

Homework 5

Due: 12:00 pm (noon), November 10, 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 5: one for the write-up, and one for the R script. Your write-up is what gets 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.

Except on problem 1, **format your answers** in the same way we've learned to do on previous homeworks, with four sections: 1) Questions; 2) Approach; 3) Results; 4) Conclusions.

Problem 1

Background. This problem is a little break from regression modeling; instead, it takes you back to the data visualization unit. Remember, data visualization is one of the most important tools of data science, and it's almost always an important part of building a regression model. So it's good to practice! In particular, this problem should remind you of the homework problem we did many weeks ago now, on ridership in Washington, DC's bike-share network. Feel free to use that script as a starting point for this problem, which involves a similar kind of situation, but a little closer to home. The basic skills of “group/pipe/summarize” and plotting are really useful for exploring data, so it's good to keep them sharp.

Data and problem: The data in `capmetro_UT.csv` contains data from Capital Metro, which runs the bus network in Austin, including shuttles (like the West Campus and 40 Acres routes) to, from, and around UT. The data tracks ridership on buses in the UT area, which is measured by an optical scanner that counts how many people get on and off 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.

- 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?
- 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. They're also the kind of figures you'd make to assist in building a model to predict ridership (even though, in this problem you won't actually be building that model).

Notes:

First, this problem need not follow our standard "Questions/Approach/Results/Conclusions" format. Just turn in the two figures and their captions. 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).

Second, 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. So if you want to re-order these variables in their usual order, try pasting the following block of code into your R script at the top, and executing it before you start further work on it.

```
# 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")))
```

Problem 2

The background: In this problem, you'll see how "p-hacking" actually works! Remember, p-hacking is not a recommended practice—quite the opposite. It's something to be avoided, and to be on the lookout for in others' work. Thus the point of this problem is to sensitize you to the range of possible choices that one can make in a data analysis; it's this sheer range of choices that makes p-hacking even possible.

The data: You'll be returning to our data set from earlier in the semester on green buildings. We've picked a subset of variables for you, in **green_hack.csv**.

- CS_PropertyID: the building's unique identifier in the CoStar database on commercial properties. This is just an ID number and isn't meant to be included in models.
- size: the total square footage of available rental space in the building.
- Rent: the rent charged to tenants in the building, in dollars per square foot per calendar year. - leasing_rate: a measure of occupancy; the fraction of the building's available space currently under lease.

- `rev_psf`: the revenue per square foot per year, which is $\text{Rent} * \text{leasing_rate} / 100$.
- `stories`: the height of the building in stories.
- `age`: the age of the building in years.
- `renovated`: whether the building has undergone substantial renovations during its lifetime.
- `Class`: A, B, or C. These are relative classifications within a specific market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.
- `green_rating`: an indicator for whether the building is either LEED- or EnergyStar-certified.
- `LEED`: indicator for a specific kind of green certification called LEED.
- `amenities`: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.
- `Utility_Costs`: a composite measure of how much utilities (gas, water, and electricity) cost in the building's geographic region. This measure has been scaled to have a mean of 0 and a standard deviation of 1. So, e.g. if `Utility_Costs` = 1, then that building's utility costs are 1 standard deviation above the national average utility costs.
- `City_Market_Rent`: a measure of average market rent per square-foot per calendar year in the building's local market. This measures the general market conditions for commercial real estate near the building.

Part A

Run the most convincing analysis you can in which you find that green certification **is** a “statistically significant” predictor of success on the commercial real estate market. You should judge statistical significance according to whether, in your model, the 95% confidence interval for the effect of green certification has only positive values (i.e. does not contain 0).

