Review and motivation

In the last lecture we learned about the survival function

$$S(t) = \text{Prob}(T > t)$$

and approximated it using the Kaplan-Meier estimate, $\widehat{S}_{KM}(t)$. (Recall that T is the random variable for the failure time and t is its observed value.) The survival function expresses the probability of surviving (not having an event) at least until time t. Although estimation of and inference pertaining to survival functions is a well-established approach to the analysis of time to event data, we want a method that allows us to model survival time (or some transformation of survival time) as a function of covariates.

The hazard function

This goal is facilitated using a function called the *hazard function*, $h(t)$. The hazard function is the *risk of failure at time* t*, given survival up to the time just before time* t*.* $h(t)$ *is the instantaneous failure rate for an individual surviving to time* t*.* It is sometimes referred to as the *intensity rate* or the *force of mortality*. It is often interpreted as an instantaneous risk of failure. The hazard function is used to answer the question (for example),

"Given that an HIV+ subject has not died of AIDS or AIDS related complications by the time they've reached five years post seroconversion, what is the probability that the subject will die at five years?".

The hazard function (cont.)

The definition of the hazard function and its relationship to the survival function are as follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left[\frac{\text{Prob}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right] \\ &= \frac{\lim_{\Delta t \to 0} \left[\frac{\text{Prob}(t \leq T \leq t + \Delta t)}{\Delta t} \right]}{\text{Prob}(T \geq t)} \quad \left(P(A|B) = P(A \cap B)/P(B) \right) \\ &= \frac{-dS(t)/dt}{S(T)} \quad \left(\text{def. of derivative and def. of } S(t) \right) \\ &= \frac{d}{dt}\{-\ln S(t)\} \quad \left(d(\ln u) = du/u \right) \end{aligned}$$

The cumulative hazard function

The cumulative hazard function, $H(t)$, is defined as

$$\begin{aligned} H(t) &= \int_0^t h(u)du \\ &= \int_0^t \frac{d}{du}\{-\ln S(u)\}du \\ &= -\ln S(t) \quad \left(\text{by the fundamental theorem of Calculus} \right) \end{aligned}$$

$H(t)$ is a measure of the "accumulated" risk of failure given survival to time t.

Key relationships

The most important concept to take home from all of this math is that there is a known relationship between the survival function and the hazard function. For the purposes of our class, we'll let the computer convert from one to the other (and vice versa) but it is important to know that one is simply a function of the other.

Key relationships (cont.)

The essential definitions and functional relationships are

1. $S(t) = \text{Prob}(T > t)$
2. $h(t) = \lim_{\Delta t \to 0} \left[\frac{\text{Prob}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \right]$
3. $H(t) = \int_0^t h(u) du$
4. $H(t) = -\ln S(t)$
5. $S(t) = \exp\{-\int_0^t h(u) du\} = \exp\{-H(t)\}$

Proportional hazards revisited

Recall in our first lecture that we said that the log-rank test (used to test the equivalence of two survival functions) is most powerful for the alternative

$$S_1(t) = [S_2(t)]^c, \ c \neq 1.$$

We said that this assumption was called the *proportional hazards assumption*. We now have the tools to demonstrate where this name comes from.

Proportional hazards revisited (cont.)

As a quick review, recall that two quantities are *proportional* if their ratio is a constant. That is, X is proportional to Y (and vice versa) if $X/Y = c$, where c is some constant.

$$\begin{aligned}
& S_1(t) = [S_2(t)]^c \\
\iff \quad & \ln S_1(t) = c \ln S_2(t) \\
\iff \quad & -\ln S_1(t) = c(-\ln S_2(t)) \\
\iff \quad & H_1(t) = cH_2(t) \\
\iff \quad & \frac{H_1(t)}{H_2(t)} = c
\end{aligned}$$

Proportional hazards revisited (cont.)

Therefore, when we plot $\ln(-\ln S_1(t))$ and $\ln(-\ln S_2(t))$ on the same set of axes, we are actually plotting $\ln(H_1(t))$ and $\ln(H_2(t))$. It follows from properties of logarithms that if $H_1(t)$ and $H_2(t)$ are proportional to one another, then their logarithms should differ by a constant.

