

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2006, The Johns Hopkins University and John McGready. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Regression for Survival Analysis

John McGready
Johns Hopkins University

Cox proportional hazards (PH) regression

Interpreting coefficients from Cox PH regression

Performing inference on Cox PH regression coefficients



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

The Cox Proportional Hazard Regression Model

In Statistical Reasoning 1 we learned why we need a different approach to analyzing time to event data in the presence of censoring

We learned to estimate the “survival function” via the Kaplan Meier method for a cohort of subjects

We learned to statistically compare the “survival functions” of two or more cohorts via the logrank and Wilcoxon-Gehan tests

However, those statistical tests did not provide an estimate of the magnitude of the difference in survival for the cohorts being compared

Further, the tests allowed for comparisons based on one grouping factor (predictor) at a time

How can we get an estimate of the magnitude of the survival-predictor relationship of interest?

How can we account for multiple factors simultaneously for each subject in a time to event study?

How can we estimate adjusted survival-predictor relationships in the presence of potential confounding?

Regression Methods for Censored Survival Data

Objective

- ★ *Relate survival times to (potentially multiple) predictors (the x 's or independent variables)*

Regression Methods for Censored Survival Data

Problem

- ★ *Can't use ordinary linear regression because how do you account for the censored data?*
- ★ *Can't use logistic regression without ignoring the time component*

Proportional Hazards Regression Model

Developed by D.R. Cox (1972)

Relates survival time to predictors

Handles incomplete follow-up

★ *Censoring*

Proportional Hazards Regression Model

It assumes the ratio of time-specific outcome (event) risks (hazard) of two groups remains about the same over time

