Laboratory Assignment 3

Katherine Wolf

November 25, 2019

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 (of your choice). Interpret the quantity that you estimated. **(20 points)**

3. Answer the following questions about the log-rank test: **(10 points total)**

   (a) Describe the specific null and alternative hypotheses that the log-rank test is considering here.

   (b) What do you conclude from this test (use 5% significance criteria)? List a 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)**

   (a) Why do we use specialized methods for survival analysis (instead of linear or logistic regression, for example)? (Hint: See readings from Vittinghoff et al. 2012 text.)

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

   (c) What assumptions, if any, does the standard **Cox** proportional hazards model make?

   (d) 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 functions, terms (covariates), and parameters in the model. (20 points)**

6. 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 (years)	Crude HR (95% CI)	Adjusted HR (95% CI)
No				
Yes				

7. Based on the model that included covariate-by-time interactions, is there evidence for a violation of the proportional hazards assumption in any of the variables? Indicate how you arrived at your conclusion. In 1-2 sentences describe in general how you would account for any violations in the proportional hazards assumption (ignoring whether or not there were significant differences here). **(10 points)**

Question 1

Participants who were current smokers at the start of the study appear to have lower probabilities of survival at all time points than participants who were not current smokers at the start of the study, and the gap between the groups appears to widen slightly over the course of the study until its end at 24 years, when the probability of survival for the smokers was 63.87% and the probability of survival for the nonsmokers was 66.18%.

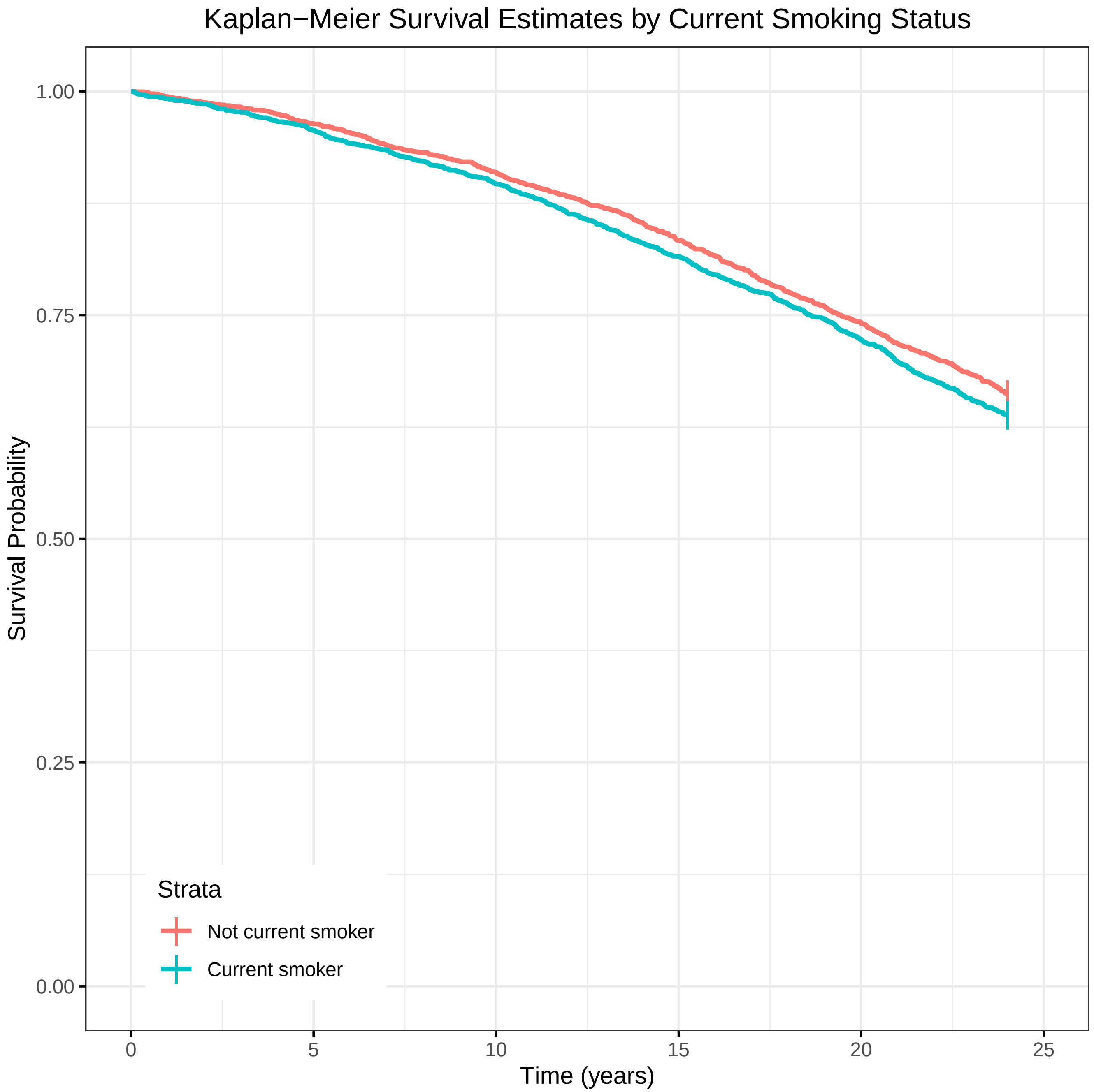
Although Kaplan-Meier estimates of survival curves offer the advantages of

