Survival analysis – part 2
Practical in R

Getting started:

- **Log in to moodle http://www.ucl.ac.uk/moodle** using your UCL username and password
- Find the course: **Research Methods for Quantitative Data**
- Go to **Survival analysis** and download the dataset **elsa_cf.rdata** to your workspace.
- Start ***R***

Reminder of our objectives and the data:

For this practical on survival analysis, we are interested in determinants of mortality in older adults who take part in the English Longitudinal Study of Ageing. Specifically, we want to assess the link between **cognitive function** and **mortality risk** in this aging population, and if this relationship exists, whether it reflects confounding by lifestyle, socioeconomic status or chronic disease (in particular cardiovascular disease).

The first assessment (wave 1) of ELSA was performed in 2002/2003. Through linkage with the Office for National Statistics mortality data, we have data on vital status with date of death over the follow-up. For this specific analysis, the follow-up ends in March 2013.

A cognitive function score, composite of memory, executive functioning and processing speed was measured at baseline (wave 1). All other potential predictors of mortality present in this dataset are measured at baseline. Full list of variables:

- **cf1**: cognitive function score composite of memory, executive functioning score and processing speed. Possible range 0-164.
- **alcohol1**: daily consumption of alcohol 0=no, 1=yes
- **cigst1** : smoking status: 0=never 1= former 2=current smoker
- **educ1**: Higher qualifications 0= high (e.g., university degree or higher), 1= intermediate (secondary school), 2= low (completed no more than compulsory schooling)
- **totwq5_bu1**: Quintile of wealth 1=highest 5=lowest
- **physact1:** Physical inactivity at w1 (0=active 1=inactive)
- **sex:** Sex of participant 0=male; 1=female
- **age1**: Age at w1
- **cancer1**: Ever been diagnosed with cancer w1
- **chd1**: Ever been diagnosed with coronary heart disease at w1

- **iintdtm1:** Month of date of interview w1 (entry in study)
- **iintdty1**: Year of date of interview w1 (entry in study)
- **dodmnth:** Month of date of death
- **dodyr:** Year of date of death
- **dead**: Mortality status 0=alive; 1=dead
- **time**: Survival time in years (time from entry in study up to death or censoring)

In this second part, the objectives are:

- To build Cox proportional hazards regression models
- To assess confounding and interactions
- To test the proportional hazards assumption

1. Quick description of the data

If you wish to remind yourself of the data, you could re-calculate some descriptive statistics as we did in Survival analyses practical 1. Otherwise, you could continue to question 2:

First look at some descriptive statistics.

How many people died? What is the mean follow up time for mortality? How many person-years?

2. Cox regression models

We found in the previous practical that there was no evidence of an association between gender and time to death (p=0.97; log rank test).
Now, let's fit some Cox proportional hazards models to the data. The command for this is coxph. For example:

```
cox1 <- coxph(Surv(time, death)~ age1 + factor(sex), data=elsa_cf)
```

First, run a model including **sex** and then run a model including **both sex and age**. Fill in the values in the Table below:

	Hazard ratio (female vs male)	95% CI	p value
Unadjusted			
Adjusted for age group			

- What happens to the hazard ratio for sex when you adjust for age? Why do you think the hazard ratio changes?

Now, we want to see if the information on morbidities, education or lifestyle are predictors of mortality. We can progressively add potential predictors to a baseline model including age and sex. We can perform a likelihood ratio test to assess whether the addition of a variable improves the model fit.

```
# Full model #
cigst1_=factor(cigst1)
educ1_=factor(educ1)
cox3 <- cph(Surv(time, death) ~ age1 + sex + chd1 + cancer1 + alcohol1 +
                        physinact1 + educ1_ + cigst1_ ,   data=elsa_cf, method="breslow")
```

- What set of covariates (i.e. predictors) will you likely include in your final model when assessing the effect of cognitive function? What are the hazard ratios for these factors and how do you interpret them?

Now, let's see if cognitive function predicts survival. Include adjustments for age, sex and your chosen covariates in the model.

- Which measure of cognitive function do you want to include (categorical or continuous)?
- What is/are the hazard ratio(s) (95% CI) for an increase of cognitive function score?
- What would you conclude from this?

