Lab 1 - MATH 141

Due Date: Tuesday 09/21/2021, 11:59 PM

Instructions:

- Please provide complete answers/solutions for each question/problem.

- If it involves mathematical computations, please provide your reasoning and/or detailed solutions.

- Please use the `echo = TRUE` option in each R code snippet to show your code.

- **Please save your work as pdf file(s), don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

Example R Code Snippets:

```r
# loading packages
library(tidyverse)
library(openintro)
```

```r
# plotting
numbers <- c(1, 3, 6, 4, 9)
plot(numbers, type="o", col="blue")
title(main="Numbers", col.main="red", font.main=4)
```
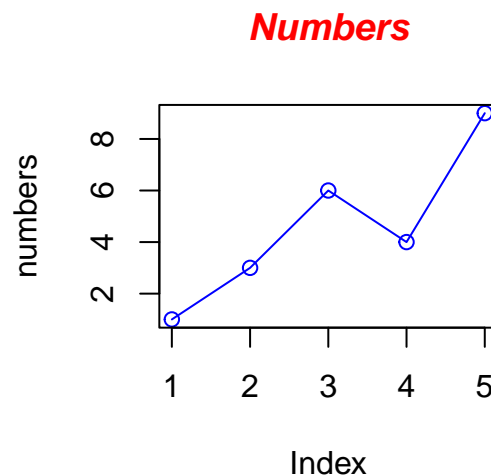

Figure 1: Figure caption

To reference a figure, use `\@ref(fig:#)` where the `#` indicates the snippet name. Fig. 1 is an example.

A. Introduction to R

I. Iris Flowers

1. Load the `iris` dataset. Note that this dataset is in the `datasets` package which is already included in the base R installation.

2. How many rows and columns does this data set have?

3. Produce 3 scatterplots and describe them. What pattern(s) do you observe? Does it look linear or nonlinear? Are any points clustered together?

4. Produce 3 histograms and describe them. What shape(s) does it show? Does it have one peak or multiple peaks?

II. Economic Regression

1. Load the `longley` dataset. Note that this dataset is in the `datasets` package which is already included in the base R installation.

2. How many rows and columns does this data set have?

3. Produce 3 scatterplots and describe them. What pattern(s) do you observe? Does it look linear or nonlinear? Are any points clustered together?

4. Produce 3 histograms and describe them. What shape(s) does it show? Does it have one peak or multiple peaks?

III. United States Counties

1. Load the county dataset from the csv file.

2. How many rows and columns does this data set have?

3. Produce 2 scatterplots and describe them. What pattern(s) do you observe? Does it look linear or nonlinear? Are any points clustered together?

4. Produce 2 histograms and describe them. What shape(s) does it show? Does it have one peak or multiple peaks?

B. Rscripting and Rmarkdown

I. Linear Function

Consider the linear function written below with parameters a and b.

$$y(x; a, b) = ax + b$$

where x is the independent variable and y is the dependent variable. Here, the slope of the line is a and intercept is b.

1. Write an R function which outputs the dependent variable y and takes in the dependent variable x, and the parameters a and b.

2. Using 3 different values of a and b (A:$a = 1, b = 1$, B:$a = 1.25, b = 1.2$, and C:$a = 2, b = 1.5$), create a dataframe with columns "line", "x", and "y". Use the `seq` command to create a vector of x values from 0 to 1 with length 10. Use the `rep` command to generate x values for the three groups. Use your R function to generate y values using your x values as inputs. The "line" column should contain the line groups A, B, and C. Use `glimpse` to show your dataframe.

3. Use `ggplot` to plot the lines on the same figure with proper line group labels.

II. United States Counties

Consider the `county` data set, which can be found in the `usdata` R package. Also, the county dataset can be accessed as a csv file.

1. Use `ggplot` to plot histograms of "pop_change" variable with categories from the "metro" variable. Make sure to label the x and y axis properly. What can you tell about the change in population whether the county has access to a metro or not?

2. Use `ggplot` to plot histograms of "per_capita_income" variable with categories from the "median_edu" variable. Make sure to label the x and y axis properly. Describe the distributions between the levels of the "median_edu" variable. Is there a strong association between "median_edu" and "per_capita_income"? Explain why.

3. Use `ggplot` to plot a scatterplot of "unemployment_rate" versus "poverty" variables with categories from the "median_edu". Make sure to label the x and y axis properly. Is there an association between "unemployment_rate" and "poverty"? Are there any differences across the "median_edu" levels?

4. Create a subset of the data where we only take rows with the states Washington, Oregon, and California.

5. Using the subset you just created, create boxplots of "pop_change" variable with categories from the "state" variable. Make sure to label the x and y axis properly. Based on the medians shown on the boxplots, which state has the lowest and highest population change?

6. Using the subset you just created, create boxplots of "per_capita_income" variable with categories from the "median_edu" variable. Make sure to label the x and y axis properly. Based on the medians shown on the boxplots, which state has the lowest and highest per capita income?