- estimating survival probabilities accurately even in the presence of censored observations during the study period, and
- being nonparametric, i.e., making no assumptions about the distribution of survival times in the cohort,

major limitations of simple visual comparison of the graphs of the survival curves of the two groups include

- the failure to account or control for the effects of any covariates that might have confounded or mediated the relationship between smoking and survival time, and
- the absence of a formal test of the hypothesis that the survival probabilities are systematically different between the groups over the study period (versus to the null hypothesis the apparent difference on the graph between the two curves resulted from chance alone).

Figure 1. Kaplan-Meier Survival Estimates by Current Smoking Status



Question 2

We cannot calculate the overall median survival time in this case because less than half of this sample had the event (death) by the end of the study. (This dataset has no censored observations except for those right-censored at the end of the study, so the the number of participants observed to have died in this dataset equals the number who did actually die during the study period.) In other words, over half of the observations are right-censored, and thus the true median survival time is somewhere in that majority of unobserved full survival times and unknown to us. As over 25% of the cohort did die over the course of the study, however, we can estimate the **25th percentile of survival time**, the first time point at which at least 25% of the cohort had died and 75% or less of the cohort remained alive, which was **19.09** years.

(For another dataset containing censored observations throughout the study period instead of only at the end, a more accurate wording of the above might be ”the first time point at which the probability of being dead in this cohort is estimated to be at least 25%. Both of those interpretations assume that all participants start counting follow-up time at the same time, whether in real or statistical time, which appears true for this dataset since all study entry times are 0.)

In other words, in this specific cohort with no censored observations until the end of the study and identical times of entry among all partiicpants, that last person who needed to die to bring the total percentage of the cohort dead to or above 25% did so 19.09 years after the start of the study.

(The 25th percentile of survival time among current nonsmokers was 19.44 years, and the 25th percentile of survival time among current smokers was 18.61 years. In other words, among members of the cohort who were nonsmokers at the beginning of the study, the first time at which at least 25% of them had died was 19.44 years after the start of the study. Similarly, the first time at which at least 25% of the members of the cohort who were smokers at the beginning of the study had died was 18.61 years after the start of the study.)

Question 3

Sub-Question 1

The null hypothesis for the log-rank test here is that the survival curve for the smokers is equivalent to that for the nonsmokers. More specifically, the null hypothesis is that the probability of survival for a nonsmoker is the same as the probability of survival for a smoker at all times during the study period, or formally:

$$H_0 : S_1(t) = S_2(t)\ \forall(t)$$

where $S_k(t)$ is the survival probability for a member of group k at time t. (Here, $k = 1$ designates the nonsmoking group and $k = 2$ designates the smoking group.)

