# HW 2

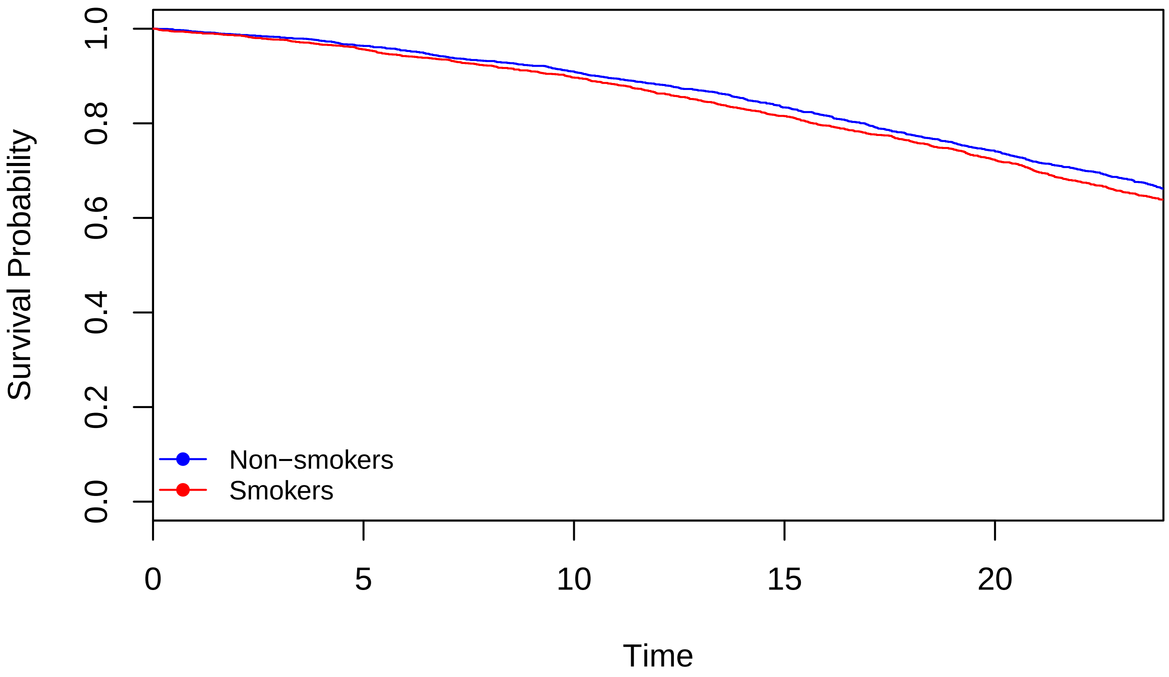
*Steph Holm and Shelley Facente*

*February 19, 2019*

## Question 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)

**Figure 1**



*Upon visual inspection, it appears that people who smoked at baseline have slightly lower probability of survival during the 24 years of follow up, compared with people who were non-smokers at baseline. However, the difference between the two appears small, so we would want to more carefully investigate whether there are meaningful differences between the two groups.*

## Question 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)

*The exact median survival can not be calculated, as more than 50% of the population was stil alive (aka without the event) at the end of the follow up time. This can be seen in our graph above. The time when 0.75 of the population is surviving can however be calculated (see Table 1) for the whole sample, or stratified by smoking status.*

*However, we could potentially model the data out further if we make an assumption of parametricity. We note that without further information this is a big assumption, as we are extrapolating past the collected data. To do this, we used the number of events in each smoking group (or in the full sample), and the person-time of follow up in each of those groups to calculate $\lambda$ the rate parameter for each of these groups. If we then assume the classic exponential survival function, $S(t) = e^{-\lambda * t}$ , we can solve for the median survival, $S(t^*) = 0.50$ as well as the time when 75% are still surviving, $S(t^*) = 0.75$. Those results are in the Table below.*

Table 1: Estimates of Median Survival for the Framingham cohort from both Kaplan-Meier survival fit as well as from classic exponential survival function. Framingham Cohort Study. 1948-1972, Framingham, MA.

| Smoker | $t^*_{75}$ from KM | $t^*_{75}$ from parametric $S(t^*)$ | $t^*_{50}$ from KM | $t^*_{50}$ from parametric $S(t^*)$ |
|---|---|---|---|---|
| No | 19.44 | 17.62 | NA | 42.46 |
| Yes | 18.61 | 16.22 | NA | 39.09 |
| Total (combined) | 19.09 | 16.91 | NA | 40.75 |

## Question 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.

*The null hypothesis, $h_0$, is that the number at risk at any time, $j$, multiplied by the probability of death at that time is the same for both groups (those who were current smokers at study enrollment versus not.)*

*If we use the indicator $k = 1$ for the smoking group and $k = 0$ for the nonsmoking group, this could be stated mathmatically as:*

$$\frac{n_{1j} * d_j}{n_j} = \frac{n_{0j} * d_j}{n_j}$$

### 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?

*Using 5% as our cut-off for statistical significance, we would conclude that these curves are* not *significantly different as the p-value is* `0.088`*, meaning that if the null hypothesis is true (there is truly no difference in survival between groups) we have a* `8.8` *percent chance of generating data such as these by chance. The log-rank test provides no measure of effect, so you can only infer the level of heterogeneity between groups, and nothing more.*

## Question 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)?

*We use specialized models for survival analyses because they deal with time; time-to-event must be defined as the probability of an event happening at any particular time* `t` *given that the event has not already happened prior to that time* `t`.

