STA130H1S – Winter 2021

Week 8 Problem Set

N. Moon & S. Caetano

Instructions

How do I hand in these problems for the 11:59 a.m. ET, March 11th deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link: https://q.utoronto.ca/courses/206597/assignments/563151) by 11:59 a.m. ET, on March 11th. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Question 0

The purpose of this question is to guide you through the steps of how to access the data for your final project, how to knit the example template for the first time, and how to save yourself a personal copy to work from as you complete your project. It should not take you longer than 5 minutes to complete (unless you have a very slow internet connection!). There is nothing for you to write here for this question - just some instructions to get you set up for your final project.

(a) Go to the project overview page on Quercus (https://q.utoronto.ca/courses/206597/pages/project-overview-2) - in the table at the top, you'll see a link to access the JupyterHub "project" which includes the project data and a .Rmd template for you to work from. Click on this link, open the project folder (in the Files pane in the bottom right corner), and knit the file called "project-template.Rmd"

(b) Make a copy of the "project-template.Rmd" file, with your name in the filename, and knit this new file. For your project, you'll be editing this customized version of the file, so that you can refer to the original template for details about formatting and such. In your actual file, you'll be deleting the instructions and replacing this with your project content.

(c) In your new .Rmd file (the one you renamed), use the glimpse function to take a first look at the data. We'll talk more about the project in class on Monday March 8th. If you aren't / weren't able to attend, I suggest watching the Zoom recording to make sure you get all the details so you can get started! Project deadlines are posted on Quercus (https://q.utoronto.ca/courses/206597/pages/project-overview-2), but the next big deadline is your project slides, due on April 1st.

Once you're done with Question 0, you'll need to navigate back to the Week8 problem set folder by using the File pane in JupyterHub (bottom right corner).

Question 1

In a 1965 article, George Moore predicted the number of transistors on processors would double every year. He projected that level of growth would continue for at least another decade. A decade later, in 1975, he revised the forecast to doubling every two years. This is now commonly known as Moore's law.

The `processors.csv` dataset contains a sample of data on processors, scraped from Wikipedia about the number of transistors in central processing units (CPUs) and general processing units (GPUs). It also shows the name of the processor and the year it was introduced.

```
processors <- read_csv("processors.csv")
glimpse(processors)
```

```
## Rows: 299
## Columns: 4
## $ processor        <chr> "Intel 4004 (4-bit, 16-pin)", "Intel 8008 (8-bit, ...
## $ year             <dbl> 1971, 1972, 1973, 1973, 1974, 1974, 1974, 1974, 19...
## $ transistor_count <dbl> 2250, 3500, 2500, 11000, 3000, 4100, 6000, 8000, 4...
## $ unit_type        <chr> "CPU", "CPU", "CPU", "CPU", "CPU", "CPU", "CPU", "...
```

(a) Create an appropriate plot for the number of transistors per year, faceted by `unit_type`.

(b) Briefly explain whether you think it is appropriate to fit a straight line through this plot as it is displayed?

(c) Add a new variable called `log_transistors` to the dataset. You can use `mutate()` and the `log()` function.

(d) Plot the association between `log_transistors` and year, faceted by `unit_type` and use `geom_smooth(se=FALSE, method="lm")` to add a line of best fit to both plots. Describe this association in each plot.

Note: You will learn more about transforming variables in future courses and are not required to be able to explain why we've done this here. You can just treat `log_transistors` as we have other variables in class and refer to it as "the log number of transistors".

(e) Before calculating anything, do you think the correlation is stronger between log transistor count and year for GPUs or CPUs? Justify your answer.

(f) Calculate the correlation between `log_transistors` and year for CPUs and GPUs. You may find `group_by()`, `summarise()` and `cor()` to be helpful functions.

(g) Write down a simple linear regression model to predict log number of transistors in a processor based on the year it was introduced. Be sure to explain each term in the model.

Hint: If you copy math equations from another software into your .Rmd document, you'll get errors when trying to knit. Instead, you should type your math equations directly in your .Rmd document. Here are some tips and examples for doing this:

1. In a .Rmd document, math equations and symbols must be typed between dollar symbols ($).
2. If you want your equation/symbol to appear in the middle of a sentence, use only one dollar sign before and one dollar sign after. For example, we can typeset beta-hat-0 in .Rmd as $\hat{\beta}_0$.
3. If you want your equation to appear on a line on its own, type it on a separate line and put two dollar signs at the begining and the end. For example,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$$

4. A few other useful symbols you may need in this question are epsilon (ϵ), "not equal" ($\neq$), and superscripts (e.g. i^{th}).

