Lab 4, PH Regression Special Topics

Dave Harrington

May 2018

Problem 1: Plots based on cumulative hazard functions

a) Suppose the baseline hazard function is a constant λ_0. Show that a plot of $\log[-\log(\widehat{S}(t))]$ versus $\log(t)$ for a categorical covariate Z should be both approximately linear in $\log(t)$ and parallel for the different values of Z.

b) Show that if T_i (the survival time for the i^{th} individual) has survivorship function $S_i(t)$, then

- the transformed random variable $S_i(T_i)$ will have a uniform distribution on $[0, 1]$, and
- $-\log[S_i(T_i)]$ is from a unit exponential distribution.

Problem 1 Solution.

a)

b)

Problem 2: Null model residuals

In the Rossi recidivism example from lecture, what causes the "line" of residuals from the null model just below $y = -0.2$? Explore your conjecture in the data.

Problem 2 Solution.

Problem 3: Residuals for the MAC prevention trial analysis

This exercise takes a closer look at the data in the MAC prophylaxis trial, examining regression diagnostics. See the last lab for a discussion of the study background.

The last lab examined the possibility of treatment effects and the association of CD4 counts for time to death in both adjusted and unadjusted models, and explored the association of sex with outcome, including the possibility of a sex by treatment interaction. The last lab also examined the assumption of proportional hazards in the unadjusted analysis of treatment effects. This problem uses regression diagnostics to examine the quality of the fit of a PH model.

a) Fit a Cox PH model to the following variables: `age`, `sex`, `karnof`, `antiret`, `cd4`, and `treatment`. Recall that the `treatment` variable was constructed in the last lab.

b) Use martingale residuals to explore whether the variable `age` might be better modeled with a transformation.

c) Do the same with the variable `karnof` and describe what you see.

d) The model in the last lab used the categorical variable `cd4cat`. Do the residual plots suggest that was a good idea?

e) Using `cd4` rather than `cd4cat` might result in certain cases with very low or high CD4 counts having large leverage on the estimated coefficient (of `cd4`). Examine whether that might be the case with the appropriate type of residual.

f) Repeat the analysis in part e) for the variable `age`.

g) In a model that includes only the variable `sex`, use a $\log(-\log(\widehat{S}(t))$ vs $\log(t)$ plot to examine whether the PH assumption holds for `sex`.

h) Use the scaled Schoenfeld residuals to examine the proportional hazards assumption for `sex`. Does this approach suggest the same conclusion as the one from part g)?

i) Use `cox.zph()` to explore the PH assumption for all the variables in the full model fit in part a), and describe the results.

Problem 3 Solution.

a)

b)

c)

d)

e)

f)

g)

h)

i)

Problem 4: Correlated event times

Patients with superficial bladder cancer experience periodic recurrences of the disease in lesions that are generally limited to the urothelial lining of the bladder. The risk of metastatic disease is low, and these superficial lesions can be removed surgically or treated with chemotherapy. Unfortunately, the lesions return and the disease can lead to death or invasive surgery.

In their original paper on the analysis of correlated event times, Wei, Lin and Weissfeld used the data from a clinical trial of three treatments for superficial bladder cancer: thiotepa, pyridoxine (vitamin B6), and a placebo. The analysis presented in their paper examined the treatment effect of thiotepa versus placebo.

The `survival` package contains three versions of the bladder cancer dataset: `bladder1`, `bladder`, and `bladdder2`.

- The dataset `bladder1` contains the full data from the study, with all treatment arms and all 118 subjects.
- The dataset `bladder` contains data on 85 subjects with non-zero follow-up who were assigned to either thiotepa or placebo and only the first four recurrences for any patient. In this version of the dataset, the four tumors being followed are labeled with the `enum` variable, so there are four lines per case, and each line has the follow-up time measured from 0 (`stop`) and an event status indicator for that tumor (`event`). The `enum` variable is used for strata in a WLW analysis.
- The dataset `bladder2` is in the start-stop format used in the Andersen-Gill model for repeated events. The recurrences are treated as repeated events, with the start time for the next recurrence beginning at the stop time for the previous recurrence.

For more detail (and perhaps more clarity!) refer to the documentation for `bladder` in the `survival` package and have a look at `bladder` and `bladder2` in the RStudio data browser.

a) Explain the how the assumptions differ between using a marginal model based on the original WLW approach and a repeated events model based in AG.

b) Use a Cox PH model with a robust standard error to estimate the unadjusted treatment effect of thiotepa on bladder cancer in a marginal model, using the WLW approach.

c) Do the data support the assumption of proportional hazards on the time to first recurrence?

d) The variables `number` and `size` are the number of tumors and the size of the largest tumor (cm) at initial presentation. Does the estimated effect of treatment change when the model is adjusted for these variables?

e) In this model, do `number` and `size` seem to be modeled correctly with proportional hazards?

f) The WLW paper also examined the possibility of a different treatment effect on the first through fourth recurrences. There are two ways to do this (at least!). One way is to fit separate models for each of the strata defined by the `enum` variable. Another way is to use a treatment by stratum interaction for each of the strata.

   Explain the different assumptions behind the two approaches, do both analyses, and describe what you see.

g) Repeat parts b) - e), but using an Andersen-Gill model. Describe the results and explain how they differ from the WLW approach. How does the interpretation of the coefficient for treatment differ between the two models?

Problem 4 Solution.

a)

b)

c)

d)

e)

f)

g)