### ii)

What are the advantages of the Cox model over other survival analysis methods? What is a potential disadvantage of the Cox model?

*The advantage of the Cox model is that it does not specify the form of the baseline hazard (and thus the distribution of T) and estimates it non-parametrically, so we don't have to worry about mis-specification of this baseline hazard. However, a potential disadvantage is that you lose some statistial efficiency by using only the part of the likelihood that doesn't contain* $h_0(t)$.

### iii)

What assumptions, if any, does the standard Cox proportional hazards model make?

*Like other proportional hazards models, the Cox proportional hazards model assumes that the effects of the exposure and co-variates of interest are constant over the entire follow-up (i.e. the hazards are proportional over the entire time interval). It also assumes no delayed entry, truncation, or other types of censoring, unless explicitly handled in design/analysis.*

### 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?

*By running the log-rank test, we get a* $\chi^2$ *of 2.9 with a p-value of* `0.088`, *which tells us there is not a statistically significant difference between smokers and non-smokers with respect to mortality in this sample. However, the log-rank test does not provide any sort of effect measure. For this, we can use the unadjusted Cox proportional hazards model, which gives us a hazard ratio of 1.091, with a 95% CI (.9872, 1.205). This confidence interval still demonstrates that there is no statistically significant difference between the two groups; however, it is preferred because we have more useful information about the hazard ratio itself, as well as direct calculation of the* $\beta$*s from the model.*

## Question 5:

Write the equation for the log-hazard function for the adjusted model you estimated. **Clearly define all parameters in the model.** (15 points)

*The equation for the model is:*

$$ln[h(t|x)] = ln[h_0(t)] + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6$$

*where*

– $x_1$ *is an indicator of whether or not someone was smoking at enrollment*
– $x_2$ *is an indicator of sex, dichotomized into male and female*
– $x_3$ *is an indicator corresponding to educational status category 2*
– $x_4$ *is an indicator corresponding to educational status category 3*

- $x_5$ *is an indicator corresponding to educational status category 4*
- $x_6$ *is age in years*

*using the results from our model, the equation then becomes:*

$$ln[h(t|x)] = ln[h_0(t)] + 0.339x_1 + -0.55x_2 + -0.038x_3 + -0.168x_4 + -0.325x_5 + 0.096x_6$$

## Question 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 $e^{\beta_1}$ where $\beta_1$ 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)

*Using the equation from the previous question, we can calculate that the log-hazard for smokers (holding other covariates constant) is:*

$$ln[h(t|x)] = ln[h_0(t)] + 0.339 * 1 + -0.55x_2 + -0.038x_3 + -0.168x_4 + -0.325x_5 + 0.096x_6$$

*and the comparable log-hazard for non-smokers is:*

$$ln[h(t|x)] = ln[h_0(t)] + 0.339 * 0 + -0.55x_2 + -0.038x_3 + -0.168x_4 + -0.325x_5 + 0.096x_6$$

*The hazard ratio generated via the model specified in the previous question is the log-hazard of smokers divided by the log-hazard of non-smokers:*

$$\frac{e^{ln[h_0(t)]+0.339*1+-0.55x_2+-0.038x_3+-0.168x_4+-0.325x_5+0.096x_6}}{e^{ln[h_0(t)]+0.339*0+-0.55x_2+-0.038x_3+-0.168x_4+-0.325x_5+0.096x_6}}$$

*which when simplified is equal to*

$$= \frac{e^{0.339*1}}{e^{0.339*0}} = \frac{e^{0.339}}{1}$$

*exactly the same as the anti-log of the $\beta_1$ coefficient on the smoking indicator (with the $\beta_1$ coefficient =* `0.339`).

## Question 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 2: 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 | 762 | 46675.20 | *ref* | *ref* |
| Yes | 788 | 44440.38 | 1.091 (0.987, 1.205) | 1.404 (1.262, 1.562) |

*In the adjusted model, smokers have roughly a 40% increased risk of mortality during the 24 years of follow-up as nonsmokers, when adjusting for sex and educational level. While this parameter estimate is a hazard ratio, it corresponds to an incidence density ratio, a common measure of association in epidemiologic research.*

## Question 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)

*The summary of our analysis to assess the proportional hazards assumption for this dataset found no significant interactions between time and our exposure variable (smoking at baseline), nor age or any levels of education (p-values for the coefficients on the interaction of time with smoking, age, and educational levels 2, 3, and 4 were* 0.42, 0.27, 0.76, 0.99, *and* 0.73, *respectively). Therefore, we can conclude the the proportional hazards assumption has not been violated for these variables.*

*However, we do have a violation of the proportional hazards assumption for sex, a confounder variable (p-value for the coefficient on the interaction of time and sex is* 0.03*). We could account for that by including time interactions with that variable in our model, or alternatively by estimating a stratified Cox proportional hazards model for each of our defined sex categories separately. If the violation were in our exposure variable, we would use a model including the time interactions and report time-specific effects (i.e.* $HR(t) = e^{\beta + \gamma x t}$*).*