This ratio is called the hazards ratio or the relative risk

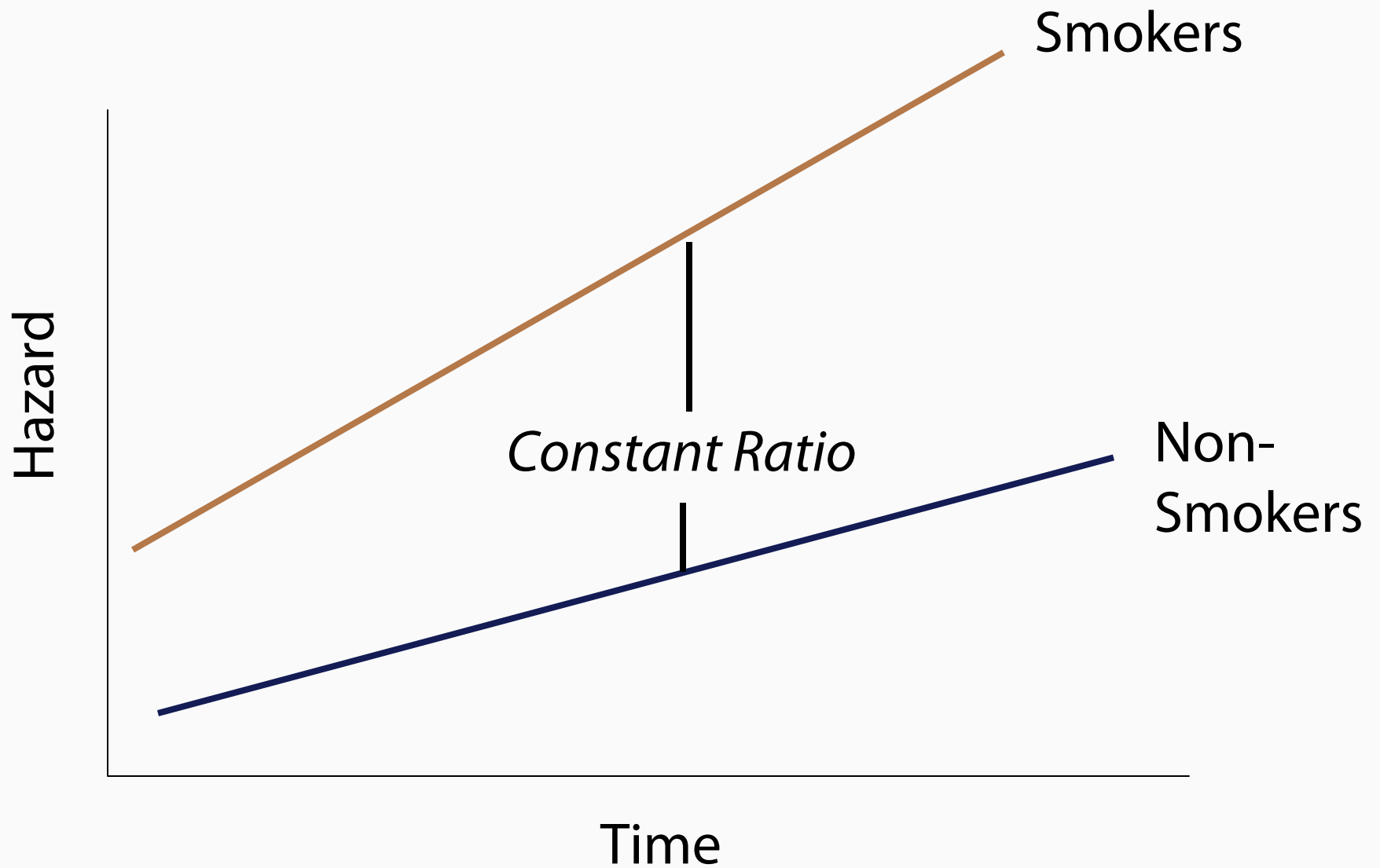
Proportional Hazards Assumption

All Cox regression requires is an assumption that ratio of hazards is constant over time across groups

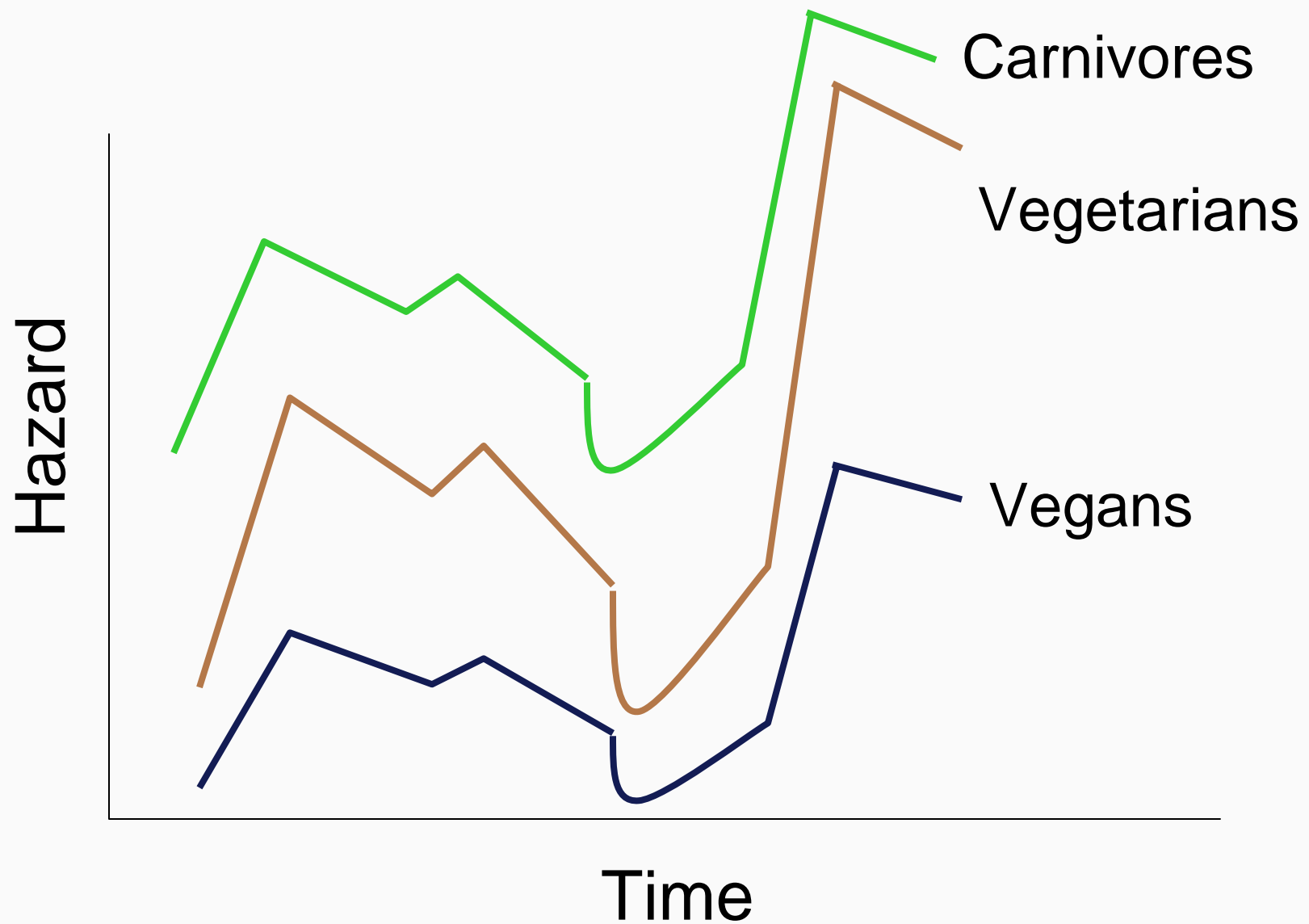
The good news—we don't need to know anything about overall shape of risk/hazard over time

