DSCI 610: Modeling Time-to-Event Data

Required packages

```
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyvers
```

```
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_con
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("survival")
library("ggfortify")
```

Set Working Directory

```
setwd("~/Box/MyDocs/Teaching/Spring/2021/DSCI 610/LectureMaterials/Week 12/Lecture")
```

Introduction

The Cox PH model is a semi-parametric modeling approach. It models the hazard function to determine which combination of potential explanatory variables affect the hazard or risk of an event. The Cox PH model is based on the assumption of proportional hazards.

Comparison of two groups

Suppose patients are randomly assigned to either a standard treatment S or a new treatment N. Let $h_S(t)$ and $h_N(t)$ be the hazards of death at time t for patients on standard and new treatments respectively. According to proportional hazard assumption: $h_N(t) = \psi h_S(t)$. This assumption implies that the survivor functions for individuals on the new and standard treatments do not cross.

- ψ is the hazard ratio or relative hazard
- If $\psi < 1$ the hazard of death at t is smaller for an individual on the new drug relative to an individual on standard drug
- If $\psi > 1$ the hazard of death at t is greater for an individual on the new drug relative to an individual on standard drug

The Cox Proportional Hazards (PH) Model

Suppose T denotes the time-to-event of interest and we observe $X \equiv \min(T, C)$ with δ as the censoring indicator variable. Corresponding to T, the survival and the hazard functions are $S(t)$ and $h(t)$ respectively. In addition, suppose $Z \equiv (Z_1, \ldots, Z_k)$ denotes the observed set of covariates/explanatory variables. For the j-th subject, time-to-event observed data are: $X_j, \delta_j, Z_j \equiv (Z_{1j}, \ldots, Z_{kj})$.

The Cox PH model (Cox, 1972) relates covariates Z to the distribution of T via the hazard function $h(t)$. The Cox proportional hazards model can be written as:

$$h(t|Z_1, \ldots, Z_k) = h_0(t) \exp(\beta_1 Z_1 + \ldots + \beta_k Z_k).$$

More concisely in matrix form,

$$h(t|Z) = h_0(t) \exp(\beta Z),$$

where

$$h_0(t) = h(t|Z = 0)$$

is the unspecified baseline hazard function.

Interpreting Cox PH Model Coefficients

Relative hazard or hazard ratio (HR) of an event comparing Z with $Z = 0$ is:

$$HR = \frac{h(t|Z)}{h(t|Z = 0)} = \exp(\beta Z) = \exp(\beta_1 Z_1 + \ldots + \beta_k Z_k).$$

Note that β_j is the log HR for a unit change in Z_j (continuous covariate), given all other covariates remain constant. That is:

$$\frac{h(t|Z_1, \ldots, Z_j + 1, \ldots Z_k)}{h(t|Z_1, \ldots, Z_j, \ldots Z_k)} = \exp\{\beta_1 * 0 + \ldots + \beta_j * (Z_j + 1 - Z_j) + \ldots + \beta_k * 0\} = e^{\beta_j}$$

For categorical covariate Z_j defined as $Z_j = 1$, say, for new drug and $Z_j = 0$, for standard drug:

$$\frac{h(t|Z_1, \ldots, 1, \ldots Z_k)}{h(t|Z_1, \ldots, 0, \ldots Z_k)} = \exp\{\beta_1 * 0 + \ldots + \beta_j + \ldots + \beta_k * 0\} = e^{\beta_j}.$$

Example: Survival of multiple myeloma patients

Multiple myeloma is a malignant disease characterized by the accumulation of abnormal plasma cells in the bone marrow. Proliferation of the abnormal plasma cells within the bone causes pain and destruction of bone tissue. Patients with multiple myeloma also experience anaemia, haemorrhages, recurrent infections and weakness. Unless treated this health condition is fatal.

