Ast2: The Dark Forest

Due Feb 18 at 11:59pm on moodle. No email submissions.

The assignment can easily be done individually but pairs/groups are *encouraged*. Only one group member should submit the assignment. Include all group member names. Other group members will have a moodle warning they didn't do an assignment; ignore that. I manually enter grades for everyone.

You're welcome (actually, you're encouraged) to use online resources to help you find facts/insights for this assignment. This includes using Generative AI. However, you are ultimately responsible for content, and you must *document* which knowledge/tool you use and where. Any content without explicit citation is assumed to be your own work.

This assignment probably not require citations but if you have any, they should be properly formatted bibliography/references. See assignment 1 requirements regarding citations.

All plots should be

Assignment objectives

- Read/lookup documentation on how to access raw/messy data and manipulate it to match another data set
- Perform a linear regression exploration/investigation to choose a model.
- Investigate slightly unintuitive results and again select a model

Your assignment doesn't need to be a report this time. All questions should be answered in no more than a couple sentences each (most questions can be answered with just one sentence), plus a plot where requested.

Do not provide any code nor output unless properly formatted and presented, and not extraneous to the information requested. As before, you may be penalized for poor presentation.

1 Part 1: Electricity consumption

Your goal is to reasonably relate the consumption by the outside temperature. I'm making you get the temperature data from the last several years yourself from the goverment's website:

i. Go to https://weather.gc.ca/city/pages/bc-45_metric_e.html

ii. At the bottom, click the link for "Historical Weather"

iii. None of the information on this page is useful. You will have to manually query the data. On the right, click the link for "Get More Data".

iv. Read through some of these poorly documented files to find the long commandline you need to use to extract the daily temperature data for the Kamloops weather station. You will have to modify the commandline for **daily** data from the appropriate **station**, and for the **years** you need to cover the consumption file in the last step. You can download more data than you need and delete the unnecessary information after. The code information given is for a bash script. In linux and the mac terminal, you should be able to copy the code directly and the file downloads into the terminal's working directory. In windows, activate bash first by typing `bash` at a terminal and then follow the code.

v. Merge this file with the electricty consumption file that I am providing for the dwelling. You can do this in R, or through a spreadsheet, save the file, and then read it back into R. Ensure that you match all 1094 days from the provided electricity data with your weather data. You can save this into a data set to read as a csv or tsv.

1.1 Questions

1. There are multiple temperature measurements from the weather data. State which one you'll use as your outdoor temperature measurement for this investigation. (There is no right answer, I just need to know which you use so I can follow.)

2. Create a simple linear regression model where the response is the consumption and the predictor is the temperature variable. Present the equation of the line from your results and justify whether consumption increases, decreases, or is not significantly related to the outside temperature. Keep this answer short and don't include any information I did not ask for.

3. Find the point estimate (not intervals, just give me the the $\hat{y}$ values for) the average electricity consumption when outside temperatures are -40C, -20C, 0C, 20C, 40C.

4. Create a plot of the data and overlay the equation of the line. Just based on the graph, you'll likely feel unsatisfied. In a couple sentences, discuss whether which of the estimated values from the previous questions seem reasonable and which you feel unsatisfied with.

5. As in lecture 2.2, create a new regression model with an additional term that does a better (but not necessarily the best) job to addresses your dissatifaction from the previous model.

    a. Identify the equation of your new linear model.
    b. *Very carefully* use the regression model to answer the naïve question "does the resident uses more or less electricty when the outdoor temperature rises?". Your answer should still be one sentence, or two short sentences.

6. Estimate the average electricity use for -40,-20, 0, 20, 40C again, but with your new model. Discuss which provide better or worse estimates than the first model.

7. Create another scatterplot and this time overlay your new linear model. Clearly identify your 5 estimated values and discuss whether the fit is better.

2 Part 2: Bushes data

An analysis of various ornamental bushes is conducted on plants which are left alone and untrimmed with common light, water, and nutrients after 3 years.

Your data is in this janky csv.

1. Ignore the find the equation of the simple linear regression line modelling height of the plant as a function of the width of the plant. Create a plot that again ignores the species and overlay your regression line. Discuss whether the result of the line matches what you see in this plot, how to interpret the slope, whether it's significant.

2. Create a new regression line using two main effects: the width and the species. This time, create the plot distinguishing species. Draw each of the submodels for each species.

3. Interpret the coefficient corresponding to the width of the second regression line, whether it's significant, and what makes the answer very very weird compared to your first regression line.

4. How would you answer "do wider bushes correspond to taller or shorter bushes?" in these data?