The bad news—the proportionality assumption can be restrictive

Proportional Hazards Assumption



Proportional Hazards Assumption



Formulation of model:

- ★ *Group hazard*
- ★ = *Baseline hazard* \times “*group factor*”

Cox Proportional Hazards Model

Formulation of the model:

Group Hazard

$$= \text{Baseline Hazard} \times e^{b_1x_1 + b_2x_2 + \dots}$$

Cox Proportional Hazards Model

Such that ...

$$\log\left(\frac{\textit{Group hazard}}{\textit{baseline hazard}}\right) = b_1x_1 + b_2x_2 + \dots$$

Cox Proportional Hazards Model

312 patients with primary biliary cirrhosis (PBC) studied at the Mayo clinic

Patients were followed from diagnosis until death or censoring

Information available includes sex and age (years) of each patient

Question—how do patient's age and sex predict survival?

Here is a snippet of the data:

```
+-----+
|      survyr      death      sex      ageyr      |
+-----+-----+-----+-----+
1. | 4.663014         0         1      51.52329      |
2. | 3.838356         0         1      46.38082      |
3. | 7.843836         0         1      49.63836      |
4. | 6.178082         0         1      62.03288      |
5. | 7.243835         0         1      55.60548      |
+-----+-----+-----+-----+
6. | 8.490411         0         1      56.60822      |
7. | 9.282192         0         1      62.56438      |
8. | 10.93699         0         1      40.23014      |
9. | 11.66027         0         0      43.92877      |
10. | 7.920548         0         1      35.01096      |
+-----+-----+-----+-----+
```

The variables:

survyr is a time measurement in years

death is an indicator of death (1) or censoring (0)

sex is an indicator (1 = female, 0 = male)

ageyr is age in years

Letting Stata Know It Is Time to Event Data

The “stset” command tells Stata that we have time to event data—Stata converts it internally and then we have use of a bunch of built in commands

