Statistical Methods for Clinical Trials
Math 150, Fall 2014, Jo Hardin
Survival Analysis Group Project

Recidivism in North Carolina

"This data collection examines the relationship between individual characteristics and recidivism for [a cohort] of inmates released from North Carolina prisons [in 1980]. The survey contains questions on the background of the offenders, including their involvement in drugs or alcohol, level of schooling, nature of the crime resulting in the sample conviction, number of prior incarcerations and recidivism following release from the sample incarceration. The data collection also contains information on the length of time until recidivism occurs." `http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/8987?geography=North+Carolina`

Note that the data are not to be redistributed. Additionally, any intentional identification of a research subject or unauthorized disclosure of his or her confidential information violates the promise of confidentiality given to the providers of the information.

There is some literature available here: `http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/8987?geography=North+Carolina`. Including a paper which gives background on survival analysis generally and uses the data as a case study. (Keeping in mind that the paper was written in 1991 when software programs like R were only just starting to become widely available. The authors used FORTRAN to analyze the data.)

Chung, Ching-Fan, Schmidt, Peter, Witte, Ann D. Survival analysis: A survey. *Journal of Quantitative Criminology.* 7, (1), 59-98, 1991.

R Hints

- The data is on Sakai. Additionally, the codebook is there (with the variable information). Note that a missing variable is coded as -9, but I think I've removed all the missing variable rows. I have only provided you with data from 1980 and file=1.
- 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. [Note: I added a variable called timefollow which is the time variable we are used to working with. If redic=0, then the censored time variable is simply the followup time. Look carefully to see that you understand why timefollow is coded as it is.]
- Come to office hours or go to the QSC R office hours on Wed from 8-10pm to get help with the R stuff.

Group Presentation:

In your groups, you should analyze the survival data using a Cox PH model to see which variables are significant factors in determining recidivism rates. You must also produce a graphic that describes something about the explanatory variables (either in relation to each other or with respect to the response variable). Please include in your presenation:

- a graphic describing something interesting about the explanatory variables (again, if you think of a plot you don't know how to make, ask me and I'll write the code for you)
- some analysis that goes beyond a Cox PH model (e.g., investigation of the proportional hazards assumption, a comparison of the Cox PH model to the results from the log-rank test, an analysis of the Schoenfeld residuals (you'll have to do some additional background work), a chi-square test to address some kind of related question, a different analysis altogether - logistic regression??, etc.).

- 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 recidivism? 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?

Assessment:

- Your primary assessment will be based on the above items (graphic, additional analysis, interpretation).
- Additionally there will be two competitions. Winning either will give you a minimum grade of 85.
  - 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.

Due dates:

- Due: Monday, Dec 8 by noon. Send me your final model (simply the coxph statement is fine unless you transformed your variables. If you did transform your variables, I need the entire markdown file). I need the model so that I can re-run all of the analyses and figure out the model prize before class on Tuesday.
- Due: Tuesday, Dec 9. Each group will present for approximately 10 minutes. You may split up the presentation as you wish, but each individual must speak.
- Due: Tuesday, Dec 9. Each member of the group needs to post (to Sakai) the presentation as well as the group R markdown file.
- Due: Tuesday Dec 9. An additional group assessment (in Word, post confidentially to your Sakai drop box):
  - Given 100 points, the point allocation for the group members should be _____.
  - Provide a few sentences justifying the point allocation above.