Applied Survival Analysis - January 2016

Solutions to Lab 4: Cox Proportional Hazards Model

(a) **Interpretation of Cox Model: Prognosis with Breast Cancer.**

(i) Taking into account the following proportional hazards model

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

The estimated hazards for a woman with ER positive tumor and for a woman with ER negative tumor, holding all other values constant, are given by:

$$\lambda(t|X_E = 1, X_A, X_S) = \lambda_{0,X_S}(t) \exp\left\{-0.783 + 0.007X_A\right\}$$
$$\lambda(t|X_E = 0, X_A, X_S) = \lambda_{0,X_S}(t) \exp\left\{0 + 0.007X_A\right\}$$

Therefore, the estimated hazard ratio for ER positive woman versus ER negative is given by:

$$HR = \frac{\lambda(t|X_E = 1, X_A, X_S)}{\lambda(t|X_E = 0, X_A, X_S)} = \exp\{-0.783\} = 0.457$$

This implies that women with ER positive tumors have an approximately 54% ((1-0.46)100%) lower risk of dying compared to women with ER negative tumors. So women with ER positive tumors have a more favorable prognosis.

(ii) The two hazards that we are interested in are the following:

$$\lambda(24|X_E = 1, X_A = 62, X_S = 2) = \lambda_{0,X_S=2}(24) \exp\{-0.783 + 0.007 \cdot 62\}$$
$$\lambda(24|X_E = 1, X_A = 67, X_S = 2) = \lambda_{0,X_S=2}(24) \exp\{0 + 0.007 \cdot 67\}$$

Thus, the corresponding hazard ratio is:

$$HR = \frac{\lambda(24|X_E = 1, X_A = 62, X_S = 2)}{\lambda(24|X_E = 1, X_A = 67, X_S = 2)} = \exp\{-0.783 - 0.007 \cdot 5\} = 0.441$$

So the hazard of death for a woman 62 years old at diagnosis with a localized ER positive tumor, 24 months beyond diagnosis is approximately 56% less than

a woman 67 years old at diagnosis with localized ER negative tumor, 24 months beyond diagnosis.

(iii) In this case the two hazards of interest are the following:

$$\lambda(36|X_E = 1, X_A = 55, X_S = 1) = \lambda_{0,X_S=1}(36)\exp\{-0.783 + 0.007 \cdot 55\}$$
$$\lambda(24|X_E = 0, X_A = 55, X_S = 1) = \lambda_{0,X_S=1}(24)\exp\{0 + 0.007 \cdot 55\}$$

Thus, the corresponding hazard ratio is:

$$HR = \frac{\lambda(36|X_E = 1, X_A = 55, X_S = 1)}{\lambda(24|X_E = 0, X_A = 55, X_S = 1)} = \frac{\lambda_{0,X_S=1}(36)}{\lambda_{0,X_S=1}(24)} \cdot \exp\{-0.783\} = \frac{\lambda_{0,X_S=1}(36)}{\lambda_{0,X_S=1}(24)} \cdot 0.457$$

So the hazard of death for a woman 55 years old at diagnosis with a in situ ER positive tumor, 36 months beyond diagnosis is approximately $\frac{\lambda_{0,X_S=1}(36)}{\lambda_{0,X_S=1}(24)} \cdot 0.457$ times the hazard for a woman with the same age at diagnosis and with in situ ER negative tumor, 24 months beyond diagnosis. We need the baseline hazard for those with an in situ tumor to get the actual hazard ratio!

(iv) The two hazards that we are interested in are the following:

$$\lambda(24|X_E = 1, X_A = 60, X_S = 3) = \lambda_{0,X_S=3}(24)\exp\{-0.783 + 0.007 \cdot 60\}$$
$$\lambda(24|X_E = 0, X_A = 60, X_S = 1) = \lambda_{0,X_S=1}(24)\exp\{0 + 0.007 \cdot 60\}$$

Therefore the corresponding hazard ratio is:

$$HR = \frac{\lambda(24|X_E = 1, X_A = 60, X_S = 3)}{\lambda(24|X_E = 0, X_A = 60, X_S = 1)} = \frac{\lambda_{0,X_S=3}(24)}{\lambda_{0,X_S=1}(24)} \cdot 0.457$$

Note that we need the ratio of the baseline hazards of the regional and in situ stage, because stratified analysis assumes that the coefficients are the same across strata but baseline hazards are unique to each stratum. So the hazard of death for a woman 60 years old at diagnosis with a regional ER positive tumor, 24 months beyond diagnosis is approximately $\frac{\lambda_{0,X_S=3}(24)}{\lambda_{0,X_S=1}(24)} \cdot 0.457$ times the hazard for a woman 60 years old at diagnosis with in situ ER negative tumor, 24 months beyond diagnosis.