Syntax . . .

```
stset time_variable, failure(event_variable =1)
```

So for our data the syntax is . . .

```
stset survyr, failure(death=1)
```

Letting Stata Know It Is Time to Event Data

Here is what Stata does . . .

```
. stset  survyr, failure(death=1)

      failure event:  death == 1
obs. time interval:  (0, survyr]
exit on or before:  failure

-----

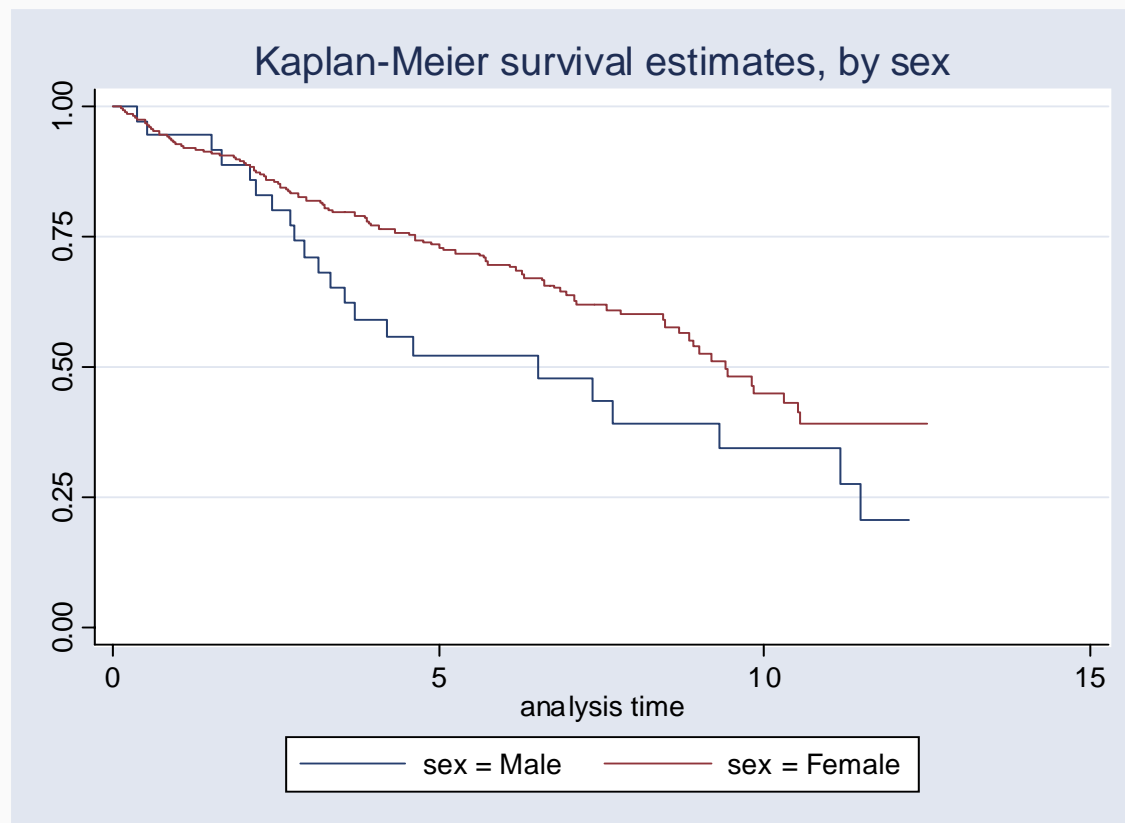
      312  total obs.
       0  exclusions

-----

      312  obs. remaining, representing
      125  failures in single record/single failure data
1715.027  total analysis time at risk, at risk from t =
                                                earliest observed entry t =
                                                last observed exit t =
```

Let's look at Kaplan-Meier curves by sex

Command: `sts graph, by(sex)`



To do a log rank test

Command: *sts test, by(sex)*

```
. sts test sex
```

```
        failure _d:  death == 1  
analysis time _t:  survyr
```

Log-rank test for equality of survivor functions

sex	Events observed	Events expected
Male	22	14.62
Female	103	110.38
Total	125	125.00

```
chi2(1) = 4.27  
Pr>chi2 = 0.0388
```

So there is visual evidence that females have longer survival than males

The results of the log-rank test show a significant difference in the survival experience of males and females

However, so far we have no measure of the association between longer survival and being female—how can we get this?

How about a regression model?

$$\log\left(\frac{\textit{Group hazard}}{\textit{baseline hazard}}\right) = b_1 x_1$$

(here x_1 is sex, 1 for females and 0 for males)

Let's figure out how to interpret b_1

★ *Model for females*

$$\log\left(\frac{\text{Hzd : Females}}{\text{baseline hazard}}\right) = b_1 * 1$$
$$= b_1$$

Let's figure out how to interpret b_1

★ *Model for males*

$$\log\left(\frac{Hzd : Females}{baseline hazard}\right) = b_1 * 0$$
$$= 0$$

Taking the difference

$$\log\left(\frac{Hzd : Females}{baseline\ hazard}\right) - \log\left(\frac{Hzd : Males}{baseline\ hazard}\right) = b_1$$

Invoking our favorite property of logs

$$\log\left(\frac{Hzd : Females}{Hzd : Males}\right) = b_1$$

So b_1 is the log of the hazard ratio (relative risk) of death for females compared to males

Such that ...

$$\left(\frac{Hzd : Females}{Hzd : Males} \right) = e^{b_1}$$

So e^{b_1} is the hazard ratio (relative risk) of death for males to females

Interpretation

- ★ $b_1 > 0$: Higher hazard (poorer survival) associated with being female
- ★ Because $e^{b_1} > 1$

Interpretation

- ★ $b_1 < 0$: Lower hazard (better survival) associated with being female
- ★ Because $e^{b_1} < 1$

Interpretation

- ★ $b_1 = 0$: *No association between hazard (and survival) and being female*
- ★ *Because $e^{b_1} = 1$*

With a sample of data, we are only going to be estimating b_1 —so the regression equation we estimate from our sample of 312 patients looks like ...

$$\log\left(\frac{\textit{Group hazard}}{\textit{baseline hazard}}\right) = \hat{b}_1 x_1$$

For this dataset, the estimated regression equation is . . .

$$\log\left(\frac{\textit{Group hazard}}{\textit{baseline hazard}}\right) = -.48x_1$$

So $\hat{b}_1 = -.48$

So we've estimated a negative association between death and being female—the hazard (risk) of death is lower for females in this sample

How to describe this association?