The alternative hypothesis is that the survival curve for at least one group is different than the others at some time during the study period, or, formally:

$$\exists(t)\ S_1(t) \neq S_2(t)$$

Sub-Question 2

The p-value of **0.09** indicates that we cannot reject the null hypothesis of no difference between the smoker and nonsmoker survival curves at the 5% significance level.

One limitation of the log-rank test is that, although it allows for censoring, it requires that censoring to be non-informative, i.e., unrelated to survival probability. (It also requires survival probabilities to be the same regardless of the entry date into the study, i.e., that survival probabilities are not affected by factors only present at specific dates when some participants were in the study and some were not.) Another limitation is that, in the case of a significant result, the log-rank test does not identify the time or times at which the survival curves differ. A third limitation of the log-rank test is its inability to distinguish, in the presence of more than two survival curves and a significant result, which curves are different without further testing. A fourth limitation is that the log-rank test does not accommodate adjustment for covariates.

Question 4

Sub-Question 1

Why do we use specialized methods for survival analysis (instead of linear or logistic regression, for example)?

We use specialized methods for survival analysis in order to fully incorporate the information contained in varying follow-up times. We cannot use linear regression (with survival time as the outcome) due to the right-censoring of the events of interest (e.g., disease incidence or death) frequently present in survival data, which often leaves many survival times unknown. Worse, the missing survival times are censored informatively, meaning that whether a survival time is unknown is associated with the outcome of survival time, i.e., the longest times are the ones most likely to be missing. Thus using linear regression on the known survival times would systematically underestimate them. Adjusting for follow-up time can partially account for variations in follow-up time but imposes assumptions on the relationship between follow-up time and event risk (e.g., that follow-up time is associated with risk in some kind of linear manner).

We cannot use logistic regression, either, even though we are interested in the binary presence or absence of an event, because logistic regression ignores variation in follow-up times, which arises both from variation in participants' entry times into the study and from differences in survival times among participants. Doing a simple logistic regression with event presence or absence at the end of the study as the outcome effectively ignores the information contained in the participant surivival times. Moreover, if systematic differences exist in entry times between the groups of interest (for example, exposed participants generally entered the study later than unexposed participants), logistic regression could produce particularly biased results. Pooled logistic regression, which looks at binary event outcomes at several times over the course of the study instead of only at the end, can help account for variations in follow-up times, but is more appropriate when events are discovered at regularly spaced intervals at which outcomes can be evaluated (e.g., study visits conducted at the same intervals for all participants) and still fails to incorporate the specific information garnered from having exact event times, producing less accurate estimates of event risk than survival analysis methods that can fully accommodate exact event times.

Sub-Question 2

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

Advantages: The Cox model does not require us to specify the distribution of the failure times T, also known as the baseline hazard. Thus the Cox model is more robust than parametric proportional hazards models because bias cannot arise from misspecification of the distribution of failure times.
Disadvantage: One potential disadvantage of the Cox model is the decreased precision (and increased variance) in the Cox model compared to fully parametric survival models that either specify the distribution of failure times accurately or are relatively robust to misspecifications.

Sub-Question 3

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

The standard Cox proportional hazards model assumes

- That the covariates have multiplicative associations with the hazard (as they reside in the parametric portion of the model, in an exponent relative to the hazard), as opposed to additive or other associations;
- That the hazard ratios are constant over time (the *proportional hazards assumption*), i.e., that the hazard in one group is a constant multiple of the hazard of any other group throughout time;
- That survival times are independent among study participants;
- That measurements and times are accurate in the dataset; and
- That any censoring is uninformative, i.e., not associated with survival time.

Sub-Question 4

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?

Both the log-rank test and the likelihood ratio test from the unadjusted Cox proportional hazards model give nearly identical results, i.e., a chi-square statistic of 2.91 on one degree of freedom ($p = 0.088$). Thus neither rejects the null hypothesis that no systematic difference in survival times exists between current smokers and nonsmokers at the 5% significance level. Were I not allowed to adjust for any other covariates, like here, I would prefer the log-rank test, as it is fully nonparametric and does not presume any prior relationship between the Kaplan-Meier survival curves on which it is based and is thus more robust to a wider variety of distributions of survival times among groups. The Cox proportional hazards model, in contrast, assumes that hazard ratios (derived from relationships among survival curves) remain constant over time between any two groups. Were I allowed to adjust for covariates, however, I would ultimately use the likelihood ratio test from the Cox proportional hazards model or some variation of it, as the Cox model can accommodate a wider variety of covariates and thus will ultimately produce more informed and accurate estimates of the relationship between a categorical exposure and survival time outcomes.

