Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Homework #5 Key

1. A randomized controlled trial is conducted to evaluate whether an experimental chemotherapy ($X = 1$) improves survival rates in leukemia patients as compared to standard of care ($X = 0$). In this example, we're pretending that we know the *true* survivor functions in each group, given as follows (there is no censoring in this problem, and we are not estimating anything):

$$\begin{aligned} S(t|X=0) &= \mathrm{P}(T > t|X=0) = \exp(-t) \\ S(t|X=1) &= \mathrm{P}(T > t|X=1) = \exp(-3t/5), \end{aligned}$$

where t is time in years. **You will not need calculus for any of the below problems**.

(a) Determine the probability of surviving beyond one year in each treatment group.

Ans: $\mathrm{P}(T > 1|X = 0) = \exp(-1) = 0.368$. $\mathrm{P}(T > 1|X = 1) = \exp(-3/5) = 0.549$.

(b) Determine the probability of dying within two years in each treatment group.

Ans: $\mathrm{P}(T \leq 2|X = 0) = 1 - \exp(-2) = 0.865$; $\mathrm{P}(T \leq 2|X = 1) = 1 - \exp(-2 \times 3/5) = 0.699$.

(c) Determine the median survival time in each group.

Ans: $S(t|X = 0) = 0.5 \Rightarrow t = 0.693$ years; $S(t|X = 1) = 0.5 \Rightarrow t = 1.16$ years.

(d) Which group has greater three-year restricted mean survival time? How do you know? As an *optional* problem, feel free to compute these (requires light calculus).

Ans: Since $S(t|X = 1) > S(t|X = 0)$ for all t, the subgroup $X = 1$ must have greater restricted mean survival time as compared to the subgroup $X = 0$. Optionally, you may also show that:

$$\begin{aligned} \mu_R^0 &= \int_{[0,3]} S(t|X=0) = \int_{[0,3]} e^{-t} dt = 0.950; \\ \mu_R^1 &= \int_{[0,3]} S(t|X=1) = \int_{[0,3]} e^{-3t/5} dt = 1.39. \end{aligned}$$

(e) State the cumulative hazard functions, $\Lambda(t|X = 0)$ and $\Lambda(t|X = 1)$.

Ans: $\Lambda(t|X = 0) = -\log(S(t|X = 0)) = t$; $\Lambda(t|X = 1) = -\log(S(t|X = 1) = 3t/5$.

(f) State the hazard functions, $\lambda(t|X = 1)$ and $\lambda(t|X = 0)$. Is the proportional hazards assumption satisfied? If so, state the hazard ratio (comparing group $X = 1$ to group $X = 0$).

Ans: $\lambda(t|X = 0) = 1$; $\lambda(t|X = 1) = 3/5$. The proportional hazards assumption is trivially satisfied because both of the hazard functions are constant. The hazard ratio is 3/5.

2. To identify predictors of survival, a group of investigators reviewed their experience with primary malignant tumors of the sternum. They classified patients as having either low-grade (26 patients) or high-grade (11 patients) tumors. The data are provided in the file `sternum.csv`.

   (a) Compute the Kaplan-Meier curve **for high-grade tumor patients** by brute force (that is, not relying on the survival-analysis capabilities of statistical software). Show your table of computations that lead to your estimate, like the one that appears in the course notes. Based on your computation, what is the estimated median survival time in this group? What is the estimated two-year survival rate?

   Ans: Indeed, this is a little bit of a pain, but it's helpful to understand the concept. Mirroring the approach we took in class

Time	At risk	Events	1 – Hazard	$\widehat{S}(t)$
0	11	0	1.0000	1.0000
3	10	2	0.8000	0.8000
4	8	1	0.8750	0.7000
7	7	1	0.8571	0.6000
8	6	0	1.0000	0.6000
9	5	1	0.8000	0.4800
12	4	1	0.7500	0.3600
14	3	1	0.6667	0.2400
22	2	1	0.5000	0.1200
26	1	1	0.0000	0.0000

   Based on this table, we have that:

   - The median survival time in this group is estimated to be 9 months (it is also acceptable to say that is estimated to be somewhere between 8 and 9 months).
   - The two-year survival rate is estimated to be 12.0%.

   (b) Using Stata, create Kaplan-Meier curves for each group; estimate the median survival and two-year survival rates, along with 95% confidence intervals. Compare your results for high-grade tumor patients to the by-hand results you computed in (a).

