THE MULTIPLE COX MODEL

A) Introduction to running descriptives

0) Premises

We would like now to include multiple predictors in a multiple Cox Regression model. Leaping in and running the model without looking at the variables is hazardous.

1) Practice in R: Getting to know your data

Earlier we only looked at the outcome variable – death – and a single predictor – age and then ethnic group.

The latter had some missing values and we want to incorporate more variables into the Cox model, so we need to summarize each of them first, to see if they too have any hidden traps.

The multiple regression model is just an extension of a simple regression model to incorporate multiple predictors. Let's say for now that we want to fit a model with five predictors because we have a good prior knowledge that they are important. Let's say we want to run a descriptive analysis for these variables:

- Age (cont.)
- Gender (binary)
- Prior opd appointments missed ("prior_dnas")
- ethnic group
- COPD (chronic obstructive pulmonary disease)

"summary" for cont variables, and "table" + "exclude=NULL" for the categorical ones.

From the analysis of the copd variable we see that nearly 3 in four patients had missed no appointments, but nearly three percent had missed five or more, with a maximum of ten.

2) How can we handle this variable in the regression model?

There are mainly 3 possibilities:

- Pretending that it is continuous and assume a linear relation with the outcome: The first option is typically preferred for ORDINAL variables with lots of values, and in particular if the relation is linear this will be the best solution.
- Categorize it using each of the values as a category
- Categorize it but combine some values.

Categorizing there is the risk of losing information. Having a few categories with lots of patients in each is, however a good way of getting around the problem of a non-linear relation. In this case, trying to assess whether the relation is linear is made harder by the sparse data for people with more than five or six missed appointments.

It is also possible to fix for example the ethnic group so the "NAs" can be turned into category 8 and those patients gets anyway included in the model.

When we run the multiple regression model (see Week 3 of the R code) Bwe note that "gender1" and "ethnicgroup1" do not appear in the above output. This is because they are the reference categories for their respective variables. By default, R sets the reference category as the lowest value for the variable, which is 1. This happens to correspond to males and to white ethnicity, but you can easily change the reference category.

3) Interpreting the output from multiple Cox model

We have just run the multiple Cox regression model and we want to understand and interpret the results.

I) Gender = 2 = females = have a hazard ratio of 0.78 = the confidence interval goes from 0.65 to 0.97, p=0.007 -> association btw women and a lower hazard of mortality. The women in this sample lived longer after their admission and their hazard is 22% lower than that of the males. Where do we get the 22% from? 100-78 = 22. A hazard ratio of 0.78 means 22% lower hazard.

II) Similarly patients with COPD have a hazard ration of 1.15 relative to those without COPD their hazard is 15% higher than that of patients without COPD. However in this case the confidence interval goes from something below one to something above 1 and a p value = 0.188, so that is not statistically significant. We do not have enough evidence for a relation btw COPD and death.

III) Ethnic group has given us hazard ratios for groups: 2, 3, 8 and 9 all relative to group 1 (= white ppl). The only statistically significant group is group 3, namely the subcontinent, with a lower hazard. Patients with Indian ethnicity after admission lived longer than white ppl.

IV) Hazard ratio for previous appointment missed which is 1.18. For each appointment missed the hazard increase by 18%, a lot. As for what concerns age the assumption is that there is a linear relation btw the predictor and the hazard for death (we will see later how to test this).

We could also dichotomize this variable. If the relation is not linear we could use categories: a simple look at the width of the confidence intervals for each category will us whether we have stretched our sample too thin and we have too few patients in our sample.

It is important to remember that each of these associations is adjusted for the other predictors in the model.

B) Introduction to Non-convergence

1) Non-convergence

We try running the cox model including the “quantile” variable. "quintile" is the neighborhood socio-economic status. If this is known, values range from 1 (most affluent) to 5 (least affluent). It's nationally weighted by population, meaning that 20% of England's population live in areas of each quintile. This message will appear

```
Warning message:

In fitter(X, Y, strats, offset, init, control, weights = weights,  :
```

```
Loglik converged before variable  4,5,6,7,8 ; beta may be infinite.
```

By asking for the model summary we see that the coefficients and particularly the standard errors for quintile are all huge. The answer can be found by simple tabulation: only 4 patients have quintile zero. This means invalid quintile, for instance when the postcode can't be mapped to a small geographical area and therefore to a socio-economic status measure. Four patients in a category can sometimes be enough to get the model to work, but there's another problem. Of those 4 patients with quintile zero, no one died. That itself might not be a problem, but we've let R choose the reference category by default, and it has chosen quintile zero. All the other 5 hazard rations are relative to this tiny group of paitents in which no one died.

2) How can we fix this problem?
   - I) Change the reference category
     By default R chooses the first (lowest) category to be the reference. Since the quintile measures socio-economic status or deprivation and we are usually interested in the effect of lower status compared with higher status, it makes sense to set higher status, say quintile = 1, as the reference category.
   - II) Combine categories
     Quintile zero is an artificial category, meaning that the patients' postcode or other geographical area identifier was missing or could not be linked to the national socio-economic status file. These patients could be for example overseas, homeless, etc.. Since they are so few we could just ignore them.
   - III) Exclude patients
     Best option if combining categories doesn't make any sense.
   - IV) Drop the offending variable
     Best option if combining categories doesn't make any sense.

Remember that the problem of non-convergence can happen in *any* kind of regression and that these simple tricks can also work there.