STA303 Summer 2018

Assignment 2

Due August 8th, 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, 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 sumbission:

- 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 should make all code visible in the knit document. 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. Nodal Involvement. From Davison (2003), Statistical Models, example 10.18 on page 490 in chapter 10 (Nonlinear Regression Models). See also the reference therein. You can get the book electronically from the U of T library website. Data can be obtained in R as follows:

```
# Load SMPracticals package
# install.packages("SMPracticals")
suppressMessages({
  suppressWarnings({
    library(SMPracticals)
  })
})
data(nodal)
dplyr::glimpse(nodal)
```

```
## Observations: 53
## Variables: 7
## $ m     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ r     <dbl> 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,...
## $ aged  <fctr> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1...
## $ stage <fctr> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1...
## $ grade <fctr> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0...
## $ xray  <fctr> 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ acid  <fctr> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1...
```

```
# ?nodal
```

The problem is to tell, given the binary variables available, whether a patient's prostate cancer has spread to the lymph nodes. Here are some thoughts to get you started:

- Can we make a causal conclusion about the relationship of each predictor with nodal involvement, based on these data?
- Is there any apparent relationship between each predictor and nodal involvement?
- Can you develop a model that incorporates these predictors to try to estimate nodal involvement for new patients? What is an appropriate model, and what are the assumptions? Are they satisfied? How would you assess whether a simpler model could potentially be better than a model including all the predictors?
- What conclusions can you draw? Do you feel comfortable recommending a treatment strategy based on your analysis, or is more data needed? Why?

2. Smoking, Age and Death. From Davison (2003), Statistical Models, example 6.18 on page 258 in chapter 6 (Stochastic Models). Data can be obtained in `R` as follows:

```
# Load SMPracticals package
# install.packages("SMPracticals")
suppressMessages({
  suppressWarnings({
    library(SMPracticals)
  })
})
data(smoking)
dplyr::glimpse(smoking)
```

```
## Observations: 14
## Variables: 4
## $ age    <fctr> 18-24, 18-24, 25-34, 25-34, 35-44, 35-44, 45-54, 45-54...
## $ smoker <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0
## $ alive  <dbl> 53, 61, 121, 152, 95, 114, 103, 66, 64, 81, 7, 28, 0, 0
## $ dead   <dbl> 2, 1, 3, 5, 14, 7, 27, 12, 51, 40, 29, 101, 13, 64
```

```
# ?smoking
```

The analysis task is: how are the three factors in this study related to each other? See the reference mentioned in the problem description in the book for more information. In your answer, you should describe any possible dependence among these factors using terms like Mutual Independence, Conditional Independence, and/or Joint Independence, like we talked about in the lectures on contigency tables. Some thoughts to get you started:

- Do you see any unintuitive relationships, for example between smoking and mortality? How can you explain this?
- What are some of the dangers of ignoring/marginalizing over variables when making conclusions? For example, what would have happened in this study if the investigators didn't measure age? How does that relate to drawing conclusions from observational studies in general?