Applied Stats II, PS4

Luke Duggan

03 April 2022

1 Question One

First we read in the data. Briefly, this is data on infant mortality in Sweden in the mid-late 19th century. Observations are grouped in threes: a case where the mother died during or shortly after childbirth and two cases having explanatory variables identical or very similar to the first case where the mother did not die. The three variables we are interested in are:

1. `exit`: The age of the infant in days at death. This is right-censored, i.e., set to 365 if the infant didn't die in their first year.
2. `age`: The mother's age in years at birth.
3. `sex`: Whether the infant was a boy or a girl.

Next, we fit a Cox proportional hazards model with mother's age and an individual's sex as explanatory variables:

```
model <- coxph(Surv(enter, exit, event) ~ age + sex)
summary(model)
```

The output from this model is:

```
Call:
coxph(formula = Surv(enter, exit, event) ~ age + sex)

  n= 105, number of events= 21

           coef exp(coef) se(coef)      z Pr(>|z|)
age    -0.04044   0.96037  0.04507 -0.897    0.370
sexboy -0.48518   0.61559  0.44224 -1.097    0.273

       exp(coef) exp(-coef) lower .95 upper .95
age       0.9604      1.041    0.8792     1.049
sexboy    0.6156      1.624    0.2587     1.465

Concordance= 0.586  (se = 0.058 )
Likelihood ratio test= 1.99  on 2 df,   p=0.4
Wald test            = 2  on 2 df,   p=0.4
Score (logrank) test = 2.03  on 2 df,   p=0.4
```

Let's interpret these results.

Let's begin with the interpretation of the coefficients. In a proportional hazards model, all units are assumed to have the same underlying "hazard function": a function expressing the rate of failure at time t conditional upon t being at least as great as some fixed time T. This is a bit abstract: a very useful concrete illustration is in the analysis of mortality, where if time t represents, say, age in years, the hazard function gives the proportion of persons who die aged t years among the class of persons aged T years and older, where $t >= T$.

The coefficients can be interpreted as follows: exponentiating the coefficient gives the *multiplicative* effect on the hazard function of a one unit increase in the explanatory variable (holding other variables constant). Happily, R provides the exponentiated coefficients in the output. Thus, we can say: a mother's being one year older at the birth of an infant has a multiplicative effect of about 0.96 on the hazard function, and the infant's being male has a multiplicative effect of about 0.62.

Note that both these estimates are less than one: that is, an increase of one unit in the explanatory variable is associated with a *decrease* in the proportion of deaths among infants. I am not very well-educated on this subject but *a priori* it does not seem unbelievable that infants born to younger mothers might be at greater risk of dying within a year, or that there was a higher rate of female infant mortality in these regions at this point in history.

Note, however, that neither of our estimates is statistically significant. R's built-in measure of significance uses an asymptotically valid normal distribution: we can also do likelihood ratio tests for either coefficient, which are asymptotically chi-square distributed.

```
drop1(model, test = "Chisq")
```

The output is:

```
Model:
Surv(enter, exit, event) ~ age + sex
       Df    AIC     LRT Pr(>Chi)
<none>    171.25
age     1 170.12 0.87104   0.3507
sex     1 170.42 1.16584   0.2803
```

We see the results are still insignificant. In a fuller investigation, we would probably start with a "maximal" model including all covariates on which we have data and perform stepwise deletion tests to exclude variables which are both statistically insignificant and theoretically unlikely to be relevant.