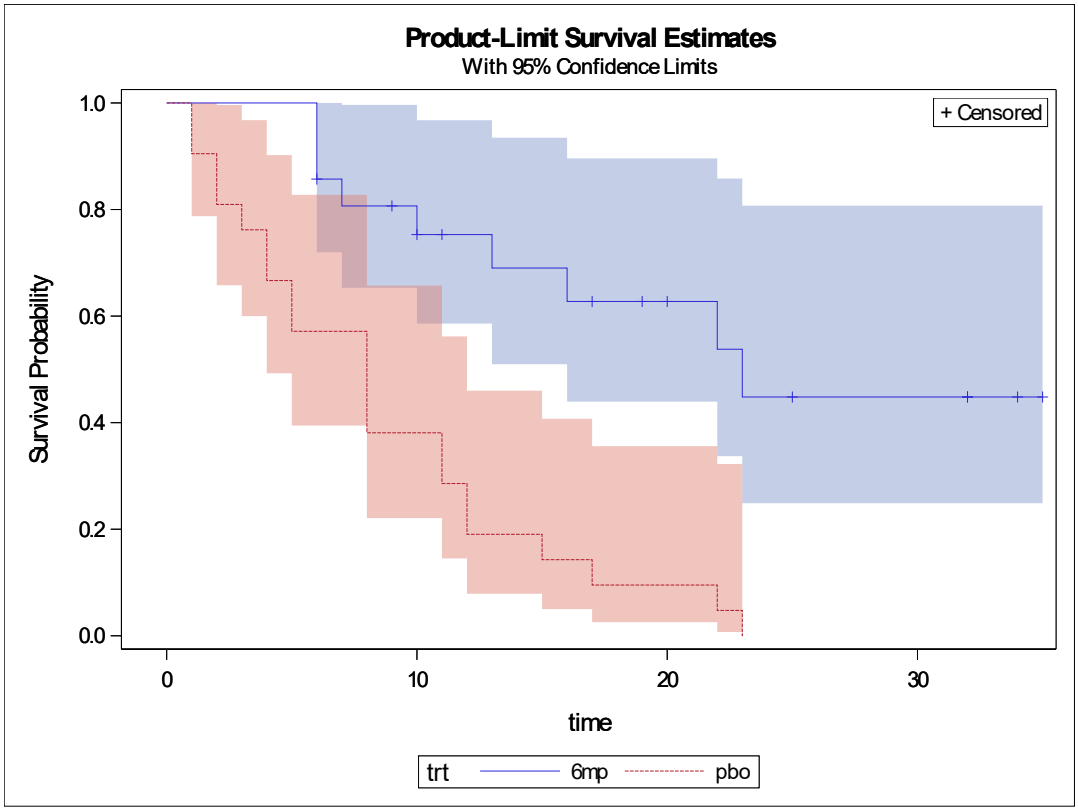
Matt Isaac
Biostatistics Homework 5
Due: 02 April 2018

1. Is there any difference in the time to relapse between the two treatment groups?
    a. (3 points) Answer this question using a nonparametric two-sample test, ignoring the presence of censoring but using all observations. State the name of the test, the p-value, and the conclusion in context.
        i. Test: Wilcoxon Rank-Sum (aka Mann-Whitney) test
        ii. P-value: 0.0042
        iii. Conclusion: Since our p-value is significant ($p < 0.05$), we have detected a significant difference in the recovery time of those who were given the 6-MP treatment and those who were given the placebo.
    b. (3 points) Answer this question using a nonparametric two-sample test, with all censored observations dropped. State the name of the test, the p-value, and the conclusion in context.
        i. Test: Wilcoxon Rank-Sum (aka Mann-Whitney) test
        ii. P-value: 0.1883
        iii. Conclusion: With the censored observations dropped, we obtained a p-value > 0.05. This indicates that there is not a significant difference in recovery time of those who were given the 6-MP treatment and those who were given the placebo.
    c. (6 points) Answer this question using a nonparametric method to appropriately account for the presence of censoring. State the name of the test, the p-value, and the conclusion in context.
        i. Test: Log-Rank test (since we only have a categorical variable)
        ii. P-value: < 0.0001
        iii. Conclusion: After accounting for censoring in the data, we have obtained a highly statistically significant result ($p < 0.0001$). This indicates that there is a highly significant difference in relapse time between patients who were given the 6-MP treatment and patients who were not.
    d. (3 points) Briefly discuss why the conclusion in part (c) is more meaningful than the conclusions in parts (a) and (b).
        i. The conclusion in part (c) is more meaningful because it is the correct way to analyze these data. The time to relapse values mean different things depending on the value of the relapse (censoring) variable. The tests in parts (a) and (b) did not account for these differences. Performing a proper survival analysis does account for this and uses all the information available. Thus, we can 'trust' the result obtained in part (c) more than the results obtained in parts (a) and (b).

e. (2 points) Report survival curves (with 95% confidence bands) for both treatment groups as fit by the model in part (c).



