Pam Lovejoy

Cox's Proportional Hazards

The Cox Proportional Hazards Model was developed by Sir David Cox in the 1970s. It is a non-parametric alternative to other parametric survival models in the sense that it depends only on the ranks of the survival times. It is a survival model that is concerned with age-specific hazard, with or without censoring. Censoring refers to whether the time of death of all individuals is known. The without-censoring model does not extrapolate beyond the last data point. Cox's Proportional Hazards Model is the most widely used regression model for survival data. In this worksheet, it is used as a way to model the death rate of drosophila based on several different concentrations of atrazine, a common herbicide.

	A	B	C	D	E
1	Treat	block	Day	Dead	SS
2	2	1	3	1	302
3	2	1	3	1	302
4	2	1	4	1	302
5	2	1	5	1	302
6	2	1	6	1	302
7	2	1	7	1	302
8	2	1	7	1	302
9	2	1	7	1	302
10	2	1	7	1	302
11	2	1	8	1	302
12	2	1	8	1	302
13	2	1	8	1	302
14	2	1	8	1	302
15	2	1	8	1	302
16	2	1	9	1	302
17	2	1	9	1	302
18	2	1	9	1	302
19	2	1	9	1	302
20	2	1	9	1	302
21	2	1	9	1	302
22	2	1	9	1	302
23	2	1	10	1	302
24	2	1	10	1	302
25	2	1	10	1	302
26	2	1	10	1	302
27	2	1	10	1	302
28	2	1	10	1	302
29	2	1	10	1	302
30	2	1	10	1	302
31	2	1	10	1	302
32	2	1	10	1	302
33	2	1	11	1	302
34	2	1	11	1	302

Data Structure

- <u>-All data must be numeric</u>
- -Each row corresponds to an individual fly
- -treatment is the concentration of atrazine the fly was exposed to
- -block is the either the first or second repetition of the experiment
- -Day corresponds to the day the fly was found dead
- -The main dependent variable, Day has to be binary (0=alive, 1=dead)
- -SS is the sample size of the corresponding block

Model

The cox proportional hazard model assumes that the hazard is in the form:

$$\lambda(t;z)=\lambda_0(t)e^{z1\beta1+z2\beta2+ z3\beta3+\ldots+ zp\beta p}$$

where: z: a p x 1 vector of covariates such as treatment indicators, prognostic factors etc.
β: a p x 1 vector of regression coefficients (effect of each z)
$\lambda_0(t)$: an unspecified baseline hazard function that will cancel out in due course, this value also has to be >0

Creating the Model

```
Coxreg=coxph(Surv(Day,Dead)~Fblock+Ftreat+SampleSize
```

Here we have created the model with all of the relevant variables. The function in R for a cox regression is coxph(). With it, we want to observe Survival by the number of dead flies per day, with the independent variables: block, treatment, and sample size.

Make sure that variables with different levels are factored aka block and treatment here. Make sure to assign the other variables.

```
Ftreat=factor(cox$Treat)
```

```
Fblock=factor(cox$block)

Day=cox$Day
Dead=cox$Dead
SampleSize=cox$SS
```

Evaluate the model

```
coxreg=coxph(Surv(Day,Dead)~Fblock+Ftreat+SampleSize)
coxreg
Call:
coxph(formula = Surv(Day, Dead) ~ Fblock + Ftreat + SampleSize)


             coef      exp(coef) se(coef)    z      p
Fblock2    -0.11626     0.89 0.027488 -4.23 2.3e-05
Ftreat2     0.95226     2.59 0.044285 21.50 0.0e+00
Ftreat3     0.86904     2.38 0.044603 19.48 0.0e+00
Ftreat4     0.92763     2.53 0.044140 21.02 0.0e+00
Ftreat5     0.40772     1.50 0.048984  8.32 1.1e-16
SampleSize  0.00378     1.00 0.000308 12.29 0.0e+00

Likelihood ratio test=1143  on 6 df, p=0  n= 6396, number of events= 5608
```

This is an evaluation of the model. The coefficients are the βs of each variable and characterize the effect z of the variable on the model. Both the β and z values are given in the table and can be entered into the model if desired.

The p-values indicate whether each variable is contributing to the model. If any of them were a good amount higher than 0.05, then you could try dropping them from the model or using the step() function with the model to attempt to make a reduced model. In this case, however, all of the variables are relevant to the model under the hypotheses H0: β=0, Ha: β=/=0, so none of them should be dropped. If the step() function were to be employed, it would say that doing nothing has the lowest AIC meaning that the model is already at its most parsimonious.

