Lab 3, PH Regression Basics

Dave Harrington

May 2018

The lab files come in two versions: a PDF with the problem statements and an Rmd file that produces the PDF. In most cases, you can work in the Rmd file and enter in your solutions. For the purely algebraic questions, you may either use LaTeX commands to enter your solutions in the Rmd file or write out the answers on paper.

The solution files (a PDF with the solutions, and the Rmd file to produce them) are contained in a separate folder under each lab. Your learning experience with the labs will be more effective if you do not look at the solutions until after you finish working on the questions.

MAC Prevention Clinical Trial

ACTG 196 was a randomized clinical trial designed to study the effects of combination regimens on the prevention of MAC (mycobacterium avium complex), one of the common opportunistic infections (OI) in patients with HIV infection. Patients were enrolled between April 1993 and February 1994, and follow-up ended August 1995.

The treatment regimens were clarithromycin (experimental), rifabutin (standard), and clarithromycin plus rifabutin (experimental combination)

This lab explores possible treatment effects.

The primary endpoint of the study was the time to development of MAC, which is associated with significant mortality. Secondary endpoints of the trial were survival, drug toxicity resulting in permanent discontinuation of study drugs, and quality of life as measured by a periodically administered questionnaire. The three problems in this lab step through some analyses of the trial.

The analysis of the primary endpoint (time to MAC) should be done with a competing risks analysis, since in this dataset, time to MAC was censored by either administrative censoring (independent censoring) or death (dependent censoring). Since the lectures have not covered the analysis of competing risks, this series of lab exercises examines the effect of treatment and other prognostic variables on time to death.

The data for this problem is in the dataset `mac`, which is in the package `eventtimedata`.

Be careful about the coding for the treatment. The coding `rif = 1` means that rifabutin was being used alone; `clari = 1` means that clarithromycin was being used alone. The combination arm (rifabutin and clarithromycin) is denoted by `rif = 0`, `clari = 0`.

Problem 1: The effect of treatment on time to death

a) Recode the treatment variables to avoid confusion. Either create three binary variables (one for each treatment arm), or create a single factor variable with three levels. Confirm that the new coding is correct using tables.

b) Explore the distribution of time to death with relevant numerical and graphical summaries, both overall and by treatment. Describe what you see.

c) Using a proportional hazards model, calculate an overall test statistic for differences in time to death among the three treatments, without adjusting for any covariates. Summarize your findings in a brief paragraph. Be sure to include a statement of what the null and alternative hypotheses are for the test.

d) Repeat the analysis in part c) using a three sample log-rank test. In this approach, what do the p-value and test statistic for differences among the three treatments correspond to in the analysis from part c)?

e) Repeat the analysis in part d), but with two pairwise log-rank tests. How does this approach differ from the one in part c)?

f) What is the estimated survival at 230 days for each of the three treatment groups?

g) Assess the assumption of proportional hazards for the three treatment groups by creating a plot of $\log[-\log(\widehat{S}(t))]$ versus $\log(t)$ for each of the three treatments. How should these plots look if the the proportional hazards assumption is approximately correct?

Problem 1 Solution.

a)

b)

c)

d)

e)

f)

g)

Problem 2: Association of CD4 cell count with time to death

The results from a clinical trial are often informative about prognosis for patients, even when the trial fails to establish a treatment effect. CD4 (cluster of differentiation 4) is a glycoprotein present on the surface of immune cells. When the number of CD4 cells in a patient is low, the immune system of the patient is less effective at fighting off infections. In the early days of the HIV epidemic, opportunistic infections in HIV+ patients were a major cause of mortality.

The CD4 count is a measure of the number of CD4 cells in the body, and is usually measured as the number of CD4 cells per cubic millimeter of blood. The variable `cd4cat` is coded `0` for patients with CD4 cell count lower than or equal to 25 cells per mm^3 of blood, and `1` for patients with CD4 cell count higher than 25 cells per mm^3 of blood.