Question 5

Equation:

$$\log(h(t|\mathbf{x},\boldsymbol{\beta})) = \log[h_0(t)] + \beta_1\mathbb{I}(x_1 = 1) + \beta_2 x_2 + \beta_3\mathbb{I}(x_3 = 1) + \beta_4\mathbb{I}(x_4 = 2) + \beta_5\mathbb{I}(x_4 = 3) + \beta_6\mathbb{I}(x_4 = 4)$$

Definitions:

- $\log(h(t|\mathbf{x},\boldsymbol{\beta}))$ is the natural logarithm of the instantaneous rate (hazard) of death at time t for a participant who has survived up to that time t, given the covariates in **x** and coefficients in $\boldsymbol{\beta}$;
- $\log(h_0(t))$ is the baseline instantaneous rate (hazard) of death at time t for a participant who was not a current smoker, was zero years old, was male, had 0-11 years of education at the first examination, and survived up to time t;
- **x** is a vector of covariates containing
  - x_1, an indicator variable for smoking corresponding to the dataset variable ‘cursmoke‘, with $x_1 = 1$ indicating that the participant was a current smoker at the first examination and $x_1 = 0$ indicating that the participant was not;
  - x_2, a continuous variable for age of the participant at the first examination in years;
  - x_3, an indicator variable corresponding to the recoded variable ‘female‘ derived from the dataset variable ‘sex‘, with $x_3 = 1$ indicating a female participant (originally ‘sex = 2‘) and $x_3 = 0$ indicating a male participant (originally ‘sex = 1‘);
  - x_4, a four-category categorical variable corresponding to the dataset variable ‘educ‘, where
    - $x_4 = 1$ corresponds to the dataset variable ‘educ = 1‘, indicating that the participant had 0-11 years of education (in this equation, this condition is the reference condition and subsumed in the baseline hazard term $\log[h_0(t)]$ tercept term $x_0 = 1$);
    - $x_4 = 2$ corresponds to the dataset variable value ‘educ = 2‘, indicating that the participant had a high school diploma or GED;
    - $x_4 = 3$ corresponds to the dataset variable value ‘educ = 3‘, indicating that the participant had some college or vocational school;
    - $x_4 = 4$ corresponds to the dataset variable value ‘educ = 4‘, indicating that the participant had a college degree or more;
- $\mathbb{I}()$ is the formula for an indicator variable, which equals 1 if the equation inside the parentheses is true and 0 if it is not (used to create indicator variables for each non-reference level of the categorical variables in **x**);
- $\boldsymbol{\beta}$ is a vector of coefficients including
  - β_1 is the log hazard ratio comparing the hazard of death for a participant who was a current smoker at the first examination to the hazard of death for a participant who was not, holding the values of all other variables constant;
  - β_2 is the log hazard ratio for a one-year increase in age, holding the values of all other variables constant,
  - β_3 is the log hazard ratio comparing the hazard of death for a participant who was a female (in the numerator) to the hazard of death for a participant who was male (in the denominator) at the first examination, holding the values of all other variables constant;
  - β_4 is the log hazard ratio comparing the hazard of death for a participant who had a high school diploma or GED (in the numerator) to the hazard of death for a participant who had 0-11 years of education (in the denominator) at the first examination, holding the values of all other variables constant;
  - β_5 is the log hazard ratio comparing the hazard of death for a participant who had some college or vocational school (in the numerator) to the hazard of death for a participant who had 0-11 years of education (in the denominator) at the first examination, holding the values of all other variables constant;
  - β_6 is the log hazard ratio comparing the hazard of death for a participant who had some college or vocational school (in the numerator) to the hazard of death for a participant who had a college degree or more (in the denominator) at the first examination, holding the values of all other variables constant.

