Lab 2 - MATH 141

Due Date: Tuesday 10/05/2021, 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:

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

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

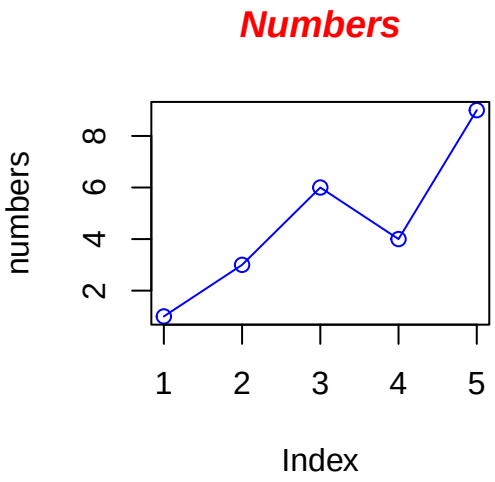

Figure 1: Figure caption

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

A. Simple Linear Regression and `dplyr`

I. Linear Regression on 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. Use `dplyr` to make subsets (Species: setosa, versicolor, and virginica) of data using variables `Sepal.Length` and `Petal.Length`. You should have three subsets. One for each species.
2. For each species find the best fit linear model to predict `Petal.Length` using `Sepal.Length` as the predictor. You should have three Linear models. One for each Species. Include writing the linear model equations with their corresponding sample statistic (intercept and slope). For example, write the equations for $\hat{y}_{setosa}$, $\hat{y}_{versicolor}$, and $\hat{y}_{virginica}$. Interpret the slopes of the models for each species and compare them with each other.
3. Using your fitted models. Produce a scatterplot which contains the data points colored according to Species. Add a line for each linear model for each species on the same scatterplot. Make sure to put legends and labels correctly.
4. For each subset you made in problem 1. Compute the correlations. Describe the correlations in this context. Does each correlation reflect what's on the scatterplot and compare them with each species?

II. Linear Regression on Housing Prices

1. Load the `duke_forest` data, which is available using the `openintro` package. Below is a table of the description of the variables of the `duke_forest` data set. For each variable, indentify whether they are numerical (discrete or continuous) or categorical (nominal or ordinal).

Variable	Description
price	Sale price, in USD
bed	Number of bedrooms
bath	Number of bathrooms
area	Area of home, in square feet
year_built	Year the home was built
cooling	Cooling system: central or other
lot	Area of the entire property, in acres

2. Produce a paired scatterplots using the `pairs` function in R. You must use the numerical variables.
3. Compute the correlations of the `bed`, `bath`, `area`, `year_built`, and `lot` variables with the `price` variable. Which explanatory variable has the highest correlation? Does it reflect a consistent pattern in the scatter plot?
4. Let the response variable (or the outcome) be the `price` variable. Create a linear model with a predictor chosen with the best coefficient of determination with the `price` variable.
5. Produce a scatterplot and residual plot for the linear model with the best coefficient of determination in problem 4. Interpret the slopes and explain the results according to this context. Based on the residuals, identify any outliers and explain why we need to resort to a more complicated model.

B. Randomizations, Simulations, and Sampling

I. Sampling from `iris` data set.

Use the `iris` data set to perform the following sampling procedures.

1. Sample 10 observations using simple random sampling.
2. Using the species as strata, sampling 30% of observation from each species.
3. Using the species as clusters, sample one species.
4. Using the `Petal.Length` variable, perform a sampling procedure where you sample 10 observations and compute its mean. Repeat this procedure for 1000 trials while recording the means. Plot the distribution of the means. What is the shape of the distribution?
5. Repeat the procedure from problem 4 but with 100 observations per sample and 2000 trials. Plot the distribution of the means. Is the distribution much more refined from the previous plot?

II. Ball Sampling from Urns.

Download the R sript `sampling-from-urns.R` and use the `source` command to initially run the R script. Then, answer the following questions.

1. Provide your initial output of the R script. Describe what the output shows you.
2. Provide output from your code (the figures comparing both probabilities) for at least these two different scenarios:

(a) `numTrials = 10^2`

(b) `numTrials = 10^4`

3. Give a short explanation of what is shown in your graphs and what you have learned from this exercise. (Example: Do you feel you agree well with the theoretical results? Which one is better? How many trials do you think you need to get a "good" agreement to the true probability?)