Statistical Methods for Clinical Trials
Math 150, Spring 2017, Jo Hardin
Survival Analysis Project

AIDS 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

Survival Analysis Project

Due Dates

April 3 Turn in an R Markdown file with the names of the individuals in the group and some sort of EDA (exploratory data analysis). Please 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.

April 14 Turn in the previous assignment, including any updates you've made along the way. Additionally: 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.

April 19 If you would like me to look at a completed rough draft, you should turn in the completed draft before April 19.

April 26 Complete assignment is due.

May 3 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 (preferably pdf with no table of contents so that the report uses the full width of the page). 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). If your group has more than one individual in it, the report should specify which new idea was learned and presented by which individual in your group.

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 3). The team whose visualization is best (secret ballot by the class) will earn an additional 5 points on the project.

R Hints

- 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?)
- Weibull PH regression (parametric survival model)
- Exponential regression (parametric survival model)
- Deriving / detailing AIC & BIC for model selection on Cox PH
- Time to event plot
- 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?)

Assessment:

- Your primary assessment will be based on the above items (graphic, modeling, 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 half of the data), I will assess your coefficients. The group whose model best describes the holdout sample will win the model prize.

Project Evaluation Assessment

The project will be assessed using a rubric which will probably be similar to that given below.

Title

_____ Does the title give an accurate preview of what the paper is about? (i.e. Is it informative, specific and precise?) 3 pts

Introduction

_____ Does the Introduction have a logical organization? Does it move from the general to the specific? 3 pts
_____ Has an explanation been given for why the research was done? Why is the work important? What is its relevance? 5 pts
_____ Is the final paragraph a brief description of the hypothesis/goals and findings of the paper? 2 pts

Methods

_____ Could the work be replicated based on the information given here? 4 pts
_____ Is the material organized into logical categories? 4 pts
The methods should be a source of detail about the approaches of the authors. Procedures that have been repeated by the authors should only be listed once. Variations to the procedure should be briefly summarized. (The methods should not read like a recipe).

Results (Model & New Ideas)

_____ Is the content appropriate for a results section? 10 pts / 5pts new idea

- Simple introduction to the scientific question
- Clear description of the results for each experiment
- Analysis of those results

_____ Are the results/data analyzed well? 10 pts overall / 5 pts new idea

- Given the data in each figure, is the interpretation accurate and logical?
- Is the analysis of the data thorough or are some aspects of the data ignored?
- Does the author make connections between ideas (graphs, models, etc.) within the text?
- Are the data interpreted in a larger context?

_____ Figures 3 pts

- Are the figures appropriate for the data being discussed?
- Are the figure legends and titles clear and concise?

Note: The entire experimental findings of a paper should be apparent from reading the results section. It should be possible to understand the question the authors are asking, the experimental approach they use to answer the question, the results of those experiments, and basic analysis of the data. Larger issues of what the research means, how it relates to other work, etc should be included in the discussion.

Discussion

_____ Does the author clearly state whether the results answer the question? (i.e. support or disprove the hypothesis?) 5 pts
_____ Were specific data cited from the results to support each interpretation? Does the author clearly articulate the basis for supporting or rejecting the hypothesis 5 pts?
_____ Does the author adequately relate the results of the current work to previous research? 5 pts
_____ Does the author appropriately discuss to whom the results can be generalized? 5 pts

References

_____ Are the references appropriate and of an adequate quantity? 2 pts
_____ Are the references cited properly (both within the text and at the end of the paper)? 3 pts

Writing Quality

_____ Is the paper well organized? (Paragraphs are organized in a logical manner) 7 pts
_____ Is each paragraph well written? (Clear topic sentence, single major point) 7 pts
_____ Is the paper generally well written? (Good use of language, sentence structure) 7 pts

Total

_____ Total number of points (100 pts possible)