Regression models for survival data

We approach the problem of modelling survival time via the hazard function. We impose a regression model-type structure on the hazard function that is the product of two components. One factor captures the effect of survival time on the hazard and the second expresses the effects of covariates associated with survival, such as age, race, sex, etc.. The form of the model is

$$h(t|X_1, \ldots, X_k) = h_0(t)e^{\beta_1 X_1 + \ldots + \beta_k X_k}.$$

Regression models for survival data (cont.)

$$h(t|X_1, \ldots, X_k) = h_0(t)e^{\beta_1 X_1 + \ldots + \beta_k X_k}$$

- $h_0(t)$ is called the *baseline hazard*. It characterizes how the hazard function changes as a function of survival time.
- $e^{\beta_1 X_1 + \ldots + \beta_k X_k}$ characterizes how the hazard function changes as a function of covariates.
- $h(t)$ is referred to as the *Cox model* or *Cox proportional hazards model* or simply the *proportional hazards model*.
- $h(t)$ is linear in the covariates on the log scale. That is, $\ln h(t) = \ln h_0(t) + (\beta_1 X_1 + \ldots + \beta_k X_k)$

Demonstrating "proportional hazards"

The proportional hazards assumption implies that the ratio of the instantaneous failure rates for two subjects is a constant. To see why this assumption is implicit in the form of the model, consider two subjects, A and B with covariates $\mathbf{X}_A$ and $\mathbf{X}_B$, respectively. Then

- $h(t|\mathbf{X}_A) = h_0(t)e^{\mathbf{X}'_A\beta}$
- $h(t|\mathbf{X}_B) = h_0(t)e^{\mathbf{X}'_B\beta}$

so that

$$\frac{h(t|\mathbf{X}_A)}{h(t|\mathbf{X}_B)} = \frac{h_0(t)e^{\mathbf{X}'_A\beta}}{h_0(t)e^{\mathbf{X}'_B\beta}} = \frac{e^{\mathbf{X}'_A\beta}}{e^{\mathbf{X}'_B\beta}}$$

Since $\frac{e^{\mathbf{X}'_A\beta}}{e^{\mathbf{X}'_B\beta}}$ is just a constant, the hazards for the two subjects are proportional to one another. Notice that the ratio of their hazards does not depend on time. The proportional hazards assumption means we assume that the ratio of the hazards is constant over time.

HIV example

Recall the HIV data presented in the last lecture.

ID Subject ID

TIME Survival time (months)

AGE Age (years) of subject at time of enrollment

DRUG Use of prior injecting drug use (1 = Yes, 0 = No)

CENSOR Censoring indicator (1 = Death observed, 0 = censored)

An additional variable, **RACE**, has been added for illustrative purposes only. It is coded 1 = African American, 2 = Other, 3 = White.

Proportional hazards model in SAS

Using the HIV data presented in the last lecture, fit the model

$$h(t|\text{DRUG}) = h_0(t)e^{\beta_{\text{DRUG}}\text{DRUG}}.$$

```
proc phreg data = one;
    model time*censor(0) = drug;
run;
```

```
                                      Percent
   Total       Event    Censored    Censored
     100          80          20       20.00

              Convergence Status
    Convergence criterion (GCONV=1E-8) satisfied.
```

Proportional hazards model in SAS (cont.)

```
         Model Fit Statistics
                  Without           With
Criterion     Covariates     Covariates
-2 LOG L         598.390        588.193
AIC              598.390        590.193
SBC              598.390        592.575

        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio        10.1973        1         0.0014
Score                   10.7432        1         0.0010
Wald                    10.3451        1         0.0013

            Analysis of Maximum Likelihood Estimates

             Parameter   Standard                           Hazard
Variable DF   Estimate      Error Chi-Square Pr > ChiSq    Ratio
drug      1    0.77919    0.24226    10.3451     0.0013    2.180
```


Interpretation of output

- The overall test of fit for the model is based on the following null and alternative hypotheses

$$H_0: \quad \beta_{\text{DRUG}} = 0$$
$$H_A: \quad \beta_{\text{DRUG}} \neq 0$$

- The overall test of model fit is tested via the likelihood ratio test, namely

$$[-2\ln L(\text{reduced})] - [-2\ln L(\text{full})] \sim \chi^2_d$$

where d is the number of variables being tested.

Interpretation of output (cont.)

- In this example, the full model is the model containing DRUG and the baseline hazard, and the reduced model contains only the baseline hazard. Therefore, under the null hypothesis (above) the likelihood ratio test statistic is distributed χ^2_1.
- The likelihood ratio test is highly significant ($p = 0.0014$) so we reject H_0 and conclude that DRUG contributes significantly to the model containing only the baseline hazard.