(h) State the null and alternative hypotheses you would use assess whether year is a useful predictor of the log number of transistors in this linear regression model.

(i) Restrict your data to CPUs and use R to fit the linear model that corresponds with your line of best fit above. Report the fitted equation of the line. Interpret the regression coefficients in the context of this data AND make a conclusion about the hypotheses you defined above.

(j) Briefly explain why or why not the interpretation of the intercept is helpful for understanding Moore's Law.

(k) Get the R^2 for your model and write one sentence interpreting it in context.

Question 2 (Adapted from Exercise 7.18 in Dietz, Barr, Cetinkaya-Rundel, “OpenIntro Statistics”, Second Edition)

The `starbucks.csv` dataset contains data on calories and carbohydrates (in grams) in Starbucks food menu items.

```
starbucksdata<-read_csv("starbucks.csv")
glimpse(starbucksdata)
```

```
## Rows: 77
## Columns: 7
## $ item     <chr> "8-Grain Roll", "Apple Bran Muffin", "Apple Fritter", "Ban...
## $ calories <dbl> 350, 350, 420, 490, 130, 370, 460, 370, 310, 420, 380, 320...
## $ fat      <dbl> 8, 9, 20, 19, 6, 14, 22, 14, 18, 25, 17, 12, 17, 21, 5, 18...
## $ carb     <dbl> 67, 64, 59, 75, 17, 47, 61, 55, 32, 39, 51, 53, 34, 57, 52...
## $ fiber    <dbl> 5, 7, 0, 4, 0, 5, 2, 0, 0, 0, 2, 3, 2, 2, 3, 3, 2, 3, 0, 2...
## $ protein  <dbl> 10, 6, 5, 7, 0, 6, 7, 6, 5, 7, 4, 6, 5, 5, 12, 7, 8, 6, 0,...
## $ type     <chr> "bakery", "bakery", "bakery", "bakery", "bakery", "bakery"...
```

(a) Produce a plot that shows the association between carbohydrates and calories in Starbucks menu items. Describe this association.

(b) Before calculating anything, estimate the correlation coefficient between carbohydrates and calorie content in Starbucks menu items based on the plot you produced in (a). Justify your answer.

(c) Calculate the correlation between carbohydrate and calorie content of Starbucks menu items. How does this compare to your estimate in part (b)?

(d) Write down a simple linear regression model to predict calories based on carbohydrate content of Starbucks menu items. Be sure to explain each term in the model.

(e) State the null and alternative hypotheses you would use assess whether there is a linear association between the two variables.

(f) Use R to fit the regression model in (d) to these data. Report the fitted regression line. Interpret the regression coefficients in the context of this study AND make a conclusion about the hypotheses you defined above.

(g) Add the estimated linear regression line that you calculated in (f) to the plot you generated in (a). Compute the coefficient of determination, R^2. How well does the linear regression line seem to capture the relationship between `carb` and `calories`? Justify your answer.

(h) Based on the Starbucks data, create a new dataset called `starbucks_lunch` which only contains food items which are of one of two types: “sandwich” and “bistro box”. Create a boxplot comparing the distribution of calories for these two types of items.

(i) Fit a linear regression model to test whether there is a difference in mean calories for items of type “bistro box” and items of type “sandwich”. Write a sentence summarizing your conclusion.

Part 2

You have just been hired as the first statistician for a start up company. Congratulations! You were hired because the owners are looking to add more credibility to their work by testing if their yo-yos are significantly better than their competitors. This is based on whether children enjoy their free time more with their yo-yo or their major competitor's (Mr. Jones). Each child enrolled in the study received either their yo-yo or one from their major competitor. An adult in the household reported how much enjoyment their children got from playing with the yo-yo on a scale from 1 to 100.

The big boss (Daisy Smith) has heard about how her competitors use linear regression for their own studies and wants you to use the same. However, the big boss does not actually know what linear regression is. Therefore, you need to craft an email explaining to Miss Smith what linear regression is, and whether it would be appropriate to use it for the proposed analysis. You should write out a hypothetical linear regression equation for the experiment and define what each part of the equation is in simple terms. Make sure to use a minimum of 2 vocabulary words and define the words for a nonstatistical audience.

Some things to keep in mind

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 but less than 400 words.
- Use full sentences.
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang or emojis should appear).
- Be specific. A good principle when responding to a prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt).

Vocabulary

Linear relationship - Approximately linear - Non-linear - Correlation - Slope - Intercept - (Simple) linear regression - Regression model - Parameter - Regression coefficients - Fitted regression line - Explanatory/independent variable - Dependent variable - Measure of model fit - Coefficient of determination - Error - Residual - Least-squares - Least-squares estimator