NHANES Sleep Trouble assignment

December 8, 2017

Today we are going to look at modelling a binary variable in the `NHANES` dataset: `SleepTrouble`. Start by reading in the data set using `library(NHANES)` and `data(NHANES)`.

1) The variables that we are interested in today are `SleepTrouble`, `Gender`, `Age`, `Poverty`, `Work`, `BMI`, `Diabetes`, `Depressed`, `PhysActive`, and `PhysActiveDays`. Make a new data frame called `sleep_data` with only these covariates.

2) What is the age of the youngest person with a non-`NA` value for `SleepTrouble`? Remove all of the people from the data set who are younger than that person. Also, change `SleepTrouble` into a binary variable where `"Yes"=1` and `"No"=0`.

3) Is `SleepTrouble` correlated with `Gender`? Find the percentage of males and females with `SleepTrouble`.

4) Are the `BMI` values different between people with and without `SleepTrouble`? Make a boxplot to see this comparison.

5) The `Depressed` variable has three levels: `None`, `Several`, and `Most` (as well as some `NA` values). Change this into a binary variable where `None` $= 0$ and `Several`/`Most` $= 1$ (while the `NA` values stay `NA`).

6) There is some discrepancy between the `PhysActive` and `PhysActiveDays`. In what ways do these two variables disagree? How should we address this going forward?

7) Finally, let's fit a logistic regression model for `SleepTrouble` with the covariates that we have in our dataset (depending on your decision for how to address `PhysActive`). What variables are driving `SleepTrouble`? How well does the model explain the variation that we see in `SleepTrouble`?