Interpretation of output (cont.)

- The chi-square tests for the MLEs test the null hypothesis that the corresponding β equals 0 given that all other covariates are in the model.
- In this example, there are no other covariates in the model besides DRUG. Therefore the test on the MLE for DRUG is asymptotically equivalent to the likelihood ratio test (LR test is the overall test for the model).

Obtaining hazard ratios

The fitted model is

$$h(t|\text{DRUG}) = h_0(t)e^{0.78}.$$

We can use the fitted model to obtain estimates of hazard ratios. Specifically, suppose we want to compare the hazards of failure for subjects with and without an injecting drug use history.

- For the subject with an injecting drug use history, $h(t|\text{DRUG} = 1) = h_0(t)e^{0.78\times 1}$.
- For the subject without an injecting drug use history, $h(t|\text{DRUG} = 0) = h_0(t)e^{0.78\times 0} = h_0(t)$.
- The hazard ratio is

$$\frac{h(t|\text{DRUG} = 1)}{h(t|\text{DRUG} = 0)} = \frac{h_0(t)e^{0.78\times 1}}{h_0(t)e^{0.78\times 0}} = e^{0.78\times 1} = 2.18.$$

Hazard ratios for categorical predictors

In general, let X be a k-level categorical variable. Let $Z_1, Z_2, \ldots, Z_{k-1}$ be the $k-1$ dummy variables associated with X. Assume β_j is the regression coefficient of Z_j obtained from a Cox-PH regression model, where $j = 1, \ldots, k-1$. Then e^{β_j} is the hazard ratio comparing the jth level of X to the reference level.

Confidence intervals for hazard ratios

A 95% confidence interval for the HR is simply

$$e^{\hat{\beta} \pm 1.96\widehat{SE}(\hat{\beta})}.$$

Therefore, from the output on Slide 15, a 95% CI for the HR of death for those with a history of IDU relative to those without a history of IDU, is

$$e^{0.77919 \pm 1.96 \times 0.24226} = (1.36, 3.50).$$

Because the CI does not contain the null value of 1, we conclude that the difference in risk between HIV+ IDUs and HIV+ non-IDUs is significant.

Confidence intervals for RRs in SAS

```
proc phreg data = one;
    model time*censor(0) = drug/rl;
run;
```

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits	
drug	2.180	1.356	3.504

Multivariable models

We now fit the model

$$h(t|\text{DRUG, AGE}) = h_0(t)e^{\beta_{\text{DRUG}}\text{DRUG}+\beta_{\text{AGE}}\text{AGE}}.$$

```
proc phreg data = one;
    model time*censor(0) = drug age/rl;
run;
```


Multivariable models (cont.)

```
         Model Fit Statistics

                  Without           With
Criterion      Covariates     Covariates
-2 LOG L          598.390        563.408

        Testing Global Null Hypothesis: BETA=0
Test                  Chi-Square       DF     Pr > ChiSq
Likelihood Ratio         34.9819        2         <.0001
```

Multivariable models (cont.)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
drug	1	0.94108	0.25550	13.5662	0.0002
age	1	0.09151	0.01849	24.5009	<.0001

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits	
drug	2.563	1.553	4.229
age	1.096	1.057	1.136

Interpreting default HRs and CIs

- The hazard of death among HIV+ subjects with a history of IDU is 2.6 times that of subjects with no history of IDU of the same age. (Can also say '... controlling for age.')
- There is a 10% increase in the hazard of death for every year increase in age for HIV+ subjects with the same history of IDU. (Can also say '... controlling for history of IDU.')

Other HRs and CIs

To compute the relative risk and corresponding 95% CI comparing subjects with values for a continuous covariate that differ by a fixed amount, say Δx, we use the following principle:

$$\widehat{\mathsf{HR}} = e^{\Delta x \hat{\beta}}$$

and the 95% CI is

$$(e^{\Delta x \hat{\beta} - 1.96 \times |\Delta x| \times \mathsf{SE}(\hat{\beta})}, e^{\Delta x \hat{\beta} + 1.96 \times |\Delta x| \times \mathsf{SE}(\hat{\beta})}).$$

Other HRs and CIs (cont.)