Data came from a study carried out at the Medical Center of the University of West Virginia. The objective of the study was to examine the association between the values of certain explanatory variables and the survival time of the patient. Data on survival times of 48 patients suffering from multiple myeloma are included in the dataset myeloma.dat

```
myeloma = read.table("myeloma.dat", header=T)
#table(myeloma$status);table(myeloma$sex);table(myeloma$protein)
#summary(myeloma$age);summary(myeloma$bun);summary(myeloma$ca); summary(myeloma$hb);summary(myeloma$pce
# Create a new age variable
myeloma<- mutate(myeloma, age.50 = age - 50)
myeloma <- mutate(myeloma, sexF = factor(sex))
myeloma <- mutate(myeloma, prF = factor(protein))
head(myeloma)
```

```
##   patient time status age sex bun ca   hb pcells protein age.50 sexF prF
## 1       1   13      1  66   1  25 10 14.6     18       1     16    1   1
## 2       2   52      0  66   1  13 11 12.0    100       0     16    1   0
## 3       3    6      1  53   2  15 13 11.4     33       1      3    2   1
## 4       4   40      1  69   1  10 10 10.2     30       1     19    1   1
## 5       5   10      1  65   1  20 10 13.2     66       0     15    1   0
## 6       6    7      0  57   2  12  8  9.9     45       0      7    2   0
```

The Cox regression model for the ith patient can be written as:

$$h_i(t) = h_0(t)\exp(\beta_1 Age.50_i + \beta_2 Sex_i + \beta_3 Bun_i + \beta_4 Ca_i + \beta_5 Hb_i + \beta_6 Pcells_i + \beta_7 Protein_i)$$

- $Age.50_i$ is the Age of ith patient minus 50
- Sex_i is the sex indicator of i the patient
- Bun_i is the blood urea nitrogen of i the patient
- Ca_i is the serum calcicum of i the patient
- Hb_i is the serum haemoglobin of i the patient
- $Pcells_i$ is the percentage of plasma cells of i the patient
- $Protein_i$ is the Bence-Jones protein indicator of i the patient

Note the baseline hazard $h_0(t)$ is the hazard function of a 50 year old male who has zero values of Bun, Ca, Hb, Pcells, and no Bence-Jones protein.

```
coxph.M_1 <- coxph(Surv(time, status)~ age.50 + sexF + bun + ca + hb + pcells + protein, data=myeloma)
summary(coxph.M_1)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age.50 + sexF + bun + ca +
##     hb + pcells + protein, data = myeloma)
##
##   n= 48, number of events= 36
##
##               coef exp(coef)  se(coef)      z Pr(>|z|)
## age.50   -0.018056  0.982106  0.027833 -0.649 0.516521
## sexF2    -0.249473  0.779211  0.403093 -0.619 0.535985
## bun       0.022661  1.022919  0.006110  3.709 0.000208 ***
## ca       0.013265  1.013353  0.132681  0.100 0.920363
## hb      -0.133017  0.875450  0.068527 -1.941 0.052249 .
## pcells  -0.001359  0.998642  0.006588 -0.206 0.836585
## protein -0.683269  0.504964  0.429395 -1.591 0.111556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age.50     0.9821     1.0182    0.9300     1.037
## sexF2      0.7792     1.2833    0.3536     1.717
## bun        1.0229     0.9776    1.0107     1.035
## ca         1.0134     0.9868    0.7813     1.314
## hb         0.8755     1.1423    0.7654     1.001
## pcells     0.9986     1.0014    0.9858     1.012
## protein    0.5050     1.9803    0.2177     1.172
##
## Concordance= 0.705  (se = 0.048 )
## Likelihood ratio test= 17.53  on 7 df,   p=0.01
## Wald test            = 20.01  on 7 df,   p=0.006
## Score (logrank) test = 25.59  on 7 df,   p=6e-04
```

Interpretation of parameter estimates

- $\hat{\beta}_1 = -0.018$. The difference in log hazard for a patient of 51 years old and a patient of 50 years old is -0.018, keeping all other covariates at fixed values
- $\log h(t|age = 51) - \log h(t|age = 50) = -0.018$ implying $\frac{h(t|age=51)}{h(t|age=50)} = \exp(-0.018) = 0.98$
- $h(t|age = 51) = 0.98h(t|age = 50)$, that is hazard for a 51 year old patient is less than the hazard for a 50 year old patient
- $\hat{\beta}_2 = -0.249$. The difference in log hazard for a female patient and a male patient is -0.249 keeping all other covariates at fixed values
- $\log h(t|sexF = 1) - \log h(t|sexF = 0) = -0.249$ implying $\frac{h(t|sexF=1)}{h(t|sexF=0)} = \exp(-0.249) = 0.78$
- $h(t|sexF = 1) = 0.78h(t|sexF = 0)$, that is hazard for a female patient is less than the hazard for male patient