Good luck! Here are your “researcher degrees of freedom”:

- 1) The choice of outcome measure. You can measure “success” on the real estate market using either the rent that tenants pay in the building (variable `Rent`) or the revenue per square foot (`rev_psf`, which is $\text{Rent} * \text{Leasing_rate} / 100$ and which accounts for the fact that some buildings aren't at full occupancy).
- 2) How to measure green certification. You can either use the `green_rating` variable (which includes both LEED and EnergyStar certifications) or you can use the `LEED` variable (which doesn't include EnergyStar).
- 3) The methodological approach. You can try to identify the effect of green certification using one of two approaches that we've covered in class: either matching or by building a multiple regression model.
- 4) The set of confounders to control for, from those on the list above. You have total freedom here, with the following caveat: you must control for at least `age`, `Class`, and `City_Market_Rent`. Approaches that don't attempt to control for these three variables will not receive full credit. (Remember, the idea of p-hacking is to come up with a plausible approach that yields the desired result; models that don't control for these variables won't be very plausible.)

Your answers to these questions should be summarized in your Approach section.

Part B

Run the most convincing analysis you can in which you find that green certification **is not** a “statistically significant” predictor of success on the commercial real estate market. Again, you should judge statistical significance according to whether, in your model, the 95% confidence interval for the effect of green certification has only positive values (i.e. does not contain 0).

You have the same research degrees of freedom as in Part A, and the same caveat: you must control for at least `age`, `Class`, and `City_Market_Rent`.

Part C

Now make a judgment. Which of your answers—Part A, or Part B—do you find more plausible? Why?

To help you answer this, you should make a big table with three columns (Part A, Part B, and Winner) and four rows corresponding to your four researcher degrees of freedom (choice of outcome, how green certification was measured, the approach, and the set of confounders adjusted for). In the Parts A and B columns, you'll briefly summarize the choices made in Parts A and B, respectively. (These are kind of like a summary, in tabular form, of the two Approach sections from Parts A and B). Then in the Winner column, you'll argue why one choice for that row (either Part A or B) makes more sense to you.

Below the table, write a single paragraph summarizing which approach, and which overall answer, seems like the best one in terms of overall plausibility.

Your table should fit on a single page, so be concise. Note that your table and paragraph for Part C fit outside the usual “Questions/Approach/Results/Conclusions” framework that you're following in Parts A and B.

Notes

Here are some important notes:

- Remember that however you approach parts A and B, you must control for at least these three variables: age, Class, and City_Market_Rent.
- Each of Parts A and B should *individually* follow the standard four-part skeleton for your write-ups: Question, Approach, Results, Conclusion. That is, don't combine these parts into a single Questions section, a single Approach section, etc; Parts A and B are like two independent write-ups. (The question in each case is the same; the approaches will obviously differ; and the conclusions should be the opposite of each other!)
- Your Approach sections here will be longer than on previous write-ups. Each approach section must summarize and attempt to justify all the choices made for each of your four researcher degrees of freedom. Remember, p-hacking involves coming up with a *plausible* approach that leads to the wished-for outcome; therefore your Approach section must at least attempt to justify the plausibility of what you've done.
- If for whatever reason you are not able to find an approach that leads to one result or the other, report the approach that gets the closest! For example, if you can't find an approach where the confidence interval for your green-certification effect doesn't contain zero, report results for the approach that comes the closest you can to a confidence interval that doesn't contain zero.
- The second point of this problem, beyond learning about p-hacking, is to make you appreciate something else important. Just because two different analyses can lead to two different answers on the same problem, it does not follow that we should just throw up our hands in despair, moan about p-hacking, and say that no one knows the answer. That's the path of cynicism. Instead, take the path of *skepticism*: when two approaches differ, one can compare those two approaches, acknowledge their relative strengths and weakness, and come to informed judgment about which approach—and therefore, which conclusion—makes more sense.

Problem 3

The data in `covid.csv` contains daily Covid-19 deaths for two of the hardest-hit European countries—Italy and Spain—during the first pandemic wave in February and March, 2020. The columns 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 separate exponential growth models for Italy and Spain, using `days_since_first_death` as the time variable, and to characterize the doubling time in each country's daily death total. (These doubling times early in the epidemic are used to estimate R_0 , the basic reproductive rate of the virus.) Are these doubling times similar, or noticeably different from one another?

Make sure that your write-up includes the following information:

- a confidence interval for the doubling time in each country.
- a line graph showing deaths over time (using `days_since_first_death`, rather than calendar date, as the relevant time variable), faceted by country

As always, format your write-up in four sections: Questions, Approach, Results, Conclusions.