The estimate hazard ratio (relative risk) of death for females relative to males is $e^{(-.48)} = .62$

Females have .62 the hazard (risk) that males have of death (or females have 38% lower hazard (risk) of death than males)

The “stcox” command will estimate the regression

General syntax: `stcox x1`

Where x_1 is the predictor of interest

Notice we do not need to specify an outcome (Stata already knows it because we have declared the time and censoring variables with the `stset` command!)

Results from the *stcox* command

```
. stcox sex

      failure _d:  death == 1
      analysis time _t:  survyr

Iteration 0:   log likelihood = -639.97989
Iteration 1:   log likelihood = -638.17435
Iteration 2:   log likelihood = -638.09351
Iteration 3:   log likelihood = -638.09345
Refining estimates:
Iteration 0:   log likelihood = -638.09345

Cox regression -- Breslow method for ties

No. of subjects =          312          Number of obs   =          312
No. of failures =           125
Time at risk    = 1715.027399

LR chi2(1)      =           3.77
Prob > chi2     =           0.0521

Log likelihood = -638.09345
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	sex	.6163198	.145757	-2.05	0.041	.3877043 .9797418

Notice `stcox` does the conversion to the hazard ratio for us!

If we wanted information about the coefficient, \hat{b}_1 , we could use the “nohr” option

General syntax: `stcox x1, nohr`

Output with “nohr” option

```
stcox sex, nohr
```

```
      failure _d:  death == 1  
analysis time _t:  survyr
```

```
Iteration 0:  log likelihood = -639.97989  
Iteration 1:  log likelihood = -638.17435  
Iteration 2:  log likelihood = -638.09351  
Iteration 3:  log likelihood = -638.09345  
Refining estimates:  
Iteration 0:  log likelihood = -638.09345
```

```
Cox regression -- Breslow method for ties
```

```
No. of subjects =          312          Number of obs   =          312  
No. of failures =          125  
Time at risk    = 1715.027399  
  
Log likelihood  = -638.09345          LR chi2(1)       =          3.77  
                                          Prob > chi2     =          0.0521
```

```
-----  
      _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
      sex |  -.4839893   .2364957   -2.05   0.041   - .9475123   - .0204662  
-----
```

Accounting for Sampling Variability

The coefficient and hazard ratio estimates are based on an imperfect sample of 312 subjects from a large population/process

To complete the story, we need to incorporate sampling error into these estimates via a confidence interval and a p-value

Accounting for Sampling Variability

Luckily, Stata does this for us!

```
. stcox sex

      failure _d:  death == 1
      analysis time _t:  survyr

Iteration 0:    log likelihood = -639.97989
Iteration 1:    log likelihood = -638.17435
Iteration 2:    log likelihood = -638.09351
Iteration 3:    log likelihood = -638.09345
Refining estimates:
Iteration 0:    log likelihood = -638.09345

Cox regression -- Breslow method for ties

No. of subjects =          312          Number of obs   =          312
No. of failures =          125
Time at risk    = 1715.027399
Log likelihood   = -638.09345          LR chi2(1)       =          3.77
                                          Prob > chi2     =          0.0521
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sex	.6163198	.145757	-2.05	0.041	.3877043 .9797418

Continued

Accounting for Sampling Variability

The confidence interval gives a range of plausible values for the true hazard ratio of death for females compared to males in the population of PBC patients

The p-values are testing ...

H_0 : hazard ratio = 1

H_A : hazard ratio \neq 1

Describing the results

In a sample of 312 PBC patients, females had a lower hazard (risk) of death than males. The estimated hazard ratio was .62 indicating that females had 38% lower hazard (risk) of death than males.

Accounting for sampling variability, the decrease in risk for females could be as large as 62% or as small as 3% (95% CI for the hazard ratio 0.38–0.97).

Accounting for Sampling Variability

Where do the CIs and p-value come from?

It turns out all inference is done on the coefficient scale

The 95% confidence interval for b_1 is $\hat{b}_1 \pm 2SE(\hat{b}_1)$

The endpoints of this 95% CI can be exponentiated to get the 95% CI for the hazard (risk) ratio

Testing

$H_0: b_1 = 0$ is equivalent to testing

$H_0: e^{b_1}(\text{hazard ratio}) = 1$

This is done by computing a test statistic: $z = \frac{\hat{b}_1}{se(\hat{b}_1)}$

And comparing to standard normal curve to get p-value

So its business as usual!

