STAD29 / STA 1007 assignment 6

Due Tuesday February 18 at 11:59pm on Quercus

Packages for this one:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.2.1     v purrr   0.3.3
## 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()
```

Hand in problems 2 and 3.

1. Work through the rest of Chapter 22 of PASIAS.

2. R has a number of built-in data sets. One of them is called `PlantGrowth`. This consists of 30 observations from an experiment to compare plant yield (measured by the dried weight of plants) under two treatment conditions and a control condition. We have two research hypotheses to consider: whether the average of the two treatments is different from the control, and whether the two treatments differ from one another.
   (a) (1 mark) Display (some of) the data set.
   (b) (2 marks) Make a suitable plot of the data.
   (c) (2 marks) Why is this a situation where contrasts would be helpful? Explain briefly.
   (d) (3 marks) Set up contrasts for the two hypotheses of interest. That is, define two vectors with mnemonic names whose values reflect what you want to compare with what. (To get the order right, think about what order the treatment groups came out on your plot.)
   (e) (2 marks) Demonstrate that your two contrasts are orthogonal.
   (f) (2 marks) Use your two contrasts to set up for `lm` to test them via `summary`.
   (g) (2 marks) Fit an appropriate model and display its summary.
   (h) (3 marks) What do you conclude? Explain briefly why that makes sense by looking at the plot you drew earlier.

3. Back in STAC32, we had some children who were learning to read. We now have some more, but the experimenters were concerned that the total income of each child's family might also affect the child's reading score. (You might imagine that a larger family income, other things being equal, would be associated with a higher (better) reading score.) There are, this time, four reading methods, labelled `method1` through `method4`. The data for this study are in `http://ritsokiguess.site/STAD29/reading_again.csv`.
   (a) (2 marks) Read in the data and display (some of) the data frame.
   (b) (3 marks) Make a suitable plot of the data. Add regression lines for each method (*without* the grey envelopes). Bear in mind that we are trying to predict reading score from everything else.

(c) (2 marks) Describe any effects of income and reading method on reading score that you see on the graph.

(d) (2 marks) Run an analysis of variance of reading score as it depends on reading method. Display the results.

(e) (3 marks) Compare, using a suitable graph or numerical summary, the reading scores for the different reading methods. What is your main conclusion?

(f) (3 marks) Repeat the previous part, but this time comparing the family income by reading method (and not the reading scores). Again, comment briefly.

(g) (3 marks) Run a suitable analysis of covariance, and use `drop1` with `test="F"` to test the significance of the two explanatory variables. What do you conclude?

(h) (3 marks) Compare the P-values for `method` from the analysis of covariance in the previous part, and the analysis of variance you did earlier. Which one do you think is more trustworthy? Explain briefly.

Notes

[1]I'm always amused at how Americans put all Asians into one group.
[2]Which is actually the last column, confusingly.