For example, to compute the HR for subjects who differ in age by 10 years (or some other meaningful time period of interest) with the same history of IDU,

$$\widehat{HR} = e^{0.09151 \times 10} \doteq 2.50.$$

The corresponding 95% CI is

$$e^{0.09151 \times 10 \,\pm\, 1.96 \times 10 \times 0.01849} = (1.74, 3.59).$$

Handling categorical variables in PROC PHREG

There is no "class" statement in PROC PHREG that allows us to conveniently handle categorical variables. However, PROC PHREG allows us to program the dummy variables 'on the fly'.

Suppose I wish to consider the model containing DRUG, AGE and RACE as covariates. It would be inappropriate to enter the variable RACE as is into the model (*Why?*). Rather, we create dummy variables within the procedure.

```
proc phreg data = two;
    model time*censor(0) = drug age race1 race2/rl;
    if race = 1 then race1 = 1; else race1 = 0;
    if race = 2 then race2 = 1; else race2 = 0;
run;
```


Handling categorical variables in PROC PHREG (cont.)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
drug	1	0.96242	0.25717	14.0055	0.0002
age	1	0.09567	0.01876	26.0054	<.0001
race1	1	0.27026	0.36474	0.5490	0.4587
race2	1	0.67565	0.36678	3.3934	0.0655

Analysis of Maximum Likelihood Estimates

Variable	Hazard Ratio	95% Hazard Ratio Confidence Limits	
drug	2.618	1.582	4.334
age	1.100	1.061	1.142
race1	1.310	0.641	2.678
race2	1.965	0.958	4.033

Testing the significance of categorical variables

To test the significance of the variable RACE in the model, it would be inappropriate to use the significance tests on the individual dummy variables. Rather, you should test the significance of the contribution of the *collection* of dummy variables that represent the effect of RACE on the hazard of death. This is accomplished in PROC PHREG using a TEST statement.

```
proc phreg data = two;
    model time*censor(0) = drug age race1 race2/rl;
    if race = 1 then race1 = 1; else race1 = 0;
    if race = 2 then race2 = 1; else race2 = 0;
    NoRace:  test race1, race2;
run;
```


Testing the significance of categorical variables (cont.)

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
NoRace	3.6047	2	0.1649

We conclude that RACE does not contribute significantly to the model.

Computing survival estimates

Suppose we would like estimates of the survival function for those with and without a history of injecting drug use based on our fitted Cox model.

```
proc phreg data = one;
  model time*censor(0) = drug age/rl;
  baseline out = survest survival = _all_/cltype = loglog;
run;

proc print data = survest; run;
```

Obs	drug	age	time	Survival	StdErr Survival	Lower Survival	Upper Survival
1	0.49	36.07	0	1.00000	.	.	.
2	0.49	36.07	1	0.88063	0.027265	0.81473	0.92416
3	0.49	36.07	2	0.83570	0.033079	0.75841	0.89003
4	0.49	36.07	3	0.73104	0.041888	0.63868	0.80340

Computing survival estimates (cont.)

Note that the BASELINE statement computes survival at the *mean* of the covariate values. Although average age is meaningful, the average value of DRUG is not meaningful since its values (0/1) represent categories. To get meaningful survival estimates, we modify the code on the previous slide as follows.

Computing survival estimates (cont.)

```
data covvals;
    input drug age;
    cards;
0 36.07
1 36.07
;
run;

proc phreg data = one;
    model time*censor(0) = drug age/rl;
    baseline out = survest covariates = covvals
    survival = _all_/nomean cltype = loglog;
run;

proc print data = survest; run;
```


Computing survival estimates (cont.)

Obs	drug	age	time	Survival	StdErr Survival	Lower Survival	Upper Survival
1	No IDU history	36.07	0	1.00000	.	.	.
2	No IDU history	36.07	1	0.92297	0.021532	0.86779	0.95570
3	No IDU history	36.07	2	0.89299	0.027546	0.82441	0.93581
4	No IDU history	36.07	3	0.82074	0.038967	0.72876	0.88397
29	IDU history	36.07	0	1.00000	.	.	.
30	IDU history	36.07	1	0.81430	0.040532	0.71871	0.88006
31	IDU history	36.07	2	0.74822	0.048153	0.63886	0.82881
32	IDU history	36.07	3	0.60275	0.056909	0.48207	0.70380

Graphing survival estimates

```
proc gplot data = survest;
    plot survival*time = drug;
    symbol1 interpol=stepLJ c=blue;
    symbol2 interpol=stepLJ c=red;
run;
quit;
```


Graphing survival estimates (cont.)