Homework 3

Due: before 12:00 pm (noon) on Tuesday, April 13. Please do not include your name on your write-up, since these documents will be reviewed by anonymous peer graders.

Do not include your raw R code in your write-up. You will submit your R script as a separate document to the write-up itself. On Canvas, you will see *two* assignments pages corresponding to Homework 2: (1) to upload your write-up PDF file and (2) to upload the R script that you used to generate your write-up. Your write-up is what will be peer graded. The R script will not be graded, but you must submit it to receive credit on the write-up.

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.

Unless the problem instructions suggest otherwise, **format your write-up responses** in the same manner as on Homework 2, with four sections: 1) Questions; 2) Approach; 3) Results; 4) Conclusions.

For problems requiring Monte Carlo simulation, use a minimum of 10,000 iterations.

Problem 1 - Summer is Coming

The file *ERCOT.csv* contains data on peak power consumption in the Gulf Coast region of Texas for every hour between 9 AM and 7 PM of every summer day (June 1 through August 31) during 2010-2015. We scraped these data from the website for ERCOT, the electricity grid operator for most of Texas. The variables in this data frame are:

- *time*: date and time stamp of the data point. Each data point covers a one-hour interval between the hours of 9 AM and 7 PM, with the window beginning at the time stamp indicated in this column.
- *COAST*: peak demand in megawatts for the entire coast region of Texas during that hour. This covers more or less from Houston and its surrounding areas down to Matagorda Bay, but not as far south as Corpus Christi.
- *temp*: average temperature recorded during that one-hour interval by the weather station at Houston's Hobby Airport, in degrees Celsius.
- *weekday*: whether the day in question is a weekday, where 1 = weekday and 0 = weekend.

Your task is to build a linear model for peak power consumption that includes main effects for temperature and weekday, as well as an interaction between temperature and weekday. Use this model to address the following questions:

A) During the summer, how much higher or lower is daytime power consumption on a weekday, versus on a weekend?

B) During the summer, how does daytime power consumption increase with temperature, on average, and does this relationship seem to differ between weekends and weekdays?

Be sure to quantify your uncertainty by quoting confidence intervals where appropriate. Estimates of coefficients that do not include appropriate error bars (i.e., confidence intervals) will not receive full credit.

Include (in the Results section of your write-up) a faceted scatter plot that shows the relationship between power consumption and temperature, faceted by weekday.

Problem 2 - Price Elasticity of Cheese

This question considers data on sales volume, price, and advertising display activity for packages of Borden sliced cheese, available in *cheese.csv*. We've worked with a slice of these data before (for Kroger stores in the Dallas area). You'll now examine the full data set. For each of 88 stores (*store*) in different US cities, we have repeated observations of the weekly sales volume (*vol*, in terms of packages sold), unit price (*price*), and whether the product was advertised with an in-store display during that week ($disp = 1$ when a display was present).

The goal of this analysis is to understand consumer behavior for Borden slice cheese, by characterizing the price elasticity of demand for this market. Remember our *milk* walkthrough example from class using sales-versus-price data: a typical model for price elasticity of demand is of the form $Q = KP^{\beta}$, where Q is quantity sold, P is price, K is a constant, and β is the elasticity—that is, the relative percentage change in sales as price changes by 1%. You should recall how to use linear regression (least squares) to fit such a model.

Build a model for Q (sales volume) in terms of P (price), store-level dummy variables, and a dummy variable for whether or not there was a display for cheese.

Use this model to answer the following questions, quoting appropriate confidence intervals:

A) What is the price elasticity of demand for Borden sliced cheese in no-display weeks? Interpret this number in a single sentence (i.e. "When price of cheese goes up by 1%. . . ").

B) Does price elasticity for Borden cheese appear to be changed by the presence of in-store display? (Hint: remember about interaction terms in models with numerical and categorical predictors; you will want to fit a modified version of the model described in the instructions above.) Summarize one economic explanation for your result in the Conclusions section of your write-up (in 2-3 sentences max).

Problem 3 - COVID-19 Exponential Growth

The file `covid.csv` contains data on daily reported COVID-19 deaths for Italy and Spain—two of the hardest-hit European countries—during the first pandemic wave in February and March of 2020. The columns in this data frame are:

- *date*: the calendar date
- *country*: Italy or Spain
- *deaths*: the number of reported COVID-19 deaths in that country on that day
- *days_since_first_death*: the number of days elapsed since the first death in that country

Your task is to fit an exponential growth models for EITHER Italy OR Spain – you choose – using `days_since_first_death` as the time variable. Use the results of your model to characterize the doubling time of the daily death total in the country that you selected for the exponential growth model (either Italy or Spain). (For context, these doubling times early in the epidemic are used to estimate R_0, the basic reproductive rate of the virus. Although you will not have to calculate R_0 here).

