

## Assignment 2

**Note:** Each of these data sets comes from Wooldridge's econometrics textbook. To use them, use the R package "wooldridge".

**Problem 1: Data visualization.** This problem will help you get under the hood in ggplot, to help you make data viz that looks good. We will use the "bwght2" data set, which contains data on prenatal behaviors and infant health.

- a. What does each observation represent? Write yourself a short description of the dataset.
- b. Choose 2-3 variables of interest here that might be related in some way. Write a paragraph or draw a diagram of your hypothesized relationship between the variables.
- c. Your goal is to create a very good and a very bad visualization of the same relationship between these variables. The very good visualization should be appropriate for any academic journal article, while the bad one should be either intentionally misleading (not fraudulent), hard to read, poorly designed, or all of the above. For both, pay attention to:
  - Chart type
  - Labels and titles, including units and formatting
  - Color schemes
  - Trends and smoothing
  - Errors

**Problem 2: Summary Tables.** This problem is meant to give you a chance to practice identifying, installing, and learning about packages you can use in R. There are lots of packages you can use to construct summary tables. For this problem, pick one (Google will be your friend!) and practice making a summary table.

Use the fertil1 data set, which is a data set detailing women and their fertility choices. Where relevant, make a well-formatted table (readable to someone with no knowledge of your data, with titles/labels/notes, well-formatted decimal places, etc.).

- a. What does each observation represent?
- b. Provide the following summary stats for the variable educ: min, first quartile, median, mean, standard deviation, third quartile, max, interquartile range, and range. Be able to describe how each of these (and their relationship) impacts your interpretation of the data.
- c. Make a frequency table for the variable kids, in both absolute and relative frequency.
- d. Choose 5 or so variables that you think are especially relevant and make a summary table that summarizes them (mean and SD/SEs are fine) across a group variable (for example, across racial groups, but you can choose your own meaningful division). The variables should be in rows with a different column for each group. Highlight any salient differences across groups.