Summarize the Model

```
summary(coxreg)
Call:
coxph(formula = Surv(Day, Dead) ~ Fblock + Ftreat + SampleSize)

  n= 6396, number of events= 5608

              coef      exp(coef)  se(coef)      z       Pr(>|z|)
Fblock2    -0.1162622  0.8902418  0.0274885 -4.229 2.34e-05 ***
Ftreat2     0.9522564  2.5915507  0.0442851 21.503  < 2e-16 ***
Ftreat3     0.8690416  2.3846243  0.0446026 19.484  < 2e-16 ***
Ftreat4     0.9276256  2.5284984  0.0441397 21.016  < 2e-16 ***
Ftreat5     0.4077192  1.5033849  0.0489844  8.323  < 2e-16 ***
```

SampleSize 0.0037809 1.0037880 0.0003078 12.285 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
Fblock2	0.8902	1.1233	0.8435	0.9395
Ftreat2	2.5916	0.3859	2.3761	2.8265
Ftreat3	2.3846	0.4194	2.1850	2.6025
Ftreat4	2.5285	0.3955	2.3189	2.7570
Ftreat5	1.5034	0.6652	1.3658	1.6549
SampleSize	1.0038	0.9962	1.0032	1.0044

Concordance= 0.641 (se = 0.004)
Rsquare= 0.164 (max possible= 1)
Likelihood ratio test= 1143 on 6 df, p=0
Wald test = 1083 on 6 df, p=0
Score (logrank) test = 1153 on 6 df, p=0

Here, R has been asked to summarize the data. The top set of comparisons shows pairwise comparisons of each variable. In the case of block 2, it is compared to block 1. All of the listed treatments are compared to the control treatment (Ftreat1). Sample Size of block 1 is compared to sample size of block 2. All of the p-values for these comparisons are significant, showing that they all have a significant effect on the model, as expected. All of the relevant comparisons are significantly different from each other under the hypotheses H0: β(of experimental treatment)=β_0(of control treatment) Ha: β(of relevant experimental treatment)=/=β_0(of control treatment).

This data also shows the increase in chance that an individual will die in each different treatment (exp coef). For example, an individual has a 2.59 higher chance of dying if they are exposed to treatment 2 compared to treatment 1 (control). The data output also provides the 95% confidence interval for that increased risk. This can also be done for the rest if the comparisons.

In R

```
> #COX REGRESSION
>
> cox=read.table("C:\\Users\\Pam\\Google Drive\\Binghamton\\Classes\\Spring 2014\\Biostats with R\\Projects\\scoxreg.txt",header=T)
> attach(cox)
The following objects are masked _by_ .GlobalEnv:

    Day, Dead
> #must install a package before you begin
>   #survival
>   #and splines
>
> #factor the variables in question
> Ftreat=factor(cox$Treat)
> Fblock=factor(cox$block)
> #this tells R that both treatment and block are factors
>
> #Assign the rest of the variables
> Day=cox$Day
> Dead=cox$Dead
> SampleSize=cox$SS
>
> coxreg=coxph(Surv(Day,Dead)~Fblock+Ftreat+SampleSize)
```

```
> coxreg
Call:
coxph(formula = Surv(Day, Dead) ~ Fblock + Ftreat + SampleSize)


               coef exp(coef) se(coef)     z       p
Fblock2    -0.11626      0.89 0.027488 -4.23 2.3e-05
Ftreat2     0.95226      2.59 0.044285 21.50 0.0e+00
Ftreat3     0.86904      2.38 0.044603 19.48 0.0e+00
Ftreat4     0.92763      2.53 0.044140 21.02 0.0e+00
Ftreat5     0.40772      1.50 0.048984  8.32 1.1e-16
SampleSize  0.00378      1.00 0.000308 12.29 0.0e+00

Likelihood ratio test=1143  on 6 df, p=0  n= 6396, number of events= 5608

> #create a model with y=surv, based on day and number dead flies. The
variables that are being compared are block, treatment, and sample size
> #creating this model compares all treatments against the control treatment
(treat 1)
> #it also compares block 1 against block 2
> #in addition, it compares samplesize 1 against sample size 2
> #because the p values of all these parts of the model are less tha 0.05,
this suggests that they all contribute to the model and should be kept in
> #if this was not the case and there was a variable with a p value higher
than 0.05, this means that the model can probably be run without that
variable
>
> summary(coxreg)
Call:
coxph(formula = Surv(Day, Dead) ~ Fblock + Ftreat + SampleSize)

  n= 6396, number of events= 5608

                 coef  exp(coef)   se(coef)      z Pr(>|z|)
Fblock2    -0.1162622  0.8902418  0.0274885 -4.229 2.34e-05 ***
Ftreat2     0.9522564  2.5915507  0.0442851 21.503  < 2e-16 ***
Ftreat3     0.8690416  2.3846243  0.0446026 19.484  < 2e-16 ***
Ftreat4     0.9276256  2.5284984  0.0441397 21.016  < 2e-16 ***
Ftreat5     0.4077192  1.5033849  0.0489844  8.323  < 2e-16 ***
SampleSize  0.0037809  1.0037880  0.0003078 12.285  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
Fblock2       0.8902     1.1233    0.8435    0.9395
Ftreat2       2.5916     0.3859    2.3761    2.8265
Ftreat3       2.3846     0.4194    2.1850    2.6025
Ftreat4       2.5285     0.3955    2.3189    2.7570
Ftreat5       1.5034     0.6652    1.3658    1.6549
SampleSize    1.0038     0.9962    1.0032    1.0044

Concordance= 0.641  (se = 0.004 )
Rsquare= 0.164   (max possible= 1 )
Likelihood ratio test= 1143  on 6 df,   p=0
Wald test            = 1083  on 6 df,   p=0
Score (logrank) test = 1153  on 6 df,   p=0

> #the summary functions shows pairwise t-tests of the different variables,
saying which are significantly different from the controls in each case
```