Make sure that your write-up includes the following:

1. a confidence interval for the doubling time in the country (either Italy or Spain)
2. a line graph showing reported daily deaths over time (using `days_since_first_death`, rather than calendar date, as the relevant time variable) in the country that you chose for the exponential model.

Problem 4 - Capital Metro UT Ridership

The file `capmetro_UT.csv` contains data from Austin's own Capital Metro bus network, including shuttles to, from, and around the UT campus. These data track ridership on buses in the UT area. Ridership is measured by an optical scanner that counts how many people embark and alight the bus at each stop.

Each row in the data set corresponds to a 15-minute period between the hours of 6 AM and 10 PM, each and every day, from September through November 2018. The variables are:

- *timestamp*: the beginning of the 15-minute window for that row of data
- *boarding*: how many people got on board any Capital Metro bus on the UT campus in the specific 15 minute window
- *alighting*: how many people got off ("alit") any Capital Metro bus on the UT campus in the specific 15 minute window
- *day_of_week* and *weekend*: Monday, Tuesday, etc, as well as an indicator for whether it's a weekend.
- *temperature*: temperature at that time in degrees F
- *hour_of_day*: on 24-hour time, so 6 for 6 AM, 13 for 1 PM, 14 for 2 PM, etc.
- *month*: July through December

Your task in this problem is **to make two faceted plots** and to answer questions about them.

1. One panel of line graphs that plots **average boardings** grouped by hour of the day, day of week, and month. You should facet by day of week. Each facet should include three lines, one for each month, colored differently and with colors labeled with a legend. Give the figure an informative caption in which you explain what is shown in the figure and address the following questions, citing evidence from the figure. Does the hour of peak boardings change from day to day, or is it broadly similar across days? Why do you think average boardings on Mondays in September look lower, compared to other days and months? Similarly, why do you think average boardings on Weds/Thurs/Fri in November look lower?
2. One panel of scatter plots showing boardings (y) vs. temperature (x) in each 15-minute window, faceted by hour of the day, and with points colored in according to whether it is a weekday or weekend. Give the figure an informative caption in which you explain what is shown in the figure and answer the following question, citing evidence from the figure. When we hold hour of day and weekend status constant, does temperature seem to have a noticeable effect on the number of UT students riding the bus?

These are exactly the kind of figures that Capital Metro planners might use to understand seasonal and intra-week variation in demand for UT bus service. These are also the kind of figures one would create in the process of building a model to predict ridership. This problem has conceptual and coding similarities to our Homework 1 problem on bike-share ridership in Washington, DC. Feel free to use that script as a template for some aspects of this problem. The basic skills of "group/pipe/summarize" and plotting will remain useful.

Notes:

First, a feature of R is that it orders categorical variables alphabetically by default. This doesn't make sense for something like the day of the week or the month of the year. Paste the following block of code into your R script at the top and execute it before you start further work on your plots for this problem:

```
# Recode the categorical variables in sensible, rather than alphabetical, order
capmetro_UT = mutate(capmetro_UT,
              day_of_week = factor(day_of_week,
                levels=c("Mon", "Tue", "Wed","Thu", "Fri", "Sat", "Sun")),
              month = factor(month,
                levels=c("Sep", "Oct","Nov")))
```

Finally, this problem need not follow our standard "Questions/Approach/Results/Conclusions" format. Simply include the two figures and their captions in your write-up. Be sure to keep each figure + caption to a single page combined (i.e. two pages, one page for first figure + caption, a second page for second figure + caption).

Problem 5 - Experimental Design

Consider the following two experimental studies.

Study 1: An economist conjectures that people will perform worse on a skill-based task when they are in the presence of an observer with a financial interest in the outcome. She conducted the following study, in which a total of 42 participants played a video game that required them to get Sonic the Hedgehog to the end of an obstacle course as quickly as possible. Subjects were randomly assigned to one of two groups. One group was told that the participant and observer would each win $10 if the participant beat a certain threshold time. The other group was told that only the participant would win the prize if the threshold were beaten, and that the observer had no stake in the outcome. Her analysis revealed that the two groups had no difference: they both met the threshold time at about the same average rate.