Question 6

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 (years)	Crude HR (95% CI)	Adjusted HR (95% CI)
No	762	46675.20	1. (reference)	1. (reference)
Yes	788	44440.38	1.091 (0.987—1.205)	1.404 (1.262—1.562)

The hazard of death during the study for a participant who was a current smoker at the start of the study was 1.404 times (95% confidence interval: 1.262, 1.562) that of a participant who was not a current smoker at the start of the study after adjusting for age, sex, and level of education. Hazard ratios, in general, correspond to incidence rate ratios. (See the following reference for a discussion of why hazard ratios do not always approximate relative risks well: *Sutradhar R and Austin PC. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. Annals of Epidemiology. 2018;28(1):54-57. doi:10.1016/j.annepidem.2017.10.014.*) That said, in the absence of competing events, rate ratios can approximate risk ratios.

Question 7

The sex variable appears to violate the proportional hazards assumption. To evaluate violations of the proportional hazards assumption in the variables, I reran the adjusted model, adding interactions with time for all the covariates:

$$\begin{aligned}\log(h(t|\mathbf{x},\boldsymbol{\beta})) = \log[h_0(t)] + \beta_1\mathbb{I}(x_1=1) + \beta_2 x_2 + \beta_3\mathbb{I}(x_3=1) + \beta_4\mathbb{I}(x_4=2) + \beta_5\mathbb{I}(x_4=3) + \beta_6\mathbb{I}(x_4=4) +\\ \beta_7\mathbb{I}(x_1=1)\times t + \beta_8 x_2\times t + \beta_9\mathbb{I}(x_3=1)\times t + \beta_{10}\mathbb{I}(x_4=2)\times t + \beta_{11}\mathbb{I}(x_4=3)\times t + \beta_{12}\mathbb{I}(x_4=4)\times t\end{aligned} \tag{1}$$

All functions, covariates, and parameters have the same defnitions as in question 5, with the addition of the coefficients $\beta_7, \beta_8, \ldots, \beta_12$ to model possible interactions with time. I then produced the table below, which shows a significant interaction at the 5% significance level ($p = 0.026$) for the interaction between sex and time. I then conducted a global likelihood ratio test comparing the log likelihood of the adjusted model with the time interaction to the same model without the time interaction, obtaining a significant chi-square value of 4.148 with one degree of freedom ($p = 0.042$), indicating that the interaction improves the model accuracy.

To account for a violation of the proportional hazards assumption: If the violation were in a confounder (e.g., in sex here), I would either include the time interaction(s) with the confounder in the model to account for its effects on the exposure-outcome association, or, if I could categorize the confounder easily and did not care about evaluating its association with the outcome, I would estimate a model stratified by the levels of that confounder to allow for a different baseline hazard function in each stratum. If the violation were in the exposure of interest (e.g., in current smoking here), I would include all the time interactions in the model and report time-specific instead of overall effects.

Table 3: Estimates of associations between baseline smoking status and mortality including adjustment for potential interactions with time. Framingham Cohort Study. 1948-1972, Framingham, MA.

Covariate	Beta	e^Beta	Standard Error	z	p
Current Smoker	0.44	1.55	0.13	3.34	0.00
Age	0.09	1.09	0.01	10.90	0.00
Female Sex	-0.29	0.75	0.13	-2.24	0.02
Education: High School	0.00	1.00	0.15	0.02	0.99
Education: Some College	-0.17	0.84	0.19	-0.89	0.37
Education: College Degree	-0.26	0.77	0.22	-1.16	0.25
Current Smoker by Time	-0.01	0.99	0.01	-0.81	0.42
Age by Time Interaction	0.00	1.00	0.00	1.10	0.27
Female by Time Interaction	-0.02	0.98	0.01	-2.22	0.03
Education: High School by Time	-0.00	1.00	0.01	-0.30	0.76
Education: Some College by Time	0.00	1.00	0.01	0.02	0.99
Education: College Degree by Time	-0.00	1.00	0.01	-0.34	0.73