   Ans: The Kaplan-Meier curves are shown on the following page. The median survival time is estimated to be 9 months in the high-grade group (95% CI: [3, 22]) and 205 months in the low-grade group (the lower bound of the 95% CI is 27). The two-year survival rate is estimated to be 12.0% in the high-grade group (95% CI: [0.66%, 40.8%] and 79.4% in the low-grade group (95% CI: [57.4% 90.9%]). These results square with those seen in part (a).

   (c) Test whether there is a difference in the survival distributions between the two patient groups. What test did you use, and what did you conclude?

   Ans: The log-rank test reveals sufficient evidence that the survival distributions are not the same between the two patient groups ($p < 0.001$).

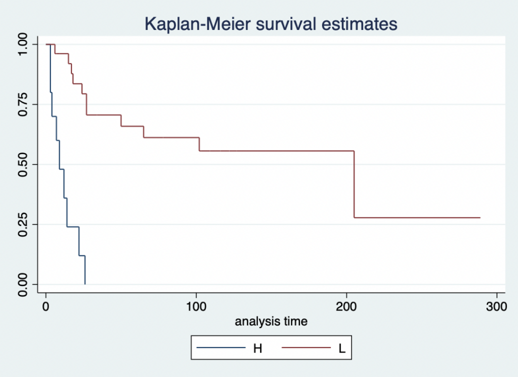

Figure 1: Kaplan-Meier curves for Problem 2.

3. A placebo-controlled clinical trial of methotrexate was conducted in a cohort of 285 primary biliary cirrhosis patients. Load the data set `pbc.csv`.

   (a) Perform an analysis to evaluate the whether the hazard of death differs between patients receiving of methotrexate or placebo, adjusting for baseline albumin level. Examine/test the proportional hazards assumption and briefly comment on your results.

   Ans: This study does not provide sufficient evidence of a hazard ratio comparing the methotrexate and placebo groups that is different from one, adjusting for baseline albumin ($p = 0.692$). We estimate the hazard of death to be 22.2% higher in the methotrexate group as compared to a placebo group of the same baseline albumin; based on a 95% confidence interval, this estimate would not be deemed atypical if in truth the hazard of death were between 54.8% lower and 231% higher in the methotrexate group. The figure below presents a plot of the log-analysis time against the estimated negative log-log-survival probability. The non-constant difference may suggest a potential departure from the assumption of proportional hazards. A test of the Schoenfeld residuals does not provide sufficient evidence of such a violation ($p = 0.966$.)

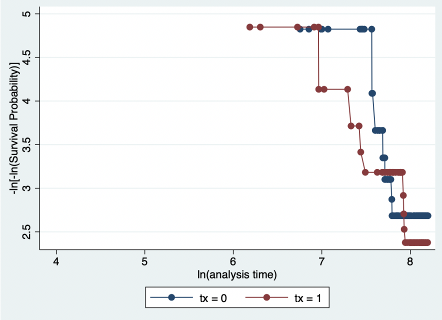

Figure 2: Diagnostic plot for Problem 3.

   (b) Using the model you fit in part (a), state point estimates of the (treatment-group adjusted) hazard ratio for baseline albumin levels of 3.0, 3.5, and 4.0, using an albumin level of 2.0 as the reference level (note that albumin is continuous and it should be treated as such).

   Ans: The estimated hazard ratios are 0.134, $(0.1339)^{1.5} = 0.0490$ $(0.1339)^2 = 0.0179$.

4. Load the data set `mri.csv`. We will examine the association between smoking and death in an adult population. Note that the variable `packyrs` denotes smoking history; those with values equal to zero correspond to those who have never smoked. Create a binary variable for any history of smoking for this problem (0 if no prior smoking history, and 1 if any prior smoking history).

   (a) On one plot, provide Kaplan-Meier curves for those with and without a history of smoking.

   Ans: The Kaplan-Meier plot is shown below. **The smoking variable should only have 734 observations; most of you did not code this correctly**.

