Exercise 3.2 Cox, logistic and conditional logistic regression

1. This exercise makes use the dataset **full_cohort.dta**, which is a simulated data set of 50,000 cohort members, where the outcome of interest is time from enrolment to coronary heart disease with event indicator. The variables recorded for this cohort include:

t (time from enrolment into the cohort to coronary heart disease, in years)
event (event indicator: 1=coronary heart disease, 0=censored)
Age (age, in years)
Gender (1=Male, 0=Female)
Chol (cholesterol in mg/dL)
HDL (high-density lipoprotein in mg/dL)
SBP (systolic blood pressure in mmHg)
Treat (antihypertensive treatment status: 1=Yes, 0=No)
Smoke (smoking status: 1=Yes, 0=No)

(i) Create a categorical variable for age with the following intervals: ≤49, 50-59, 60-69, 70-79, ≥80.

(ii) For each of the other continuous covariates (**Chol**, **HDL**, and **SBP**) center them at their mean values (i.e. subtract the mean from each value).

(iii) Run a Cox regression analysis of the whole cohort with these new variables created in part (i) and (ii) and the remaining binary variables (Gender: 1=Male, 0=Female, Treat: 1=Yes, 0=No, Smoke: 1=Yes, 0=No).

(iv) Assume the investigators wanted to study coronary heart disease by year 10 by sampling an exclusive case-control dataset and conducting a logistic regression. Draw an exclusive 1:1 case-control sample ***at year 10*** that is matched on sex and age category. Save a copy of this data set for part (v).

(v) Run a conditional logistic regression analysis of the data from (iv) and compare the results to the full cohort in part (iii).

2.

(i) Run Cox regression analyses of the whole cohort to obtain the crude HR for **Gender** and the HR adjusted for age-category.

(ii) Draw a 1:5 *nested* case-control sample (i.e. 5 controls for each case) using simple random sampling (use a seed!) and save a copy of the data before proceeding. Run a conditional logistic regression of **Gender** using your nested case-control sample and compare the results with the crude HR for gender obtained from the whole cohort Cox regression.

(iii) Draw a nested case- control sample that is matched on age category. Compare the OR for gender from a conditional logistic regression analysis to the age-adjusted HR from the Cox regression of the whole cohort.

(iv) Use the nested case-control data from (ii) to run the full model in Question 1 (i.e., all the variables listed in part (iii), categorical age, mean centered **HDL**, **Chol** and **SBP**, and binary treatment, gender and smoking), and compare the results to the Cox regression of the whole cohort.

Hints for Stata users

Before any survival analysis, you must use the **stset** command to declare the survival data. **stcox** runs a Cox regression on data that has been "stset"

The **ccmatch** command in Stata will match cases and controls on specified variables

The **sttocc** command in Stata can be used to sample a nested case-contol study. Before using the sttocc command, you need to set the time and failure variables using **stset.** Remember to also set a seed.

Hints for R users

The **Epi** package in R has a command **ccwc** to sample *nested* case-control data from a cohort. As there is no non-nested version, one could use this command by setting the **exit** time (i.e., either the event or censoring time) to the same value for all subjects such that they are in the same "risk set". Running **ccwc** on the modified data will select controls from all non-cases in the cohort (since they have the same exit time) or from controls with the same (exit time and) confounding profile, if you had matched on additional confounders.

The **clogit** command (from the survival package) allows you to run a conditional logistic regression. Strata must be specified as the case-control "Set", which was generated by the **ccwc** command when drawing the nested case-control sample by adding it as a predictor in the model "…`+strata(Set)`".