Now we are interested to know whether this association can vary according to other factors. What is this called?

```
# Create agegr (age in categories) and stratify analysis
elsa_cf$agegr[elsa_cf$age1 < 60] <- 1
elsa_cf$agegr[elsa_cf$age1>=60 & elsa_cf$age1 < 70] <- 2
elsa_cf$agegr[elsa_cf$age1 >= 70] <- 3

#Create subsets#
elsaage1 <-subset(elsa_cf, elsa_cf$agegr==1)
elsaage2 <-subset(elsa_cf, elsa_cf$agegr==2)
elsaage3 <-subset(elsa_cf, elsa_cf$agegr==3)

#Run models in each strata#
coxf.age1 <- coxph(Surv(time, death) ~ cf1_10 + age1  + sex + … ,
          data=elsaage1, method="breslow")
coxf.age2 <- coxph(Surv(time, death) ~ cf1_10 + age1  + sex + … ,
           data=elsaage2, method="breslow")
coxf.age3 <- coxph(Surv(time, death) ~ cf1_10 + age1  + sex + …,
           data=elsaage3, method="breslow")
summary(coxf.age1)
summary(coxf.age2)
summary(coxf.age3)
```

- For instance, include an interaction between age and cognitive function test for statistical significance using likelihood ratio test. What do you conclude?

When an interaction is found significant, it can be affecting the direction of the effect, or its magnitude (or both). It is more of a problem when the direction of the association is different according to a third factor (e.g. positively associated in men and negatively associated in women). We need to stratify our analysis (splitting the data into groups and run the model in each group) to assess whether the effect of cognitive function on mortality differs substantially according to age.
Conduct stratified analysis by age groups. What can you conclude? Is the interaction affecting the direction of the association?

3. Checking the proportional hazards assumption

Our first analyses suggest an effect of cognitive function independently of other factors. However the Cox model we use implies that the hazards are proportional over time. We have to check if this assumption holds.

- Remember that we plotted the Kaplan-Meier graph for different cognitive function score. What do you notice? Graphically, does the PH assumption seem to hold?

Another (and more refined) way to assess proportional hazards is to plot minus the log cumulative hazard.

```
# Plot 1- log (-log survival)
coxq <- cph(Surv(time, death) ~ cf1+age1 + sex + chd1,
        data=elsa_cf, x=TRUE,y=TRUE, method="breslow")

survplot(coxq, cf1=c(38,45,50,57), age1=mean(age1), sex=0, chd1=0,
      logt=TRUE, loglog=TRUE, xlim=c(-4, 2.5), ylim=c(-10, 0))
```

If the proportional hazards assumption is met, then the lines should be approximately parallel. Is this the case?

We can also carry out a formal test of the proportional hazards assumption for **cf1**. One way to do this is to introduce an interaction between **cf1** and time. For instance:

```
#Create a time-dependent model and test for interactions with time for all covariates
time.dep <- coxph( Surv(time, death)~age1+sex+cf1,   data=elsa_cf, method="breslow",
na.action=na.exclude)
time.dep.zph <- cox.zph(time.dep, transform = 'log')
time.dep.zph
```

This allows the effect of all covariates on survival to vary by time since entry to study.

- Test for the proportional hazard assumption in the fully adjusted model. Is there evidence that the effect of cognitive function varies by time?

- Is there any evidence to suggest that hazards are non-proportional for physical activity?

If there is evidence that the proportional hazards assumption is not met for **physical activity,** then the next step might be to re-run the analyses stratified by **physical activity**. This estimates separate baseline hazard functions for men and women thus allowing them to be non-proportional. To do this, use the **Strata** option:

Does this alter the hazard ratio for cognitive function?

Another way to check for the PH assumption is a test based on the Schoenfeld residuals. It gives a test of the deviance from the PH assumption for all variables included in the model. We can also plot these residuals and assess graphically: the line should be horizontal.

```
# Plot the residuals
plot(time.dep.zph)
```

4. Plotting estimates of survival functions after running Cox regression model

If we want to visualize the predicted survival estimated by the Cox regression, we have several ways to do it: plot the survival, the hazard (=1-survival) or the cumulative hazard. For this we use the option survplot.