Homework 3

Due: 12:00 pm (noon), October 13, 2020

Here are some general guidelines.

Do not include your name on your write-up, since these will be peer-graded anonymously.

Do not include your raw R code in your write-up unless we explicitly ask for it. You will submit your R script as a separate document to the write-up itself. In Canvas, you will see actually *two* assignments corresponding to homework 3: one for the write-up, and one for the R script. Your write-up is what get's graded, but your R scripts must also be submitted along with the homework, by the same deadline, for the purpose of audits and ensuring compliance with course policy regarding academic integrity. If you do not submit your R script, you will not receive credit for the homework.

If you use tables or figures, make sure they are formatted professionally. Figures and tables should have informative captions. Numbers should be rounded to a sensible number of digits (you're at UT and therefore a smart cookie; use your judgment for what's sensible). Rows and columns in tables should line up correctly, and tables shouldn't merely be copied and pasted in Courier (or similar) directly from the R output.

Problem 1

One of our colleagues at an Australian university ran the following little experiment with her students in an intro data science class. Everyone in the class stood up, and the professor asked everyone to fold their arms across their chest. Students then filled out an online survey with two pieces of information: 1) Did they fold their arms with the left arm on top of right, or with the right arm on top of the left? 2) Were they male or female?

The professor then asked her students to assess whether, in light of the data from the survey, there was support for the idea that males and females differed in how often they folded their arms with their left arm on top of the right. The survey data indicated that males folded their arms with their left arms on top about 5% more frequently. But was this just a "small-sample" difference? Or did it accurately reflect a population-level trend?

The data for this survey are in `armfold.csv`. There are two relevant variables:

- `LonR_fold`: a binary (0/1) indicator, where 1 indicates left arm on top, and 0 indicates right arm on top.

- `Sex`: a categorical variable with levels `male` and `female`.

(There's also a third variable indicating which hand the student writes with, but we're not using that here.)

Your task is to build a model, just as we did in class for the recidivism data, to assess support for any male/female differences in the population-wide rate of "left arm on top" folding. Make sure to quantify your uncertainty about how much more often males fold their left arms on top. (That is, it's not enough to just report the estimate for this sample; you have to provide a confidence interval that tells us how we can expect this number to generalize to the wider population. In doing so, you can treat this sample as if it were a random sample from the relevant population, in this case university students.)

Your write-up should include four sections:

1) **Question:** What question are you trying to answer?
2) **Approach:** What modeling approach did you use to answer the question?
3) **Results:** What evidence/results did your modeling approach provide to answer the question? This can include numbers, figures, and/or tables as appropriate.
4) **Conclusion:** What is your answer to the question?

Note: for a relatively simple problem like this, each of these four sections will likely be quite short. Nonetheless, these sections reflect a good general organization for a data-science write-up. So we'll start practicing with this organization on a simple problem, even if it seems a bit overkill at first. (It is certainly possibly in this case for each of them to be only 1 or 2 sentences long. Although you might feel you need more, and although nobody on our end is breaking out a word counter, it shouldn't be *too much* longer than that.)

Problem 2

For this exercise, you'll need to download the `GasPrices.csv` data set from the class website. This data set came from a student project in the spring of 2016 in the second McCombs statistics course (formerly STA 371, now STA 235). It was a pretty awesome project! We'll let the students who did the project describe things in their own words:

> Have you ever been driving through town looking to make a quick stop to fill up your car with gas and noticed that different gas stations are advertising different gas prices? Have you ever stopped to wonder why this might be the case? Could there be some underlying factors responsible for this noticeable difference in price, specifically for the same, regular unleaded mix of gas on the same day at the same time?
>
> To observe prices and other traits of gas stations firsthand, we visited 101 gas stations in the Austin area. We split the city into east and west sections with Lamar Blvd.~serving as the dividing line. At each gas station, we observed all necessary characteristics while staying in the car. We used the Maps app to determine the address and zip codes of the gas stations and the transportation feature within Maps on the iPhone to locate the gas stations themselves. We input the data directly into an Excel spreadsheet. Once we had visited all 101 gas stations, we used the US Census Bureau's American Fact Finder to input the median income for each zip code.

Needless to say, these students knocked it out of the park for effort. (While there's no project in this course, you'll likely be asked to do one when you get to STA 235.) Let's look at their data set and use it to answer some questions. There are lots of variables in this data set, but for our purposes here, the important ones are as follows:

- ID: Order in which gas stations were visited
- Name: Name of gas station
- Price: Price of regular unleaded gasoline, gathered on Sunday, April 3rd, 2016
- Highway: Is the gas station accessible from either a highway or a highway access road?
- Stoplight: Is there a stoplight in front of the gas station?
- Competitors: Are there any other gas stations in sight?
- Zipcode: Zip code in which gas station is located
- Income: Median Household Income of the ZIP code where the gas station is located based on 2014 data from the U.S. Census Bureau
- Brand: ExxonMobil, ChevronTexaco, Shell, or Other.

The theories

People have a lot of pet theories about what explains the variation in prices between gas stations. Here are several such theories:

A) Gas stations charge more if they lack direct competition in sight.

B) The richer the area, the higher the gas prices.

C) Gas stations at stoplights charge more.

D) Gas stations with direct highway access charge more.

E) Shell charges more than all other non-Shell brands.

Which of these theories seem true, and which are unsupported by data? Take each theory one by one and assess the evidence for the theory in this data set.

