STAD29 / STA 1007 assignment 5

Due Tuesday Feb 11 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 questions 2 and 3 of the following:

1. Work through, or read through, chapter 22 of PASIAS. This will prepare you for the two questions following.

2. The number of hours that a battery operates might depend on the material it is made of, and the temperature at which it is operated. In an experiment, three materials (labelled A, B, and C) were tested, and three temperatures: Low (-10 ˚ C), Medium (20 ˚ C) or High (45 ˚ C). Twelve batteries were randomly selected from each material type and were then randomly allocated to each temperature level. The resulting life of all 36 batteries is shown in `http://ritsokiguess.site/STAD29/batteries.txt`, with the data values separated by spaces.

   (a) (2 marks) Read in and display (some of) the data.

   (b) (3 marks) Which order are the temperatures in, as far as R is concerned? Does that make sense? If not, put them in the right order by creating a column in the data frame that has the right order. (Hint: what order are they in in the data frame?)

   (c) (2 marks) Make a suitable plot of these data, given the number and types of variables you have. Put temperature on the x-axis.

   (d) (3 marks) Make an interaction plot, again putting temperature on the x-axis.

   (e) (2 marks) What do you conclude from your interaction plot? By looking at your first plot, explain briefly why your conclusion from your interaction plot makes sense.

   (f) (3 marks) Run a suitable analysis of variance, including interaction, and display the results. Was your interaction significant?

   (g) (2 marks) Explain briefly why simple effects would be a useful technique for this data set, and give an example of a comparison you would be able to make with them.

   (h) (4 marks) Find the simple effects of material at each temperature. That is, for each temperature, compare the materials at that temperature using `aov` and (if necessary) Tukey, and state your conclusions in the context of the data.

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]Otherwise, why make it?

[2]It is not enough to be able to parrot the definition; you need to *understand* the definition well enough to be able to use it.

[3]You should not look at the Tukey comparisons when the ANOVAs were non-significant, but here none of those are significant anyway.

[4]As for regression summaries, the Tukey output is for looking at rather than computing with.