Study 2: A financial planner conjectures that consumers with poor credit histories will benefit from receiving free credit-counseling services. She spent several weeks recruiting study participants at car dealerships. She ultimately collected a sample of 155 consumers who were shopping for a new car, but had reported difficulty in securing a loan to make the purchase. All 155 of these consumers were offered free credit-counseling services, and 94 accepted the offer. Of the 94 consumers who received credit counseling, 84% managed to secure a car loan within 6 months. Of the 61 who did not receive counseling, only 49% managed to get a loan.

One of these studies has a notably superior experimental design. Which is it, and why?

(a) Study 1, because it has both a treatment group and a control group.
(b) Study 2, because it has a larger sample size.
(c) Study 1, because it randomized participants to treatment and control groups.
(d) Study 2, because it showed a difference between the two groups (whereas Study 1 showed no difference).

In your write-up, indicate which answer choice (a, b, c, or d) is correct and provide a short explanation (1-2 sentences) to justify your answer.

Problem 6 - Redlining

The term "redlining" dates to the late 1960s, when it was first used by community activists to describe the practice wherein banks would "draw a red line on a map" to mark an area where they simply refused to grant loans. In areas that banks were allegedly redlining, people would find it difficult to get access to credit, regardless of circumstance or merit. A 1934 map of Austin that was used by realtors and businesses reveals redlining with marks for large swathes of East and South Austin (with predominantly Black and Latino/a residents) as "Hazardous".

Through originally coined in the context of bank loans, the term "redlining" has since acquired a broader meaning, and can apply to any denial of service based on ethically dubious grounds—race, gender, sexual orientation, and so forth. No financial institution openly admits to redlining, of course. It's always a question of whether such practices are happening behind closed doors, with the denial of service then being rationalized in public using other, more socially acceptable reasons.

We will now consider a famous example of possible redlining in the insurance industry, from Chicago in the 1970s. To measure access to the private home insurance market, we will use a proxy: the number of policies issued by a federal program called FAIR, which offers insurance to those who cannot find coverage from private firms. Homeowners are required to carry insurance as a condition for receiving a mortgage, so if someone wants to buy a house, they must get insurance somewhere, whether through the private market or through FAIR. So the logic here is that the more FAIR policies we see per capita in a ZIP code, the less the residents in that ZIP code are able to turn to the private market.

However, private insurance companies can have perfectly reasonable, non-racially motivated grounds for denying insurance to people. Some areas of town, for example, are more subject to fire because of nearby industry, mix of building materials, or prevailing wind patterns. Other areas have mainly older houses that pose greater liability risks. In individual cases, it is therefore hard to tell whether an insurance company's decision to deny coverage was the result of redlining, or a non-race-based actuarial calculation. Nonetheless, we can try to adjust for these ZIP-code differences and look for city-wide patterns that might constitute evidence of discrimination.

A second issue with this proxy is that it is difficult to tell whether people must turn to a FAIR policy because they are denied coverage outright or because they cannot afford the private coverage offered. Hence we also have to adjust for differences in income among different areas that might explain differences in FAIR policies, even in the absence of racism.

In `redline.csv`, you are given data on 40 different ZIP codes in Chicago. (All we have is the ZIP-code-level data, rather than data for individual homeowners.) For each ZIP code, you have the following information:

- `policies`: new FAIR plan policies and renewals per 100 housing units in that ZIP code. Remember that more FAIR policies means that residents in that ZIP code are buying private insurance at lower per-capita rates. It's our proxy for access to the private market, where more policies implies less access.
- `minority`: percentage of residents in that ZIP code who, on the last US census, self-identified as a member of racial/ethnic minority (i.e. other than non-Hispanic white.)
- `fires`: fires per 100 housing units in that ZIP code.
- `age`: percent of housing units in that ZIP code built before WWII
- `income`: median family income in thousands of dollars

Use a linear regression model to assess whether there is an association between the number of FAIR policies and the racial/ethic composition of a ZIP code, adjusting for the `fire`, `age`, and `income` variables. Write a short report summarizing your analysis and interpreting results, using the usual format you used before (Question/Approach/Results/Conclusion). Provide confidence intervals where appropriate. In your Results section, report and interpret a measure of model fit to convey the percentage of variability in policies explained by the predictor variables in your regression model.