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
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?
Objective

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

Hazard vs. Time

Smokers
Non-Smokers

Constant Ratio
Proportional Hazards Assumption

- Carnivores
- Vegetarians
- Vegans

Hazard vs. Time
Formulation of model:

- Group hazard
- $= \text{Baseline hazard} \times \text{“group factor”}$
Formulation of the model:

\[ \text{Group Hazard} = \text{Baseline Hazard} \times e^{b_1x_1 + b_2x_2 + \ldots} \]
Cox Proportional Hazards Model

Such that ...

\[
\log \left( \frac{\text{Group hazard}}{\text{baseline hazard}} \right) = b_1 x_1 + b_2 x_2 + \ldots
\]
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:

<table>
<thead>
<tr>
<th></th>
<th>survyr</th>
<th>death</th>
<th>sex</th>
<th>ageyr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.663014</td>
<td>0</td>
<td>1</td>
<td>51.52329</td>
</tr>
<tr>
<td>2</td>
<td>3.838356</td>
<td>0</td>
<td>1</td>
<td>46.38082</td>
</tr>
<tr>
<td>3</td>
<td>7.843836</td>
<td>0</td>
<td>1</td>
<td>49.63836</td>
</tr>
<tr>
<td>4</td>
<td>6.178082</td>
<td>0</td>
<td>1</td>
<td>62.03288</td>
</tr>
<tr>
<td>5</td>
<td>7.243835</td>
<td>0</td>
<td>1</td>
<td>55.60548</td>
</tr>
<tr>
<td>6</td>
<td>8.490411</td>
<td>0</td>
<td>1</td>
<td>56.60822</td>
</tr>
<tr>
<td>7</td>
<td>9.282192</td>
<td>0</td>
<td>1</td>
<td>62.56438</td>
</tr>
<tr>
<td>8</td>
<td>10.93699</td>
<td>0</td>
<td>1</td>
<td>40.23014</td>
</tr>
<tr>
<td>9</td>
<td>11.66027</td>
<td>0</td>
<td>0</td>
<td>43.92877</td>
</tr>
<tr>
<td>10</td>
<td>7.920548</td>
<td>0</td>
<td>1</td>
<td>35.01096</td>
</tr>
</tbody>
</table>
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)

Continued
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: \textit{sts test, by(sex)}

\begin{verbatim}
. sts test sex

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

Log-rank test for equality of survivor functions

<table>
<thead>
<tr>
<th></th>
<th>Events</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>observed</td>
<td>expected</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>22</td>
<td>14.62</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>103</td>
<td>110.38</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
<td>125.00</td>
<td></td>
</tr>
</tbody>
</table>

\textit{chi2(1) = 4.27}
\textit{Pr>chi2 = 0.0388}
\end{verbatim}
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?

\[ \log \left( \frac{\text{Group hazard}}{\text{baseline hazard}} \right) = b_1 x_1 \]

(here \( x_1 \) is sex, 1 for females and 0 for males)
Let’s figure out how to interpret $b_1$

- *Model for females*

\[
\log \left( \frac{Hzd: Females}{baseline \ hazard} \right) = b_1 \times 1
\]

\[= b_1\]
Let’s figure out how to interpret $b_1$

- Model for males

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

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

\[ \log\left( \frac{Hzd: \text{Females}}{Hzd: \text{Males}} \right) = b_1 \]

So \( b_1 \) is the log of the hazard ratio (relative risk) of death for females compared to males
Such that \( \frac{Hzd: Females}{Hzd: Males} = e^{b_1} \)

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

Interpretation

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

Interpretation

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

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 . . .

\[
\log\left( \frac{\text{Group hazard}}{\text{baseline hazard}} \right) = \hat{b}_1 x_1
\]
For this dataset, the estimated regression equation is . . .

\[
\log \left( \frac{\text{Group hazard}}{\text{baseline hazard}} \right) = -.48 x_1
\]

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 $x_1$

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!)
The Stata Output

Results from the *stcox* command

```
. stcox  sex

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

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                          LR chi2(1)     =  3.77
Time at risk   = 1715.027399                 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
------------------------------------------------------------------------------
```

Continued
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 \( x_1 \), 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
---------------------------------------------------------------------
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:  suryry

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
No. of failures = 125
Time at risk = 1715.027399
Log likelihood = -638.09345

Number of obs = 312
LR chi2(1) = 3.77
Prob > chi2 = 0.0521

+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+
|     _t | Haz. Ratio | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|-------+-----------+-----------+-----+-----+-----------------+-----------------+-----------------+
| sex   | 0.6163198 | 0.145757 | -2.05 | 0.041 | 0.3877043 | 0.9797418 |
|-------+-----------+-----------+-----+-----+-----------------+-----------------+-----------------+
```

Continued 43
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   | -0.4839893 |  0.2364957 | -2.05  | 0.041 | -0.9475123 -.0204662   |