Try it out using results below !

```
stcox sex, nohr
```

```
      failure _d:  death == 1  
analysis time _t:  survyr
```

```
Iteration 0:   log likelihood = -639.97989  
Iteration 1:   log likelihood = -638.17435  
Iteration 2:   log likelihood = -638.09351  
Iteration 3:   log likelihood = -638.09345  
Refining estimates:  
Iteration 0:   log likelihood = -638.09345
```

```
Cox regression -- Breslow method for ties
```

```
No. of subjects =           312           Number of obs   =           312  
No. of failures =           125  
Time at risk    = 1715.027399  
  
Log likelihood  = -638.09345           LR chi2(1)        =           3.77  
                                           Prob > chi2      =           0.0521
```

```
-----  
      _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
      sex |  -.4839893   .2364957   -2.05   0.041   - .9475123   - .0204662  
-----
```



JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section B

Multiple Predictors in a Cox Regression Model

Death/Sex in PBC Patients

We've already seen that female PBC patients have a lower risk of death than male PBC patients

We also have info on patients age as well—is it possible that age is an important predictor of death above and beyond sex?

Is it possible that age confounds the death/sex relationship?

We can investigate this via Cox regression!

Relationship Between Death and Age

Cox regression relating hazard (risk) of death to age

```
. stcox ageyr

      failure _d:  death == 1
      analysis time _t:  survyr

Iteration 0:  log likelihood = -639.97989
Iteration 1:  log likelihood = -629.73129
Iteration 2:  log likelihood = -629.72592
Refining estimates:
Iteration 0:  log likelihood = -629.72592

Cox regression -- Breslow method for ties

No. of subjects =          312          Number of obs   =          312
No. of failures =          125
Time at risk    = 1715.027399
Log likelihood  = -629.72592          LR chi2(1)       =          20.51
                                          Prob > chi2     =          0.0000

-----+-----
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      ageyr |   1.040781   .0091647    4.54   0.000    1.022973    1.0589
-----+-----
```

Estimated hazard ratio for *ageyr*—1.04, with 95% CI 1.02 to 1.05

Interpretation—in this sample a one-year increase in age is associated with a 4% increase in the hazard (risk) of death

Accounting for sampling variability, this increase could be as small as 2% or as large as 5%

In other words, if we were to compare two groups of persons who differ by one year in age, the older groups have a 5% higher risk of death than the younger group

Age and Sex Together as Predictors

We can easily estimate the results from a Cox regression model that includes both age and sex as predictors!

$$\log\left(\frac{\textit{Group hazard}}{\textit{baseline hazard}}\right) = \hat{b}_1 x_1 + \hat{b}_2 x_2$$

Where $x_1 = \text{sex}$ (1 = female, 0 = male)

$x_2 = \text{age (years)}$

Interpretation of Regression Coefficients

\hat{b}_1 is the estimate log of the age-adjusted hazard (risk) ratio of death for females compared to males

\hat{b}_2 is the estimated log of the sex-adjusted hazard (risk) ratio of death for two groups of subjects who differ by one year in age

Age and Sex Together as Predictors

Results from Stata with hazard ratios (exponentiated coefficients)

```
. stcox sex ageyr

      failure _d:  death == 1
      analysis time _t:  survyr

Iteration 0:   log likelihood = -639.97989
Iteration 1:   log likelihood = -628.87866
Iteration 2:   log likelihood = -628.77147
Iteration 3:   log likelihood = -628.77138
Refining estimates:
Iteration 0:   log likelihood = -628.77138

Cox regression -- Breslow method for ties

No. of subjects =          312          Number of obs   =          312
No. of failures =          125
Time at risk    = 1715.027399

Log likelihood  = -628.77138          LR chi2(2)       =          22.42
                                          Prob > chi2    =          0.0000

-----+-----
      _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      sex |   .7107455   .1695274   -1.43   0.152     .4453332     1.13434
      ageyr |  1.039023   .0091919    4.33   0.000     1.021162     1.057195
-----+-----
```

Continued

Results

Age-adjusted hazard ratio of death, females to males:
0.71 (95% CI 0.44–1.13)

Sex-adjusted hazard ratio associated with a one-year
difference in age: 1.05 (95% CI 1.02–1.06)