NHANES cleaning assignment

Stephen Lauer

November 10, 2017

Today we are going to look at blood pressure in the `NHANES` dataset. Our goal is to look at the observed blood pressures and try to predict the unobserved blood pressures. Start by reading in the data set using `library(NHANES)` and `data(NHANES)`.

1) The variables that we are interested in today are `ID`, `Gender`, `Age`, `Weight`, `Height`, `BPSys1`, `BPSys2`, `BPSys3`, `BPSysAve`, `BPDia1`, `BPDia2`, `BPDia3`, `TotChol`, and `Diabetes`. Make a new data frame with only these covariates.

2) What is the age of the youngest person with a blood pressure reading? Remove all of the people from the data set who are younger than that person.

3) Make a new data frame called `BP_NA` with all of the people who still have `NA` values for `BPSysAve`. With your original data frame, remove these people and remove the column `BPSysAve`.

4) `gather` the columns that start with `BPSys` into two columns: `SysMeasure` and `BPSys`.

5) Keep only the numbers in the `SysMeasure` column. Make sure that they are class numeric.

6) Repeat 4 and 5 for the `BPDia` columns. Then only keep the rows where `SysMeasure` is the same as `DiaMeasure`.

7) Since systolic and diastolic blood pressures are measured at the same time, we're going to remove the strange values of either. Remove all values of systolic that are greater than 190 or less than 70. Remove all values of diastolic that are greater than 100 or less than 40. Also remove all `NA` values for both.

8) Find the average values of `BPSys` for each `ID` while keeping `Gender`, `Age`, `Weight`, `Height`, `TotChol`, and `Diabetes`. Save this as `BPSysAve2`.

9) Fit a linear model for `BPSysAve2` using the other covariates (but not `ID`) and save it to `BP_fit`. Check out the linear model fit.

10) Predict the values in `BP_NA` using `BP_fit` and save it to `BP_preds`.

I-can't-believe-you-made-it-this-far bonus) Compare the values of `BPSysAve2` that we generated versus the original values of `BPSysAve`. Are they the same? We haven't covered this as much, but you can use `left_join` to combine two data frames by matching variables (such as `ID`), which easily allows us to compare the averages by person.