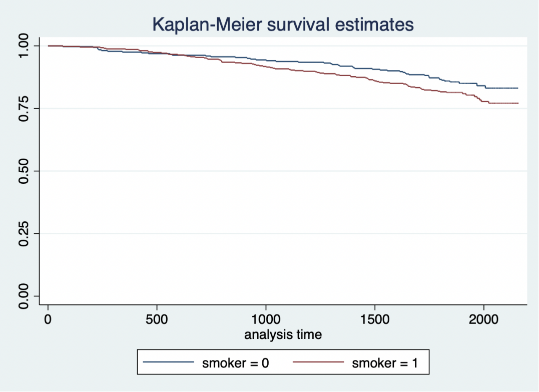

Figure 3: Diagnostic plot for Problem 3.

   (b) Perform an analysis to evaluate whether any history of smoking is associated with the hazard of death, adjusting for age.

   Ans: This study provides sufficient evidence that smoking is associated with a higher hazard of (all-cause) death ($p = 0.016$). Comparing subgroups of the same age by differing in their smoking status, we estimate that the smokers have a 55.9% higher hazard as compared to nonsmokers. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the hazard among smokers were between 8.79% and 123% higher.

   (c) Repeat part (b), this time using cardiovascular death as the outcome. Briefly describe the relative advantages and disadvantages of the cause-specific outcome *cardiovascular death* as compared to *all-cause* death.

   Ans: This study provides sufficient evidence that smoking is associated with a higher hazard of (cardiovascular) death ($p = 0.023$). Comparing subgroups of the same age by differing in their smoking status, we estimate that the smokers have a 87.5% higher hazard as compared to nonsmokers. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the hazard among smokers were between 9.20% and 222% higher.

   Though cardiovascular death may be considered more clinically relevant in some cases, it suffers from the challenge of competing risks; moreover, all-cause death encompasses at least as many events as cause-specific deaths and hence can have higher power to detect a fixed alternative hypothesis.

5. Phase-I randomized trials often spend considerable efforts calibrating treatment doses. Consider a Phase-I study of long-term prednisone use for subjects with chronic pulmonary disorders deemed to be at extraordinarily high risk of death (these data are provided in the file `prednisone.csv`). Doses are given as 20, 40, and 60 mg over certain periods of time; note that the doses change over time within individuals randomized to the treatment group. Included in these data are the subject ID, randomization group, as well as the interval dose and cumulative dose, along with an indicator of whether that observation time corresponded to an observed death.

   (a) Perform an analysis to evaluate whether cumulative prednisone dose is associated with hazard of death. Please make a comparison between subgroups differing in their cumulative dose by, say, 20 milligrams, rather than a single milligram.

   Ans: This analysis does not provide sufficient evidence of an association between long-term prednisone use and hazard of death in this patient population ($p = 0.114$). Comparing subgroups differing in cumulative dose by 20 mg, we estimate the hazard to be 10.0% lower in the group receiving a higher dose as compared to the group receiving the lower dose. Based on a 95% confidence interval, this estimate would not be considered surprising if in truth the hazard in the higher dose group were between 21.2% lower and 2.58% higher.

   (b) Consider the reduced version of the data set, `prednisone-reduced.csv`, which excludes the granular details of time-dependent treatment. Perform an analysis based on the total cumulative dose and compare your results (a); very briefly comment on the limitations of this approach.

   Ans: This analysis provides evidence of an association between long-term prednisone use and hazard of death in this patient population ($p = 0.017$). Comparing subgroups differing in cumulative dose by 20 mg, we estimate the hazard to be 9.50% lower in the higher dose group. Based on a 95% confidence interval, this estimate would not be judged unusual if the hazard were between 1.77% and 16.6% lower in the higher dose group. This analysis is limited in that it does not take into account the time-dependent nature of treatment; it in effect assumes that the cumulative dose is constant over time and that subjects remain on the same hazard trajectory over time. In this analysis, our strength of evidence regarding efficacy is likely overstated, though the point estimate is similar.

   (c) Now, again considering the *reduced* data set, perform an analysis based whether *any* treatment is associated with the hazard of death (any treatment is defined by the `Group` variable). Comment on how your results compare to (a) and (b); very briefly comment on the limitations of this approach.

   Ans: This analysis does not provide evidence of an association between long-term prednisone use and hazard of death in this patient population ($p = 0.305$). Comparing subgroups differing in whether they receive prednisone, we estimate the hazard to be 56.6% lower in the group receiving prednisone. Based on a 95% confidence interval, this estimate would not be judged unusual if in truth the hazard were between 91.2% lower and 114% higher in the prednisone group. Not only is the treatment considered as time-stable, but the variation in dose magnitude is also lost in this analysis; the effect is estimated to be much larger.