(b) **Fitting Cox Model and handling of ties: Nursing Home Data**.
First, let's have a look at the data

```
> ############################################
> ### LAB 4: Cox Proportional Hazards Model ###
> ############################################
```

```
>
> # Fitting Cox Model and handling of ties: Nursing Home Data
> # Import data
> nurshome = read.csv("C:/Applied_Survival_Analysis_Jan2016/lab4/data/nurshome.csv")
>
> # See few lines of the data
> nurshome[1:5,]
  los age rx gender married health fail
1 665  83  1      0       0      4    1
2 697  80  0      0       0      3    0
3   7  92  1      0       0      4    1
4 217  69  0      0       0      3    1
5 153  74  1      1       0      3    1
```

(i) In R, a Cox model can be fitted via the function `coxph` in the `survival` package. The *Efron* method is the default way of handling the ties (keep in mind that STATA uses the *Breslow* method by default). Again we must use the `Surv` function to specify the survival time along with the failure indicator. Here we are interested in investigating the effect of marital status on the length of stay. Thus, the formula to be used has to be `Surv(los,fail) ~ married`.

```
> # (a): The Efron method is the default!
> library(survival)
> fit = coxph( Surv(los,fail) ~ married,data = nurshome)
> summary(fit)
Call:
coxph(formula = Surv(los, fail) ~ married, data = nurshome)

  n= 1591, number of events= 1269

           coef exp(coef) se(coef)     z Pr(>|z|)
married 0.31128   1.36517  0.07216 4.314  1.6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
married     1.365     0.7325     1.185     1.573

Concordance= 0.523  (se = 0.006 )
Rsquare= 0.011   (max possible= 1 )
Likelihood ratio test= 17.42  on 1 df,   p=2.989e-05
Wald test            = 18.61  on 1 df,   p=1.605e-05
Score (logrank) test = 18.76  on 1 df,   p=1.484e-05
```

The estimate of β is $\hat{\beta} = 0.311$, and the corresponding Cox model is

$$\lambda(t|z) = \lambda_0(t)\exp(0.311 \cdot z), \quad z = \begin{cases} 1, & \text{Married} \\ 0, & \text{Unmarried} \end{cases} \Rightarrow$$

$$\text{HR(Married vs Unmarried)} = e^{0.311} = 1.365.$$

So the risk of discharge for married individuals is approximately 36.5% greater than single individuals. In other words, the length of stay for married individuals is shorter than for single individuals.

(ii) In the `coxph` function, you can specify the method for tie handling using the option `ties`. The available methods are

- The *Efron* method, `ties = "efron"` (default)
- The *Breslow* method, `ties = "breslow"`
- The *exact (discrete)* method, `ties = "exact"`.

So,

```
> # (b): Methods for tie handling
> # Breslow method
> fit.breslow = coxph( Surv(los,fail) ~ married,data = nurshome, ties = "breslow")
> summary(fit.breslow)
Call:
coxph(formula = Surv(los, fail) ~ married, data = nurshome, ties = "breslow")

  n= 1591, number of events= 1269

           coef exp(coef) se(coef)     z Pr(>|z|)
married 0.31024   1.36376  0.07216 4.299 1.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
married     1.364     0.7333     1.184     1.571

Concordance= 0.523  (se = 0.006 )
Rsquare= 0.011   (max possible= 1 )
Likelihood ratio test= 17.31  on 1 df,   p=3.171e-05
Wald test            = 18.48  on 1 df,   p=1.713e-05
Score (logrank) test = 18.63  on 1 df,   p=1.585e-05

>
> # Efron method
```

```
> fit.efron = coxph( Surv(los,fail) ~ married,data = nurshome, ties = "efron")
> summary(fit.efron)
Call:
coxph(formula = Surv(los, fail) ~ married, data = nurshome, ties = "efron")

  n= 1591, number of events= 1269

          coef exp(coef) se(coef)     z Pr(>|z|)
married 0.31128   1.36517  0.07216 4.314  1.6e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
married     1.365     0.7325     1.185     1.573

Concordance= 0.523  (se = 0.006 )
Rsquare= 0.011   (max possible= 1 )
Likelihood ratio test= 17.42  on 1 df,   p=2.989e-05
Wald test            = 18.61  on 1 df,   p=1.605e-05
Score (logrank) test = 18.76  on 1 df,   p=1.484e-05

>
> # Exact (discrete)
> fit.exact = coxph( Surv(los,fail) ~ married,data = nurshome, ties = "exact")
> summary(fit.exact)
Call:
coxph(formula = Surv(los, fail) ~ married, data = nurshome, ties = "exact")

  n= 1591, number of events= 1269

          coef exp(coef) se(coef)     z Pr(>|z|)
married 0.31226   1.36651  0.07242 4.312 1.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
married     1.367     0.7318     1.186     1.575

Rsquare= 0.011   (max possible= 1 )
Likelihood ratio test= 17.42  on 1 df,   p=2.994e-05
Wald test            = 18.59  on 1 df,   p=1.618e-05
Score (logrank) test = 18.74  on 1 df,   p=1.496e-05
```

