STAT40810 — Stochastic Models

Brendan Murphy

Week 7

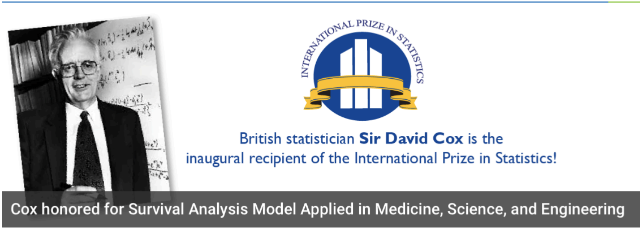
Proportional Hazards Model

Covariates

- A covariate is any quantity recorded in respect of each observation such as age, sex, type of treatment etc.
- The covariates partition the population into groups (with the same values).
- If there are a small number of groups, then the Kaplan-Meier estimates can be compared.
  We did this with the Maintained/Non-Maintained leukemia groups.
- Alternatively, a regression model that uses the covariates to compare survival probabilities can be constructed.
- The most widely used regression model has been the proportional hazards model.
- This model is also known as the Cox model.

International Statistics Award

- Sir David Cox was recently awarded the International Prize in Statistics for this model.



British statistician **Sir David Cox** is the inaugural recipient of the International Prize in Statistics!

Cox honored for Survival Analysis Model Applied in Medicine, Science, and Engineering

http://imstat.org/news/2016/10/20/1477003179504.html

Cox Proportional Hazards Model

- Let observation i have covariates $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
- A survival time T_i follows a Cox proportional hazards model if the hazard function for the ith observation can be written as:

  $$\lambda(t; x_i) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = \lambda_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}),$$

  where β is a vector of regression parameters.
- The function $\lambda_0(t)$ is known as the baseline hazard.
- The covariates then act multiplicatively on the baseline hazard.

Proportional Hazards

- Under the Cox model the hazards of different observations with covariates x_{i_1} and x_{i_2} are in the same proportion at all times:

$$\frac{\lambda(t; x_{i_1})}{\lambda(t; x_{i_2})} = \frac{\lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{i_1 j}\right)}{\lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{i_2 j}\right)} = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{i_1 j}\right)}{\exp\left(\sum_{j=1}^{p} \beta_j x_{i_2 j}\right)}$$

- Hence, the *proportional hazards model* name.
- This means that the general shape of the hazard is determined by the baseline hazard while the exponential term accounts for the differences between observations.

Estimation

- It turns out that we can fit a Cox proportional hazards model without ever specifying the baseline hazard $\lambda_0(t)$;
  this means that we can compare survival for different observations without ever estimating the hazard.
- The Cox proportional hazards model is fitted by maximizing the *partial likelihood*.
- Let $t_1 < t_2 < \ldots < t_K$ be the times at which events are observed.

Estimation 2

- Let $R(t_k)$ denote the set of observations which are at risk just before the kth observed survival time, and for the moment assume that there is only one event at each observed survival time. Let x_k be the covariates for the observation that had an event at time t_j.
- The partial likelihood is:

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta x_k^T)}{\sum_{i \in R(t_k)} \exp(\beta x_i^T)}.$$

Estimation (Ties)

- Suppose we observe more than one event at time t_k.
- Then, we can change the partial likelihood to have the form

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta s_k^T)}{\left(\sum_{i \in R(t_k)} \exp(\beta x_i^T)\right)^{d_k}},$$

where s_k is the sum of the covariates for all observations that had an event at time t_k and d_k is the number of events at time t_k.

Example: Leukemia Data

- We can fit a Cox proportional hazards model to the leukemia data in R.

```
#Load survival package
library(survival)

#Load leukemia data
data(leukemia)

#Fit the Cox PH model to the data
fit <- coxph(Surv(time,status)~x,data=leukemia)
summary(fit)
```

- The code is very like previous regression code in its syntax.

