binomial distribution for outcome variable (event/non-event).
  - Complementary log-log link.
- Common intervals for everyone (although don't need to be equally spaced).
- Side benefit: easy to incorporate time-varying covariates.
  - Can't allow for exposures to change within an interval.

Complementary Log-Log Model

- Data structure: Must have an observation in the dataset for each subject *at each interval* they are still at risk.
- Example:
  Data from Western Collaborative Group Study (Rosenman et al. JAMA 1964)*.
  - Large epidemiologic study of coronary heart disease (included $>$ 3500 men, age 39-59)
  - **Outcome**: indicator of CHD (over 10 yrs f/u), and time of event
  - **Predictors**: BMI (WHO categories), age (continuous)

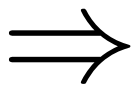
*Data from Vittinghoff et al. *Regression Methods in Biostatistics.* 2012

Complementary Log-Log Model

- The outcome variable (chd) indicates if the subject was diagnosed at that visit (=1), or was disease free/censored (=0).
- Decide on number of intervals, and width.
- Expand the dataset so each subject has one entry for each interval (multiple records per subject)
  A subject who developed CHD at the 8 year assessment (id=455) would have 4 entries in the dataset:
  1. 0-2 years
  2. 2-4 years
  3. 4-6 years
  4. 6-8 years

Complementary Log-Log Model

id2	chd69	inter	age	bmi
218	no	10	41	22.59412
455	yes	8	53	27.36592
1752	no	8	46	22.73934



id2	chd	year2	intnew	age	bmi
218	0	2	1	41	22.59412
218	0	4	2	41	22.59412
218	0	6	3	41	22.59412
218	0	8	4	41	22.59412
218	0	10	5	41	22.59412
455	0	2	1	53	27.36592
455	0	4	2	53	27.36592
455	0	6	3	53	27.36592
455	1	8	4	53	27.36592
1752	0	2	1	46	22.73934
1752	0	4	2	46	22.73934
1752	0	6	3	46	22.73934
1752	0	8	4	46	22.73934

Complementary Log-Log Model

- We then model the probability of the event (diabetes) happening within interval j (denoted θ_j):

$$\ln\left[-\ln\left(1-\theta_j\right)\right] = \alpha_j + \mathbf{x}\beta$$

  where α_j is the baseline effect for interval j (intercept terms). The β represent the log-hazard ratios for this model.
- Generalized linear model: **binomial** distribution, **complementary log-log** link function.
  - Not necessary, but will estimate all interval-specific parameters (α_j)–remove default intercept from model.

Complementary Log-Log Model

See posted R code for data recoding:

```
data.wcgs.long$bmi.cat <- cut(data.wcgs.long$bmi,
     c(0,18.5,25,30,1000), right=FALSE)

data.wcgs.long$bmi.cat <- relevel(as.factor(data.wcgs.long$bmi.cat),
     "[18.5,25)")

# Remove obs w/ underweight BMI
data.wcgs.long.nouw <-
     data.wcgs.long[data.wcgs.long$bmi.cat != "[0,18.5)",]

wcgs.cloglog.fit <- glm(chd ~ factor(intnew) + factor(bmi.cat) + age -1,
                        data=data.wcgs.long.nouw,  family=binomial(cloglog))
hr.wcgs<-exp(cbind(coef(wcgs.cloglog.fit), confint(wcgs.cloglog.fit)))
colnames(hr.wcgs)<-c("HR","95% LL","95% UL")
hr.wcgs.expci <- round(hr.wcgs,digits=4)
hr.wcgs.expci[6:7,]
```

Complementary Log-Log Model

```
> hr.wcgs.expci[6:7,]
                              HR 95% LL 95% UL
factor(bmi.cat)[25,30)    1.24  0.963   1.59
factor(bmi.cat)[30,1e+03) 2.22  1.162   3.83
```

Delayed Entry/Left Truncation

A few words on left truncation:

- **Left truncation** (delayed entry): when we don't observe everyone from the time origin.
  - Here we assumed everyone was observed from $t = 0$, but is that correct origin?
  - If you consider age as appropriate time scale, then maybe risk periods were different lengths.
- Left truncation can cause you to miss early events.
- Can bias hazard ratios from regression models.
- Especially problematic if delayed entry associated with exposure.