This problem examines the association of CD4 cell count with time to death.

a) Produce a plot showing the estimated Kaplan-Meier survival functions for both CD4 > 25 and CD4 ≤ 25.

b) Calculate a log-rank test of the effect of CD4 category (> 25 vs ≤ 25) on the risk of death.

c) Calculate the generalized Wilcoxon test (Peto and Prentice, Fleming-Harrington test with $\rho = 1$) for the effect of `cd4cat` on risk of death. How does it compare to the log-rank test? When would you expect it to be less powerful than the log-rank test under the alternative suggested by the data?

d) Fit a Cox proportional hazards model to survival time, with `cd4cat` as the only covariate. Provide the Wald, score, and likelihood ratio tests for the effect of CD4 level on the HR for death. Are any of these test statistics equivalent to the log-rank or Wilcoxon test statistics from the above calculations?

e) Summarize the effect of CD4 in the model from part d) using the estimated hazard ratio and 95% confidence interval. Write a short interpretation of the hazard ratio.

Karnofsky score (`karnof`) is an overall health assessment made by the primary care physician, and has possible values (for this study) of 50, 60, 70, 80, 90 or 100, where 100 represents a lack of any symptoms and a score of 50 indicates impairment in daily function to the point of requiring considerable assistance and frequent medical care.

f) Compare the test statistics for the effect of CD4 category on the risk of death from the following models. Explain the assumptions behind each approach and differences in interpretation.

- Score test for `cd4cat` from Cox PH model, stratifying by Karnofsky score
- Score test for `cd4cat` from Cox PH model, controlling for Karnofsky score
- Log-rank test for `cd4cat`, stratifying by Karnofsky score

g) The dataset also contains the CD4 cell count as a numeric variable, `cd4`. Describe the association between `cd4` and time to death, and explain how this association compares to the association examined between `cd4cat` and death in the previous parts of this problem.

Problem 2 Solution.

a)

b)

c)

d)

e)

f)

g)

Problem 3: Adjusted analyses of treatment and CD4 counts

Fit a Cox PH model to the following variables: `age`, `sex`, `karnof`, `antiret`, `cd4cat`, and `treatment`.

The `antiret` variable indicates history of antiretroviral use, with `0` indicating never/unknown and `1` indicating previous or current use.

a) How is the "baseline" group defined in this model, in terms of the covariates? Does the baseline group correspond to any of the observations in the dataset?

b) Do the results from this model change the earlier conclusion about the possible differences among these treatments? Be sure to include both the global test of a treatment effect in this adjusted analysis, as well as the pairwise tests of the two experimental treatments versus the rifabutin control.

c) What is the estimated hazard ratio for death associated with a higher CD4 count (> 25 vs ≤ 25), adjusting for all other covariates? Give a 95% confidence interval for the hazard ratio for `cd4cat`, adjusting for the other covariates, and provide a verbal interpretation of the confidence interval for a non-statistician. How does this compare to the unadjusted effect estimated in Problem 2?

d) What is the interpretation of the estimated hazard ratio for age? What is the estimated hazard ratio for death for a subject aged 45 years versus a subject aged 30 years, holding all other covariates constant?

e) Based on this model, calculate the estimated hazard ratio of death for a subject aged 50 with baseline CD4 ≤ 25 and Karnofsky score of 70 versus a subject aged 30 with baseline CD4 $>$ 25 and Karnofsky score equal of 100.

f) In the early days of the HIV epidemic, it was thought that the prognosis differed between men and women. Based on these data, describe the association of sex with risk of death, using the summary statistics you believe are best suited for this.

g) It was also thought that the response to treatment might differ by sex; that is, that there might be a sex by treatment interaction for the outcome of time to death. Do these data support that hypothesis, at least with these treatments? Explain why the main effect for sex appears to be different in the model with the interaction versus the model without the interaction.

Problem 3 Solution.

a)

b)

c)

d)

e)

f)

g)