Results: Leukemia

```
Call:
coxph(formula = Surv(time, status) ~ x, data = leukemia)

  n= 23, number of events= 18

                 coef exp(coef) se(coef)     z Pr(>|z|)
xNonmaintained 0.9155    2.4981   0.5119 1.788   0.0737 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

               exp(coef) exp(-coef) lower .95 upper .95
xNonmaintained     2.498     0.4003    0.9159     6.813

Concordance= 0.619  (se = 0.073 )
Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.06581
Wald test            = 3.2  on 1 df,   p=0.07371
Score (logrank) test = 3.42  on 1 df,   p=0.06454
```

Hazards and Survival

- Let's recall the connections between hazard and survival.

$$h(t) = -\frac{d}{dt} \log S(t).$$

- Thus,

$$S(t) = \exp\left(-\int_t^\infty h(s)ds\right).$$

- Thus,

$$\text{Hazard} \uparrow \quad \Rightarrow \quad \text{Survival} \downarrow$$

$$\text{Hazard} \downarrow \quad \Rightarrow \quad \text{Survival} \uparrow$$

Breast Cancer

- The German Breast Cancer Study Group completed a randomized trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients.
- The data is available in the TH.data package in R.
- We can model the data using a Cox proportional hazards model.

```
#Load survival package
library(survival)

#Load the GBSG2 data
library(TH.data)
data("GBSG2")
GBSG2

# Change tumor grade to unordered for interpretation
GBSG2$tgrade <- factor(GBSG2$tgrade,ordered=FALSE)

#Fit a Cox proportional hazards model
fit <- coxph(formula = Surv(time, cens) ˜ ., data = GBSG2)

# Examine the fit
print(fit)
summary(fit)
```

Results 1

```
Call:
coxph(formula = Surv(time, cens) ˜ ., data = GBSG2)

                  coef exp(coef)  se(coef)     z       p
horThyes     -0.346278  0.707316  0.129075 -2.68 0.00730
age          -0.009459  0.990585  0.009301 -1.02 0.30913
menostatPost  0.258445  1.294915  0.183476  1.41 0.15895
tsize         0.007796  1.007827  0.003939  1.98 0.04779
tgradeII      0.636112  1.889121  0.249202  2.55 0.01069
tgradeIII     0.779654  2.180718  0.268480  2.90 0.00368
pnodes        0.048789  1.049998  0.007447  6.55 5.7e-11
progrec      -0.002217  0.997785  0.000574 -3.87 0.00011
estrec        0.000197  1.000197  0.000450  0.44 0.66131

Likelihood ratio test=105  on 9 df, p=0
n= 686, number of events= 299
```

Results 2

```
Call:
coxph(formula = Surv(time, cens) ~ ., data = GBSG2)

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

             exp(coef) exp(-coef) lower .95 upper .95
horThyes        0.7073     1.4138    0.5492    0.9109
age             0.9906     1.0095    0.9727    1.0088
menostatPost    1.2949     0.7723    0.9038    1.8553
tsize           1.0078     0.9922    1.0001    1.0156
tgradeII        1.8891     0.5293    1.1591    3.0788
tgradeIII       2.1807     0.4586    1.2885    3.6909
pnodes          1.0500     0.9524    1.0348    1.0654
progrec         0.9978     1.0022    0.9967    0.9989
estrec          1.0002     0.9998    0.9993    1.0011

Concordance= 0.692  (se = 0.018 )
Rsquare= 0.142   (max possible= 0.995 )
Likelihood ratio test= 104.8  on 9 df,   p=0
Wald test            = 114.8  on 9 df,   p=0
Score (logrank) test = 120.7  on 9 df,   p=0
```

Drug Treatment

A proportional hazards regression model was used to study the duration that recovering drug addicts spend undergoing treatment in a clinic.

The variables included in the study are:

Time (t_i)	-	Time undergoing treatment
Clinic (c_i)	-	0=Clinic A; 1=Clinic B
Status (s_i)	-	Censored or not
Prison (p_i)	-	1=Prison record; 0=No prison record
Dose (d_i)	-	Dose of medication administered at the clinic

The model was fitted using a statistical software package and the following output was produced.

```
          coef exp(coef) se(coef)     z        p
clinic -1.0099     0.364  0.21489 -4.70 2.6e-06
prison  0.3266     1.386  0.16722  1.95 5.1e-02
dose   -0.0354     0.965  0.00638 -5.54 2.9e-08
```