Delayed Entry/Left Truncation

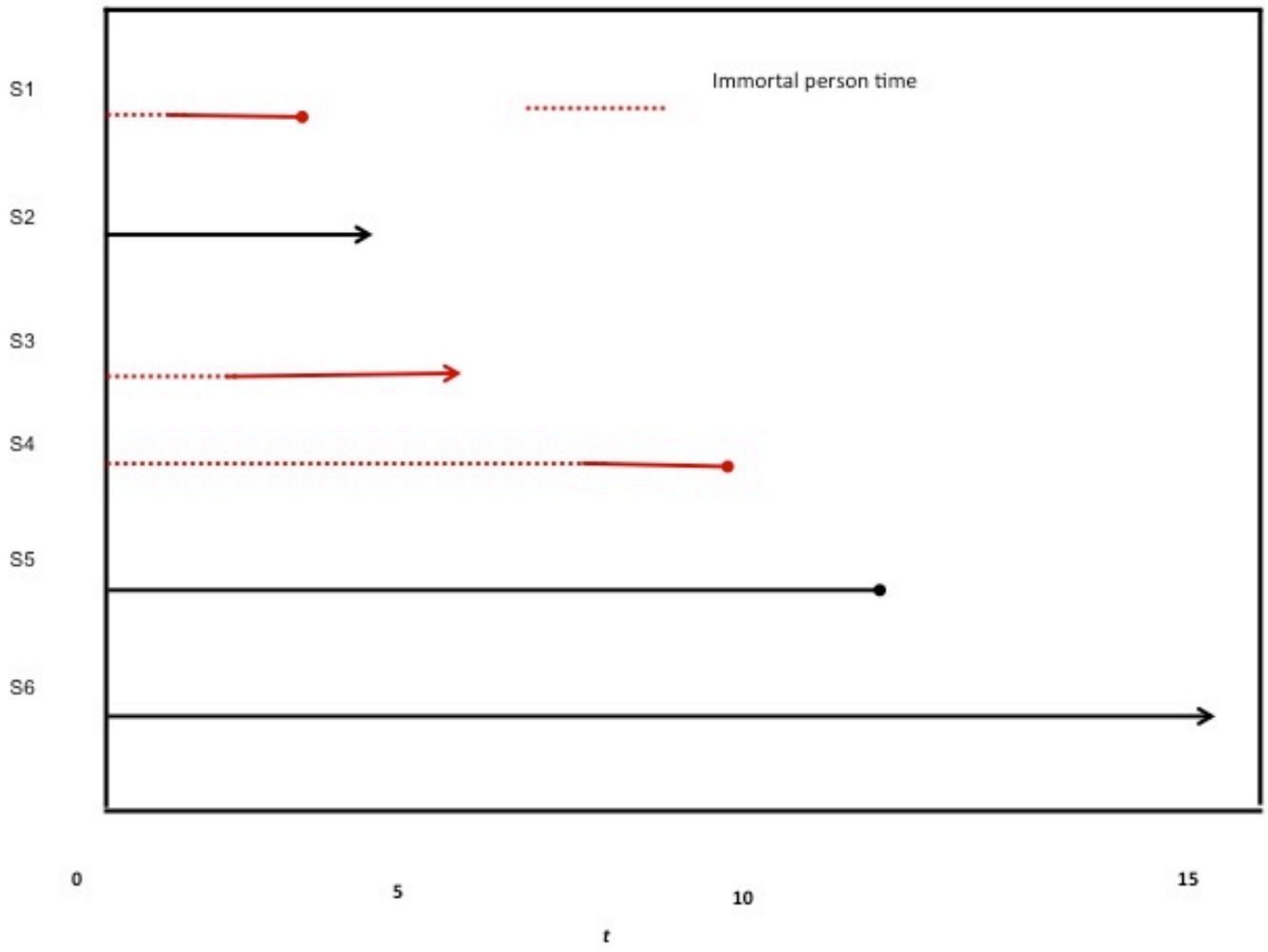
- Common in studies of cancer survivorship (and other substantive areas):
  - Subjects enrolled at varying points after diagnosis: some subjects enrolled at 2 years, some at 4 years.
  - Time between diagnosis and enrollment is *immortal person-time*.
  - Some people redefine the research question: "Survival from study entry until end of follow-up."
  - *As meaningful as survival from date of diagnosis?*

Delayed Entry/Left Truncation

- If assumption of *independent truncation* is reasonable, then analysis is straightforward.
  - **Independent truncation:** timing of entry into study is independent of subsequent survival.
  - Everyone (no matter how soon they died/had event) had a chance at being in the study.
  - Need to only count their time under observation.
- Data structure should account for time of study entry relative to meaningful origin.

Delayed Entry/Left Truncation

Consider 6 subjects, some who enter study after time origin.



Delayed Entry/Left Truncation

Each observation should include entry time (`t0`) and exit time (`t1`):

ID	Exposed	Enter (t0)	Exit (t1)	Event
S1	0	1	3	1
S2	1	0	4	0
S3	1	2	5	0
S4	0	7	9	1
S5	0	0	11	1
S6	1	0	15	0

Delayed Entry/Left Truncation

- Under independent truncation, analysis proceeds using same code (`coxph` with (entry, exit) format).
- In Stata, need to declare entry/exit times in `stset` command–see manual for examples.
- Data does not have to (but can) have multiple records per subject (accomodates TVC).

5. Estimating Measures of Association: Proportional Hazards Modeling

Patrick T. Bradshaw, Ph.D.

PUBHLTH 250C: Advanced Epidemiological Methods
School of Public Health
University of California, Berkeley