Homework 9

Problem 1: Get out the vote

The data in turnout.csv contain information from a major party's voter database about a "get out the vote" campaign in advance of the 1998 midterm Congressional elections. The question of interest is whether receiving a "get out the vote" (GOTV) call from a volunteer in advance of the 1998 election increased the chances that someone actually voted that year. But from the standpoint of causal identification, the issue is that voters were not called randomly. Some voters were more likely to receive a GOTV call than others, and the recipients and non-recipients might differ in their underlying propensity to vote.

Each row in turnout.csv is about a single person. The variables relevant to our purposes are:

- `voted1998`: whether the person voted in the 1998 Congressional election. This is our outcome variable (1=yes, 0=no).
- `GOTV_call`: whether the person received a "get out the vote" call prior to the 1998 election (1=yes, 0=no). This is our treatment variable of interest.
- `voted1996`: whether the person voted in the 1996 Congressional election (1=yes, 0=no)
- `AGE`: the person's age in years
- `MAJORPTY`: whether the person is registered as a member of either one of the two major U.S. political parties (1=yes, 0=no)

Part A. How much more likely are GOTV call recipients to have voted in 1998? As a preliminary analysis, calculate the following quantities.

- The proportion of those receiving a GOTV call who voted in 1998.
- The sample proportion of those *not* receiving a GOTV call who voted in 1998.
- A large-sample 95% confidence interval for the **difference in these two proportions**: that is, the proportions of voting in 1998 (`voted1998==1`) for those who received a GOTV call versus those who didn't.

Part B. Consider the `voted1996`, `AGE`, and `MAJORPTY` variables. Provide evidence that at all three of these variables are **confounders** that prevent the difference you observed in Part A from representing the true causal effect of the GOTV call on the likelihood that a person voted in 1998. Confounders here would be factors that make someone more likely to receive a GOTV call *and* to have voted in 1998. Your evidence here can consist of any appropriate plot, table, or set of summary statistics.

Part C. Now let's get a better estimate of the effect of the GOTV call on the likelihood that a person voted. Use matching to construct a data set with `GOTV_call` as our treatment variable, and with `voted1996`, `AGE`, and `MAJORPTY` as our "matching" or "balancing" variables. Use 5 control cases for each treated case in your matching (`ratio=5`).

Provide evidence that your "matched" data set is, indeed, balanced with respect to the three confounders of `voted1996`, `AGE`, and `MAJORPTY`. (That is, show that these variables are no longer confounders for the matched data.) Then repeat your analysis from Part A, except using the matched data only. For this matched data set, calculate:

- The proportion of those receiving a GOTV call who voted in 1998.
- The sample proportion of those *not* receiving a GOTV call who voted in 1998.
- A large-sample 95% confidence interval for the **difference in these two proportions**: that is, the proportions of voting in 1998 (`voted1998==1`) for those who received a GOTV call versus those who didn't.

What do you conclude about the overall effect of the GOTV call on the likelihood of voting in the 1998 election?

Problem 2: Manufacturing flaws in circuit boards

In this problem, you'll look at data from a quality-control experiment from AT&T's process for manufacturing printed circuit boards. The data are in solder.csv.

The response variable, `skips`, is the number of solder skips on the circuit board (i.e. manufacturing flaws) that are apparent from a visual inspection. The remaining variables reflect different choices in the manufacturing process. The goal is to understand which combination of choices leads to the lowest number of skips, and therefore the most reliable manufacturing process. While there are many variables of possible interest, we'll focus on two grouping variables in this problem:

- `Opening`: the size of the opening on the solder gun (small, medium, or large)
- `Solder`: the thickness of the alloy used for soldering (thick or thin)

Part A: Make two plots. The first plot should provide evidence that the size of the opening on the solder gun is related to the number of skips. The second should provide evidence that the thickness of the alloy used for soldering is related to the number of skips. Give each plot an informative caption describing what is shown in the plot.

Part B: Build a regression model with `skips` as the outcome and with the following terms as predictors:

- a main effect for `Opening`
- a main effect for `Solder` type
- an interaction between `Opening` and `Solder` type

Make a table that shows the estimate and 95% large-sample confidence interval for each coefficient in your model.[1]

Part C: Interpret each estimated coefficient in your model in no more than 1-2 sentences. A good template here is provided in the course packet, when we fit a model for the video games data that had an interaction in it and interpreted each coefficient in a sentence or two.

Part D: If you had to recommend a combination of `Opening` size and `Solder` thickness to AT&T based on this analysis, which one would it be, and why? (Remember, the goal is to minimize the number of skips in the manufacturing process.)

[1] If you're using RMarkdown, you might find it easiest to use the `get_regression_table` function in the `moderndive` library, described here in our course packet.