Methods in Biostatistics
Math 150, Spring 2019, Jo Hardin
Survival Analysis Project

HIV Clinical Trial

The data come from a double-blind, placebo-controlled trial that compared the three-drug regimen of indinavir (IDV), open label zidovudine (ZDV) or stavudine (d4T) and lamivudine (3TC) with the two-drug regimen of zidovudine or stavudine and lamivudine in HIV-infected patients (Hammer et al., 1997). Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outcome measure was time to AIDS defining event or death. Because efficacy results met a pre-specified level of significance at an interim analysis, the trial was stopped early.

REFERENCE: Study Information: `https://clinicaltrials.gov/ct2/show/NCT00000841`;
Study Outcomes (Hammer et al. (1997) NEJM *A Controlled Trial of Two Nucleoside Analogues plus Indinavir in Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less*): `http://www.nejm.org/doi/full/10.1056/NEJM199709113371101`

DATA: is comma delimited at: `http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv`

Variable	Name	Description	Codes/Values
1	id	Identification Code	1-1156
2	time	Time to AIDS diagnosis or death	Days
3	censor	Event indicator for AIDS defining diagnosis or death	1 = AIDS defining diagnosis or death 0 = Otherwise
4	time.d	Time to death	Days
5	censor.d	Event indicator for death (only)	1 = Death 0 = Otherwise
6	tx	Treatment indicator	1 = Treatment includes IDV 0 = Control group (treatment regime without IDV)
7	txgrp	Treatment group indicator	1 = ZDV + 3TC 2 = ZDV + 3TC + IDV 3 = d4T + 3TC 4 = d4T + 3TC + IDV
8	strat2	CD4 stratum at screening	0 = CD4 $\leq$ 50 1 = CD4 $>$ 50
9	sex	Sex	1 = Male 2 = Female
10	raceth	Race/Ethnicity	1 = White Non-Hispanic 2 = Black Non-Hispanic 3 = Hispanic (regardless of race) 4 = Asian, Pacific Islander 5 = American Indian, Alaskan Native 6 = Other/unknown
11	ivdrug	IV drug use history	1 = Never 2 = Currently 3 = Previously
12	hemophil	Hemophiliac	1 = Yes 0 = No
13	karnof	Karnofsky Performance Scale	100 = Normal;no complaint no evidence of disease 90 = Normal activity possible; minor signs/symptoms of disease 80 = Normal activity with effort; some signs/symptoms of disease 70 = Cares for self; normal activity/ active work not possible
14	cd4	Baseline CD4 count (derived from multiple measurements)	Cells/milliliter
15	priorzdv	Months of prior ZDV use	Months
16	age	Age at Enrollment	Years

Due Dates

Mon April 8 Post to GitHub both the R Markdown file and the corresponding html document. Include the names of the individuals in the group and some sort of EDA (exploratory data analysis). Do not print the data, but do something that indicates you've uploaded the data and know what some of the variables are. You might have some summary statistics or a graph. This is *not* an extensive assignment. Email me the URL for your project repository.

Fri April 19 Add on to the previous file(s) which you have posted to GitHub. Outline the "something new" part of the assignment. You should indicate who is doing what, what resources each of you will use to learn about your new topic, and a few sentences on what the topic is or how it relates to survival analysis / the analysis at hand. Additionally, for each "new" thing, provide 1-2 sentences describing what will be challenging about learning something new.

Wed April 24 If you would like me to look at a completed rough draft, you should turn in the completed draft before April 24.

Wed May 1 Complete assignment is due.

Wed May 8 Last day of class, each person / group will present their visualization, and there will be secret ballot voting for the best visualization of the data / model.

Assignment

The analysis should be done in a complete R Markdown file and posted to GitHub. If you would like a private repo, that can easily be set-up. The report should address the following tasks below, but the report should not enumerate the tasks. See below for grading rubric which speaks to how the report should be communicated.

1. Exploratory data analysis. Graphical and numerical summaries of the data (both explanatory and response variables). Do *not* attach every single graph you try. Instead, choose only a few images that give the reader a sense of the variables and their relationships.

2. Cox PH model. Using model building techniques, choose a subset of variables (including transformations, interactions, etc.) to find the model which produces the best survival information. The team whose model is best on a hold-out sample will earn an additional 5 points on the project.

   The Cox PH analysis should include: an interpretation of your final survival model including a discussion of the sign of the coefficients (note: feel free to use interactions) Which variable(s) are in? Which are out? What do you conclude about AIDS/death? Is there anything worth mentioning about how you got to your final model? What can you say about causation? What can you say about generalizing to a larger population?

3. Something new. Each individual in the class will learn a new idea related to survival analysis (see below). The new part will be assessed on your ability to explain the new idea. That is, you must indicate a complete understanding of the topic as well as an adequate use of the method applied to the dataset.

4. A visualization to enter into the competition. The visualization could be of the EDA variety, or it could be based on the model / results. The different visualizations will be presented to the class on the last day of the semester (May 8). The team whose visualization is best (secret ballot by the class) will earn an additional 5 points on the project.

R Hints

- Use R Markdown to facilitate working in GitHub. There is an amazing amount of help available here: `https://happygitwithr.com/`. If you work within the R Studio server, you will not need to install Git. You do, however, need to connect RStudio with GitHub. Let me know if you would prefer your work to be private (the default on GitHub is for it to be public).
- If you are working in pairs, the project is extended in two ways. 1. You must *both* be able to work on the files which means that you should be careful about merge conflicts and whatnot. 2. You must do *two* new items.
- Note that the event of interest is either "death" or "AIDS defining diagnosis".
- Be as creative as possible trying to think about how you might like to graphically display the data. If you come up with a cool idea for a graph but don't know how to implement it, *please email me and I'll write the code for you.*
- Please do not re-code the variables or change the variable names. You may, however, transform the variables within your R code (that is, for example, if you wanted to divide months by 12 to have years, or square a variable, etc.).

Something New:

Each individual should have some analysis that goes beyond a Cox PH model. For your analysis, you should give details of what is going on, how it is relevant, what are the assumptions, what are the conclusions, etc. Your analysis should indicate a sense that you understand and that you can communicate the results to a possible client.

Some possible topics to investigate include:

- Investigation of the proportional hazards assumption (what does the R function `cox.zph` do?)
- Exponential or Weibull PH regression (parametric survival model)
- Deriving / detailing AIC & BIC for model selection on Cox PH
- Power analysis (a simulation?)
- Derivation of the sample size calculation for the log rank test (and application to the data)
- An analysis of the Schoenfeld residuals (how are they calculated and why is that calculation relevant?)
- Bootstrapping the survival model (what are the assumptions? what do you conclude?)
- An analysis of possible time dependent covariates (do transformations help?)
- A description / rationale / derivation of the (early) stopping criteria and why it was used in this study
- An analysis / understanding / simulation of the multiple comparisons issues for checking whether the trial should end

Assessment:

- Your primary assessment will be based on the above items (graphic, modeling, understanding of new topic, additional analysis, interpretation).
- Additionally there will be two competitions. Winning either will add 5 points to your score.
  - Graphic: the class will vote on who has the best graphic.
  - Model: using a holdout sample (I only gave you part of the data), I will assess your coefficients. The group whose model best describes the holdout sample will win the model prize.