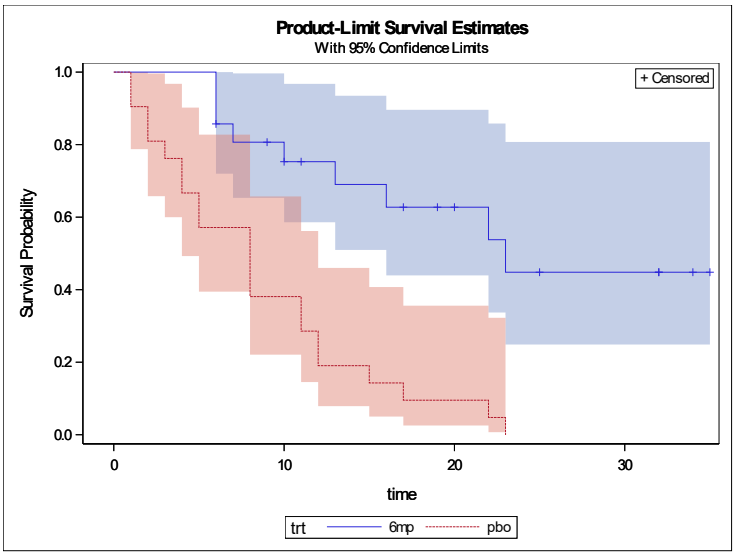
f. (4 points) Provide a brief interpretation of the plot reported in part (e)
   i. On average, the group which received the 6-MP treatment had a higher probability of 'survival' (no relapse) than the placebo group. In other words, the placebo group experienced recurrence of leukemia sooner than the 6-MP treatment group.

2. (7 points) Suppose that the researchers in the leukemia study sent their data to a statistician (you) and asked for a statistical recommendation of whether 6-MP should be used on leukemia patients. Write a short (one paragraph) report for the researchers, including a justification for your recommendation with a brief summary / interpretation of the statistical methods you used. Note that while this report will be an item in the homework document you submit, it should be a self-contained, stand-alone item, not referring to other items in your homework document. In other words, respond to this Exercise as though it is the only thing the researchers would see from you.

Report:

In analyzing and understanding these data from the leukemia treatment study, a log-rank test was performed. This technique was selected for two main reasons. First, the log-rank test makes no assumptions about the distribution of the time to the event of interest (in this case, the time to leukemia relapse). The log-rank test also takes into account and uses information from those in the study who either dropped out of the study or did not have a relapse during the duration of the study.

The log-rank test yielded a highly statistically significant result ($p < 0.0001$). This indicates that there is a highly significant difference in relapse time between patients who were given the 6-MP treatment and patients who were not. A visualization of the estimated survival curves provides additional insight into the impact that the 6-MP has on the time to a relapse (as compared with the control treatment). As shown in the graph to the right, the control group had leukemia relapses sooner on average than the 6-MP group. These data are significant evidence that the 6-MP treatment is effective at increasing time to leukemia relapse.



3. What affects survival following transplant?
    a. (6 points) Define indicator variables male (gender=1) and white (race=1). Fit an appropriate model for survival times, accounting for censoring, and including predictor variables male, white, and age. Report the significance results for each predictor.
        i. Male: P-value = 0.8794 – there is no significant change in survival/hazard due to gender.
        White: P-value = 0.5822 – there is no significant change in survival/hazard due to race.
        Age: P-value < 0.0001 – there is a highly significant effect on survival due to age.
        (2 points) What is the name of the model you used in part (a)?
        ii. Cox Proportional Hazaards Model
    b. (5 points) Provide a correct interpretation of the age coefficient from the model you fit in part (a).
        i. $\hat{\beta} = 0.05104$ for Age, with hazard ration $HR = 1.052$. After accounting for all other covariates, every year increase in age results in a 5.2% increase in hazard of death. As noted in part (a), this 5.2% increase in hazard of death is significant.
    c. (4 points) Report and test this model's major underlying assumption for the age predictor.
        i. In order to test the proportional hazards assumption for age, I defined 'AgeTime' as the interaction between age and time (defined within PROC PHREG). In testing this assumption, I found that the proportional hazards assumption is satisfied, because the interaction AgeTime was not significant (i.e. no effect of time), with a p-value of 0.2770.