Your discussion of each theory should include the same four mini-sections as in the first problem: 1) Question; 2) Approach; 3) Results; 4) Conclusion. No single theory should require more than a single typed page to assess, including any figures, tables, or numbers you use to make your case. Less than 1 page per theory is perfectly acceptable—indeed, preferable, as long as you can cover all the bases.

In assessing the evidence for each theory, make sure you appropriately deal with a major issue: **uncertainty.** Remember, this is just a sample of gas stations, and there is always uncertainty when generalizing from a specific sample to a wider population. Therefore, you need to quote confidence intervals rather than just single-number estimates for these differences. So, for example, it's not enough to say something like, "the difference in price between gas stations on and off the highway is X cents," and to draw your conclusion from this statement. Instead, you have to say something like, "the difference in price between gas stations on an off the highway is somewhere between L and U, with 95% confidence" (where you supply the lower and upper bounds L and U), and to draw your conclusion from this interval.

Problem 3

In this problem, you'll analyze data from an experiment run by EBay in order to assess whether the company's paid advertising on Google's search platform was improving EBay's revenue. (It was certainly improving Google's revenue!)

Google Ads, also known as Google AdWords, is Google's advertising search system, and it's the primary way the company made its $162 billion in revenue in fiscal year 2019. The AdWords system has advertisers bid on certain keywords (e.g. "iPhone" or "toddler shoes") in order for their clickable ads to appear at the top of the page in Google's search results. These links are marked as an "Ad" by Google, and they're distinct from the so-called "organic" search results that appear lower down the page.

Nobody pays for the organic search results; pages get featured here if Google's algorithms determine that they're among the most relevant pages for a given search query. But if a customer clicks on one of the sponsored "Ad" search results, Google makes money. Suppose, for example, that EBay bids $0.10 on the term "vintage dining table" and wins the bid for that term. If a Google user searches for "vintage dining table" and ends up clicking on the EBay link from the page of search results, EBay pays Google $0.10 (the amount of their bid). [1]

For a small company, there's often little choice but to bid on relevant Google search terms; otherwise their search results would be buried. But a big site like EBay doesn't necessarily have to pay in order for their search results to show up prominently on Google. They always have the option of "going organic," i.e. **not** bidding on any search terms and hoping that their links nonetheless are shown high enough up in the organic search results to garner a lot of clicks from Google users. So the question for a business like EBay is, roughly,

[1] There's huge variability in the market price of different search terms. The market price per click for a search term like "insurance" or "attorney" or "MBA programs" might be $50 or more. For stuff you might buy on EBay, it's usually a lot less.

the following: does the extra traffic brought to our site from paid search results—above and beyond what we'd see if we "went organic"—justify the cost of the ads themselves?

To try to answer this question, EBay ran an experiment in May of 2013. For one month, they turned off paid search in a random subset of 70 of the 210 designated market areas (DMAs) in the United States. A designated market area, according to Wikipedia, is "a region where the population can receive the same or similar television and radio station offerings, and may also include other types of media including newspapers and Internet content." Google allows advertisers to bid on search terms at the DMA level, and it infers the DMA of a visitor on the basis of that visitor's browser cookies and IP address. Examples of DMAs include "New York," "Miami-Ft. Lauderdale," and "Beaumont-Port Arthur."

In the experiment, EBay randomly assigned each of the 210 DMAs to one of two groups:

- the treatment group, where advertising on Google AdWords for the whole DMA was paused for a month, starting on May 22.

- the control group, where advertising on Google AdWords continued as before.

In `ebay.csv` you have the results of the experiment. The columns in this data set are:
- DMA: the name of the designated market area, e.g. New York
- rank: the rank of that DMA by population
- tv_homes: the number of homes in that DMA with a television, as measured by the market research firm Nielsen (who defined the DMAs in the first place)
- adwords_pause: a 0/1 indicator, where 1 means that DMA was in the treatment group, and 0 means that DMA was in the control group.
- rev_before: EBay's revenue in dollars from that DMA in the 30 days before May 22, before the experiment started.
- rev_after: EBay's revenue in dollars from that DMA in the 30 days beginning on May 22, after the experiment started.

The outcome of interest is the **revenue ratio** at the DMA level, i.e. the ratio of revenue after to revenue before for each DMA. If EBay's paid search advertising on Google was driving extra revenue, we would expect this revenue ratio to be systematically lower in the treatment-group DMAs versus the control-group DMAs. On the other hand, if paid search advertising were a waste of money, then we'd expect the revenue ratio to be basically equal in the control and treatment groups.

Two explanatory notes here:

- We use the ratio rather than the absolute difference because the DMAs differ enormously in population and therefore revenue.

- We wouldn't necessarily expect the before-and-after revenue ratio to be 1 (i.e. similar revenue before and after the experiment), even in the control-group DMAs. That's because, like any retailer, EBay's sales exhibit a lot of seasonal patterns and might be lower in some months across the board, regardless of paid search. That's why the important question isn't whether the revenue is the same before and after in the treatment-group DMAs, but whether the before-and-after **ratio** is the same for the treatment group as for the control group.

Your task is to fit a model and use a randomization test to assess the evidence for whether the revenue ratio is the same in the treatment and control groups, or whether instead the data favors the idea that paid search advertising on Google creates extra revenue for EBay. Make sure you use at least 10,000 Monte Carlo simulations in your randomization test. As in the previous two problems, you're write-up should include the same four sections: 1) Question; 2) Approach; 3) Results; 4) Conclusion.