Applied Survival Analysis
Lab 4: Cox Proportional Hazards Model

In today's lab, we are going to review the basic interpretation of a Cox proportional hazards model. Then we are going to learn how to fit a Cox PH model using STATA and evaluate the implication of tied failure times.

1. ***<u>Interpretation of Cox Model: Prognosis with Breast Cancer</u>***

A follow-up study of post-menopausal women diagnosed with breast cancer was performed to examine whether the estrogen receptor (ER) status of the tumor was related to prognosis, adjusting for stage of disease at diagnosis and age at diagnosis.

Let $\lambda(t, X)$ be the hazard of death at month *t* after diagnosis for an individual with covariates **X**, **X** = (X_E, X_A, X_S) where

$$X_E = \begin{cases} 1 & \text{ER positive tumor} \\ 0 & \text{ER negative tumor} \end{cases}$$

$$X_A = \text{age in years at diagnosis}$$

$$X_S = \text{stage} = \begin{cases} 1 & \textit{in situ} \\ 2 & \text{local} \\ 3 & \text{regional} \\ 4 & \text{distant} \end{cases}$$

Suppose the following proportional hazards model was found to fit the data:

$$\lambda(t, X) = \lambda_{0X_S}(t)\exp(-0.783X_E + 0.007X_A)$$

where $\lambda_{0X_S}(t)$ is the baseline hazard function, specific for stage (i.e. value of X_S). (This is a *stratified* proportional hazards model, stratified by tumor stage.)

(For the following questions, get an actual number if you can, showing how you got it. If you can't get an actual number, then write an expression for how it would be calculated if you had additional information)

(a) Based on this model, what is your best estimate of the hazard ratio (i.e. relative risk) of death for a woman with an ER positive tumor relative to a woman of the same age and with the same stage ER negative tumor, the same number of months beyond diagnosis? Do women with ER positive tumors have a more favorable or less favorable prognosis?

(b) Based on this model, what can you say about the hazard ratio of death for a woman 62 years old at diagnosis with a localized ER positive tumor, 24 months beyond diagnosis, relative to a woman 67 years old at diagnosis with a localized ER negative tumor, 24 months beyond diagnosis?

(c) Based on this model, what can you say about the hazard ratio of death for a woman 55 years old at diagnosis with an *in situ* ER positive tumor, 36 months beyond diagnosis, relative to a woman 55 years old at diagnosis with an *in situ* ER negative tumor, 24 months beyond diagnosis?

(d) Based on this model, what can you say about the hazard ratio of death for a woman 60 years old at diagnosis with an ER positive tumor with regional spread, 24 months beyond diagnosis, relative to a woman 60 years old at diagnosis with an ER negative *in situ* tumor, 24 months beyond diagnosis?

2. *<u>Fitting Cox Model and handling of ties: Nursing Home Data</u>*
(Morris et al., *Case Studies in Biometry*, Ch 12)

Now we are going to move into STATA. We are going to consider the same example as last time (*nurshome.dta*). The National Center for Health Services Research studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay. “Treated” nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient's health and sending them home.

The study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

<u>Variables include:</u>
`los` - Length of Stay of Resident(days)
`age` - Age of Resident
`rx` - Nursing Home Assignment
`gender` - Sex
`married` - Marital Status
`health` - Health Status
`fail` - Event Indicator

Again before starting any analysis we have to *stset* our data: **`stset  los fail`**

The command to fit a Cox proportional hazards model in STATA is **`stcox`** , e.g we want to evaluate the effect of marital status, we would type the following command:

`stcox  married`

```
         failure _d:  fail
   analysis time _t:  los

Iteration 0:   log likelihood = -8556.5713
Iteration 1:   log likelihood = -8548.0345
Iteration 2:   log likelihood =  -8547.915
Iteration 3:   log likelihood =  -8547.915
Refining estimates:
Iteration 0:   log likelihood =  -8547.915

Cox regression -- Breslow method for ties

No. of subjects =         1591                     Number of obs   =      1591
No. of failures =         1269
Time at risk    =       386211
                                                   LR chi2(1)      =     17.31
Log likelihood  =    -8547.915                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 married |   1.363757   .0984086      4.299   0.000       1.183898     1.57094
```

If would like to display the coefficient instead of the hazard ratio we would add the option **`nohr`** :

`stcox  married, nohr`

```
         failure _d:  fail
   analysis time _t:  los

Iteration 0:   log likelihood = -8556.5713
Iteration 1:   log likelihood = -8548.0345
Iteration 2:   log likelihood =  -8547.915
Iteration 3:   log likelihood =  -8547.915
Refining estimates:
Iteration 0:   log likelihood =  -8547.915

Cox regression -- Breslow method for ties

No. of subjects =         1591                     Number of obs   =      1591
No. of failures =         1269
Time at risk    =       386211
                                                   LR chi2(1)      =     17.31
Log likelihood  =    -8547.915                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 married |   .3102432   .0721599      4.299   0.000       .1688124     .451674
```

Note that the default way of handling the ties is the *Breslow* method.

(a) Using the estimate of β, write out the appropriate Cox PH model. What is the estimated hazard ratio? How would you interpret this hazard ratio?

Now to get the three other methods for tied failure times we add the options:
`efron` for the *Efron* method
`exactp` for the *Discrete* method
`exactm` for the *Exact Marginal Likelihood* (approximation)

So we have (we will run the **`exactp`** last, since it takes more time)

`stcox married, nohr efron`

```
         failure _d:  fail
   analysis time _t:  los

Iteration 0:   log likelihood = -8553.0704
Iteration 1:   log likelihood = -8544.4793
Iteration 2:   log likelihood = -8544.3581
Iteration 3:   log likelihood = -8544.3581
Refining estimates:
Iteration 0:   log likelihood = -8544.3581

Cox regression -- Efron method for ties

No. of subjects =         1591                     Number of obs   =      1591
No. of failures =         1269
Time at risk    =       386211
                                                   LR chi2(1)      =     17.42
Log likelihood  =   -8544.3581                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 married |   .3112771   .0721591      4.314   0.000         .169848    .4527063
```

`stcox married, nohr exactm`

```
         failure _d:  fail
   analysis time _t:  los

( more iterations )
Refining estimates:
Iteration 0:   log likelihood = -7224.9616

Cox regression -- exact marginal likelihood

No. of subjects =         1591                     Number of obs   =      1591
No. of failures =         1269
Time at risk    =       386211
                                                   LR chi2(1)      =     17.42
Log likelihood  =   -7224.9616                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 married |   .3112798   .0721594      4.314   0.000        .1698499    .4527096
```

```
stcox  married, nohr exactp

         failure _d:  fail
   analysis time _t:  los

( more iterations )
Refining estimates:
Iteration 0:   log likelihood = -7224.9632

Cox regression -- exact partial likelihood

No. of subjects =         1591                     Number of obs   =      1591
No. of failures =         1269
Time at risk    =       386211
                                                   LR chi2(1)      =     17.42
Log likelihood  =   -7224.9632                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 married |   .3122564   .0724154      4.312   0.000       .1703248    .4541881
```

(b) How much impact do the different approaches have on the estimate of β? How much impact do the different approaches have on the computing time required? Would you expect tied failure times to be a big issue in this dataset? Which approach would you recommend?

(c) Compare the test statistics from parts (a) and (b) above to the log-rank and Wilcoxon tests you obtained in part (b) of Lab 3. Do any of them match? (Try also the command **`sts test married, cox`**). If not, how would you get the same test statistic from a log-rank test as for one of the tied options in a Cox model? Which type of test statistic is this (score, likelihood ratio, or Wald)?