Homework 2

Due Friday, 22 February 2019 at 10am (before lab)

Read all questions carefully before answering. You may work in small groups of no more than 3 individuals and turn in a single assignment (and everyone in the group will receive the same grade). Work through the entire assignment individually first, then come together to discuss and collaborate. Please type your responses, **show your work, and please keep answers brief.**

Directions:

Use the dataset `frmgham_recoded.Rdata` and code provided herein to explore the relationship between smoking status at baseline and time to death in the Framingham cohort.

```r
load("frmgham_recoded.Rdata")
library(survival)

#CREATE A SINGLE-RECORD DATASET (retain 1st observation)
frmgham_recoded <- frmgham_recoded[which(frmgham_recoded$period == 1),]
```

The relevant variables for this analysis are:

- `time_yrs` (time of entry into study)
- `timedth_yrs` (time of death)
- `death` (indicator of death [=1] or censored [=0])
- `cursmoke` (indicator of current smoking status: yes (1) vs. no(0))
- `age` (age in years)
- `sex` (variable denoting male (1) or female (2); **use 1 as referent category**)
- `educ` (educational status, nominal categories 1-4; **use 1 as referent category**)

Adapting the code presented in the lecture, and the additional piece below, complete the following tasks:

Describing survival data (Questions 1-2, part of 7)

- Referring to the code from the lecture notes, **plot the Kaplan-Meier estimate of the survival function** for each smoking category (using the variable `cursmoke`) for these data:

- Using the code below, **calculate the number of events and number of person-years in each exposure group**:

```r
# The "scale" option set to 1 tells pyears() that the time is already in yrs
# (default is to divide time by 365.25)
py.smoke <- pyears(Surv(timedth_yrs,death)~cursmoke, frmgham_recoded,
                   scale=1)[c("event","pyears")]
simplify2array(py.smoke) # Make it pretty
```

Proportional hazards modeling (Questions 3-8)

- Referring to the code from the lecture notes, **use the logrank test to determine if there are any differences in survival between the smoking groups.**
- **Using a Cox proportional hazards regression model,** estimate the association between current smoking status (at baseline) and time to death. Estimate 2 models:
  - An **unadjusted** model (only including smoking status), and
  - An **adjusted** model that also includes age (continuous), sex (binary) and education (4-category) in this model. (For nominal categorical variables, you may need to use the `factor()` operator in the formula as demonstrated in class.)
- Estimate a model with an interaction between **linear follow-up time** and each of the covariates in the model.

Questions:

1. Using the Kaplan-Meier plots, graphically assess the relationship between baseline smoking status and time to death. **Briefly interpret what you see.** In 1-2 sentences describe the limitations of this approach. [include the graph, labeled **Figure 1**] **(10 points)**
2. **Referring to the code from lecture**, are you able to calculate the overall median survival time in this case? If so, provide an estimate of this quantity, if not, describe why and provide an estimate of a percentile of survival time. Interpret the quantity that you estimated. **(15 points)**
3. Answer the following questions about the log-rank test: **(10 points total)**
   i) Describe the specific hypothesis that the logrank test is considering here.
   ii) What do you conclude from this test (use 5% significance criteria)? What is the limitation of the inference that you obtain from the log-rank test?
4. Answer the following questions about the Cox models estimated above: **(20 points total)**
   i) Why do we use specialized methods for survival analysis (instead of linear or logistic regression, for example)?
   ii) What are the advantages of the Cox model over other survival analysis methods? What is a potential disadvantage of the Cox model?
   iii) What assumptions, if any, does the standard **Cox** proportional hazards model make?
   iv) Compare the test of the smoking-mortality association between the log-rank test and the likelihood ratio test from the unadjusted Cox proportional hazards model. What do you observe? Between these two analytic approaches, which one would you prefer, and why?
5. Write the equation for the log-hazard function for the *adjusted* model you estimated. **Clearly define all parameters in the model. (15 points)**
6. Using the model you specified in the previous question, show that the **hazard ratio** comparing current smokers to non-smokers, holding all other covariates constant, is $\exp(\beta_1)$ where β_1 is the coefficient on the smoking indicator. (**Hint:** Start by showing the log-hazard for smokers and the log-hazard for non-smokers and use the fact that the log of the hazard ratio is the difference between two log-hazards.) **(10 points)**
7. Complete the following table. How would you interpret the parameter estimate that compares smokers to non-smokers in the **adjusted model**? What measure of association common in epidemiologic research does this correspond to? **(10 points)**

Table 1: Crude and adjusted hazard ratio (HR) estimates of the association between baseline smoking status and mortality. Framingham Cohort Study. 1948-1972, Framingham, MA.

Smoker	Events	Follow-up Time (yrs)	crude HR (95% CI)	adj. HR (95% CI)
No				
Yes				

8. Is there evidence for a violation of the proportional hazards assumption in any of the variables?

Indicate how you arrived at your conclusion. Describe how you would account for any noted violations in the proportional hazards assumption. **(10 points)**