

STAC32

Assignment 7

Due Thursday November 7 at 11:59pm

To begin:

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.2.1 --
## v ggplot2 3.2.1    v purrr  0.3.2
## v tibble  2.1.3    v dplyr  0.8.3
## v tidyr   1.0.0    v stringr 1.4.0
## v readr   1.3.1    v forcats 0.4.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(broom)
```

You may or may not need **broom**. If running `library(broom)` gives you an error like “no such package”, you’ll need to install it first, with `install.packages`.

1. Work through Chapter 14 of PASIAS. There are *lots* of questions there, so work through enough of them to get the idea.

Hand in the next (rather long) question:

2. Are graduation rates better from colleges where students enter with higher SAT scores? How do student-related expenditures (tuition, textbooks etc.) come into the picture? A sample of US colleges and universities was taken (only of institutions with between 10,000 and 20,000 students). For each college or university, three things were recorded:
 - the median SAT score (of current students when they first enrolled)
 - student-related expenditure in \$ per full-time student
 - six-year graduation rate, in percent. (This is the percentage of students who graduate within six years of first enrolling at the institution.)

The data are in http://www.utsc.utoronto.ca/~butler/assgt_data/graduation-rates.csv as a CSV file.

- (a) (2 marks) Read in and display the data.
- (b) (2 marks) Make a suitable plot of graduation rate and median SAT score.
- (c) (2 marks) Comment briefly on what you learn from your plot. (Hint: linear or curved? Up or down? Strong or weak?)
- (d) (3 marks) Make a (similar) suitable plot for graduation rate against expenditure. Comment briefly on what you see.

- (e) (2 marks) Fit a (multiple) regression predicting graduation rate from the other two variables and display the results.
- (f) (3 marks) One of the **expenditure** values was much lower than the others. It turns out that this value was an error. Create a new data frame that *excludes* this observation, and give the new data frame a name.
- (g) (3 marks) Re-run your model of (e), but on your new data set. What would you say is the most important difference between the output of the two models? Explain briefly.
- (h) (3 marks) Produce the one of the standard residual plots that most clearly indicates a problem with the most recent regression, and describe the problem it indicates. (Hint: look at plots of residuals: normal quantile plot, against fitted values, against each of the explanatory variables including non-significant ones. Also note that you may need to do some extra work to obtain a data frame with the data and the stuff from the regression in it.)
- (i) (4 marks) How might you modify your previous regression model to take care of the problem you found in the previous part? Make that modification, re-fit the model, and describe the principal change that you see in the results.
- (j) (3 marks) Plot the residuals from your last regression against expenditure. Does the problem seem to have been solved?