Test of Hypotheses

$$H_0 : \beta_4 = \beta_6 = 0$$
$$H_a : \beta_4 \neq 0 \;\text{ or }\; \beta_6 \neq 0$$

Likelihood Ratio Test

```
coxph.M_2 <- coxph(Surv(time, status)~ age.50 + sexF + bun +  hb  + protein, data=myeloma)
anova(coxph.M_2,coxph.M_1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ age.50 + sexF + bun + hb + protein
##  Model 2: ~ age.50 + sexF + bun + ca + hb + pcells + protein
##    loglik  Chisq Df P(>|Chi|)
## 1 -98.599
## 2 -98.574 0.0495  2    0.9756
```

$$-2[loglike(reduced) - loglik(full)] = 0.05$$

is not statistically significant at 5% level. Therefore failed to reject the null hypothesis. This indicates that patients' serum calcium and percentage of plasma cells do not have any significant impact on their hazard of survival.

$$H_0 : \beta_1 = \beta_2 = \beta_7 = 0$$
$$H_a : \beta_1 \neq 0 \ \text{ or } \ \beta_2 \neq 0 \ \text{ or } \ \beta_7 \neq 0$$

```r
coxph.M_3 <- coxph(Surv(time, status)~  bun +  hb, data=myeloma)
anova(coxph.M_3,coxph.M_2)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status)
##  Model 1: ~ bun + hb
##  Model 2: ~ age.50 + sexF + bun + hb + protein
##     loglik  Chisq Df P(>|Chi|)
## 1 -100.349
## 2  -98.599 3.5006  3    0.3207
```

```r
summary(coxph.M_3)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ bun + hb, data = myeloma)
##
##   n= 48, number of events= 36
##
##           coef exp(coef)  se(coef)      z Pr(>|z|)
## bun  0.020043  1.020245  0.005816  3.446 0.000569 ***
## hb  -0.134952  0.873758  0.061956 -2.178 0.029391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## bun    1.0202     0.9802    1.0087    1.0319
## hb     0.8738     1.1445    0.7738    0.9866
##
## Concordance= 0.67  (se = 0.052 )
## Likelihood ratio test= 13.98  on 2 df,   p=9e-04
## Wald test            = 16.11  on 2 df,   p=3e-04
## Score (logrank) test = 19.54  on 2 df,   p=6e-05
```

Since the likelihood ratio test is not significant, we failed to reject the null hypothesis and exclude age, sex and Bence-Jones protein indicator from the model.

Interpretation of parameter estimates in the final model

- $\hat{\beta}_3 = 0.02$. For 1 unit change in blood urea nitrogen the difference in log hazard 0.02, keeping serum haemoglobin at fixed value
- $\log h(t|bun+1, hb) - \log h(t|bun, hb) = 0.02$ implying $\frac{h(t|bun+1,hb)}{h(t|bun,hb)} = \exp(0.02) = 1.02$
- $h(t|bun+1, hb) = 1.02h(t|bun, hb)$, that is hazard for a patient with 1 unit increase in blood urea nitrogen is greater than the hazard for a patient with no increase
- $\hat{\beta}_5 = -0.135$. For 1 unit change in serum haemoglobin the difference in log hazard -0.135, keeping blood urea nitrogen at fixed value
- $\log h(t|hb+1, bun) - \log h(t|hb, bun) = -0.135$ implying $\frac{h(t|hb+1,bun)}{h(t|hb,bun)} = \exp(-0.135) = .874$
- $h(t|hb+1, bun) = 0.874h(t|hb, bun)$, that is hazard for a patient with 1 unit increase in serum haemoglobin is less than the hazard for a patient with no increase