The estimate of β is identical in all three methods, which implies that the tied failure times is not a big issue in this dataset. The computing time required for

the discrete method (`ties = "exact"`) wasn't that great, and since it is the best (exact) estimation of the hazard function we might as well use it.

(iii) We can also compute the *Logrank* or *Wilcoxon* test to explore the effect of marital status on the length of stay.

```
> # (c): Logrank and wilcoxon test
> # Logrank
> survdiff(Surv(los,fail) ~ married,data = nurshome)
Call:
survdiff(formula = Surv(los, fail) ~ married, data = nurshome)

             N Observed Expected (O-E)^2/E (O-E)^2/V
married=0 1319     1032     1086      2.68      18.7
married=1  272      237      183     15.90      18.7

 Chisq= 18.7  on 1 degrees of freedom, p= 1.5e-05
> chi2.logrank = survdiff(Surv(los,fail) ~ married,data = nurshome)$chisq
> chi2.logrank
[1] 18.74239
>
> # Peto & Peto modification of the Gehan-Wilcoxon test
> survdiff(Surv(los,fail) ~ married,data = nurshome, rho = 1)
Call:
survdiff(formula = Surv(los, fail) ~ married, data = nurshome,
    rho = 1)

             N Observed Expected (O-E)^2/E (O-E)^2/V
married=0 1319      615      649      1.82      17.1
married=1  272      150      116     10.23      17.1

 Chisq= 17.1  on 1 degrees of freedom, p= 3.58e-05
```

The *Logrank* test statistic (18.74) is expected to be close to the wald or LR test statistic obtained from the cox models of part (ii), as the *Logrank* test and the Cox model have the same underlying assumptions (proportional hazards). However, these test are not exactly identical.

During the lectures we saw (page 50) that the *Logrank* test is equivalent to the score test in the Cox model with the discrete tied option. To convince yourself, just type the following

```
> # The logrank test is equivalent to
> # the cox model using the "exact" option
> c(fit.exact$score,chi2.logrank)
[1] 18.74239 18.74239
>
```

```
> fit.exact$score - chi2.logrank
[1] 3.268497e-13
```

(c) **Optional:** You can use the following code to simulate the data

```
> # Simulated data from a clinical trial
> # Treament group: T ~ exp(rate = 2)
> # Control group: T ~ exp(rate = 4)
> set.seed(5)
> n = 5000
>
> # These are uncensored survival times
> timeTrt = rexp(n,rate = 2)
> timeCont = rexp(n,rate = 4)
> surv = c(timeTrt,timeCont)
>
> # Censoring time
> # Mean = 1/rate for exponentially distributed times
> cens = rexp(2*n,rate = 1/5)
>
> # Failure indicator
> delta = 1*(surv < cens)
>
> # Observed survival time
> time = pmin(surv,cens)
>
> # Save data in a data frame
> data = data.frame(time = time,fail = delta, trt = c(rep(1,n),rep(0,n) ))
>
> # Fit a cox model
> fit.cox = coxph(Surv(time,fail) ~ trt,data = data)
> summary(fit.cox)
Call:
coxph(formula = Surv(time, fail) ~ trt, data = data)

  n= 10000, number of events= 9330

        coef exp(coef) se(coef)      z Pr(>|z|)
trt -0.68274   0.50523  0.02161 -31.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
trt    0.5052      1.979    0.4843    0.5271
```

```
Concordance= 0.584  (se = 0.003 )
Rsquare= 0.094   (max possible= 1 )
Likelihood ratio test= 992.6  on 1 df,   p=0
Wald test            = 998.3  on 1 df,   p=0
Score (logrank) test = 1031  on 1 df,   p=0
```

The estimated hazard ratio is 0.5052 (95%CI: 0.48-0.52), which is very close to the true value of 0.50. Thus, the cox model seems appropriate for analysing such data. However, you have to keep in mind that the cox model is valid as long as

- the hazards are proportional across the treatment groups and
- the censoring times are independent of the survival times.

Both requirements are fulfilled in our example.