STA303 Summer 2018

Assignment 1

Due July 18th, 2018

Instructions

The assignment: this assignment consists of analyzing two datasets and briefly writing up your results. Specifically, for each dataset described below,

- Read the chapter of Stat Labs corresponding to the dataset (posted on Quercus)
- Write a paragraph that describes the problem at hand, how your anlaysis addresses this problem, and your conclusions. You may reference tables, numbers and plots contained in your appendix. This write-up must not exceed one half-page, 12-point times new roman font, single spaced, 8.5 x 11 inch paper, 1 inch margins. Failiing to adhere to the formatting requirements will result in severe mark penalties.
- Write an appendix of no more than 3 pages in length, containing all analyses needed to support your write-up. This is where you do summary statistics and plots, fit models and check assumptions, and justify your analysis choices.

Instructions for submission:

- Your assignment must consist entirely of one executable `.Rmd` file that we can knit in `RStudio` by pressing `Cmd+Shift+K`.
- Submit the `.Rmd` file via Quercus, to the "Assignment 1" assignment
- Knit your file into a `.pdf`, and submit this via crowdmark. You will receive your personalized crowdmark link to your U of T email.

Grading Criteria

This assignment will be graded based on the effort and reasonableness with which you analyze the datasets and answer the questions provided. It is not to be done by simply recreating topics discussed in lecture with no context. You will be graded on

- The clarity of your explanation of why you chose to analyze the data in the way you did. It's not enough to make a boxplot (you can just copy my code from class for that); you are being marked on your understanding of why a boxplot is appropriate, and what you are answering/investigating by creating it
- The quality and context of your graphical summaries. All plots should be clearly labelled (axes, title, and usually subtitle), and through the data and axis labels and titles, should clearly describe what you are trying to describe, with no additional contextual information needed. You have to use ggplot. Plots done in any other platform (e.g. base R) will not be graded.
- The thoroughness of your model checks. Try transformations of the response and (where appropriate) the covariates; investigate qualitatively and/or quantitatively which subset of variables is the most reasonable; check your model assumptions graphically and numerically. Clearly explain each assumption

that you are checking, how you are checking it, and what the impact would be if the assumption were broken

- Your conclusions. Are they reasonable given the data and the model? Did you answer the original question, or provide a convincing argument as to why the question cannot be answered using the data/methods described?

These are just guidelines to help you understand the attitude with which the assignment is marked; the exact rubric that the TAs will use is posted on Quercus.

All graphs must be done using the **ggplot()** function in the **ggplot2** package, as described many times in lecture. Base **R** graphs will not be graded by the TAs. You don't have to use **dplyr** for data manipulation and summarizing, but you should.

You may use code snippets from lecture with citation. All other work must be your own. You may discuss the assignment with your classmates in general terms, but you should not use any code or anything else written or created by another person. This is a 3rd year course; you are responsible for understanding the University's policy on academic misconduct.

Datasets

1. Stat Labs, chapter 8: Calibrating a Snow Gauge. Information about these data can be found on the course webpage, file statlabs-chapter8-snowgauge.pdf. The problem is calibrating a gauge used to estimate snow density; this information is used by environmental officials to assess flood risk, among other things.

The data is posted on the course webpage file snow-gauge.dat. It is also available from the Stat Labs Data page, and can be read in as follows:

```
readr::read_table("https://www.stat.berkeley.edu/~statlabs/data/gauge.data",col_types = "dd")
```

```
## # A tibble: 93 x 2
##    density  gain
##      <dbl> <dbl>
##  1   0.686  17.6
##  2   0.686  17.3
##  3   0.686  16.9
##  4   0.686  16.2
##  5   0.686  17.1
##  6   0.686  18.5
##  7   0.686  18.7
##  8   0.686  17.4
##  9   0.686  18.6
## 10   0.686  16.8
## # ... with 83 more rows
```

The statistical problem you should focus on is that of estimating mean **density** at a given **gain**. Some points to get you started:

- Is there a relationship between **gain** and **density**, and what does it look like?
- Can you build a simple statistical model to estimate mean **density** at a given **gain**? Can you interpret this model in the context of the problem? Are any data transformations or modifications required?
- What are the limitations and assumptions required for this model to give valid estimates? How accurate are these estimates?
- In the context of the problem -gauge calibration- is a statistical model necessary? Is there any other way you could use these data to answer the question?
- What do you conclude- if the scientists asked you whether you could estimate the mean **density** at a particular **gain** reading, what would you say to them?

2. Stat Labs, chapter 7: Dungeness Crab Growth. Information about these data can be found on the course webpage, file statlabs-chapter7-crabs.pdf. While the chapter talks extensively about premolt and postmolt shell size, we are going to use the other dataset in this chapter, which relates shell size to whether the shell was clean or not. This is described on page 142, "The second set of data…". In this question we are going to consider whether mean carapace width is different among crabs who have molted during the most recent molting season, and those who have not.

The data is posted on the course webpage, file crabpop.dat. It is also available from the Stat Labs Data page, and can be read in as follows:

```
readr::read_table("https://www.stat.berkeley.edu/users/statlabs/data/crabpop.data",col_types = "dc")
```

```
## # A tibble: 362 x 2
##     size shell
##    <dbl> <chr>
##  1  117. 1
##  2  117. 1
##  3  118. 1
##  4  120. 1
##  5  120. 1
##  6  120. 1
##  7  121. 1
##  8  123. 1
##  9  124. 1
## 10  124. 1
## # ... with 352 more rows
```

The statistical problem you should focus on is that of estimating mean carapace size for crabs who have recently molted, for those who have not, and telling whether any observed differences between these numbers are meaningful. Some thoughts to get you started:

- What do the two variables in the dataset represent? The documentation does not actually explicitly say. Can you figure it out from the broader context given in the chapter? Like it or not, this task of figuring out what the data even is before analyzing it is very common in practice.
- Are the two variables in the data related? How do you assess the relationship between a continuous and categorical variable? Can you use your answer to the previous bullet to interpret this relationship in the context of the problem?

- What type of statistical technique might you use to estimate these means, and tell whether their difference is meaningful? What assumptions are required, and do you think they are satisfied?
- What do you conclude?