GR5058 Assignment 2

Due: Tuesday, October 9, 2018 by 6PM

Instructions

Create a RMarkdown file that contains the answers to the following questions. Upload both the RMarkdown file and the HTML or PDF file it generates to Canvas when you are finished. Work on this problem set by yourself, but you can ask questions on CampusWire. Just remember to click on the options that says your question will be visible to "Instructors and TAs only".

Matrix Algebra

Question 4 on pages 301 – 302 of Moore and Siegel. You can use R.

Inverses of Matrices

Is it true that

$$\left(\mathbf{W}+\mathbf{x}\mathbf{y}^\top\right)^{-1} \quad = \quad \mathbf{W}^{-1} - \frac{\mathbf{W}^{-1}\mathbf{x}\mathbf{y}^\top\mathbf{W}^{-1}}{1+\mathbf{y}^\top\mathbf{W}^{-1}\mathbf{x}}$$

presuming that all matrices and vectors are conformable for multiplication and that $\mathbf{W}^{-1}$ exists? Show your work as to why or why not. Although you do not have to use boldface letters in your answer, it may help to use "math mode" in your RMarkdown file where mathematical expressions are contained within $$...$$, $\mathbf{W}^{-1}$ can be written as W^{-1} and $\mathbf{y}^\top$ can be written as y^{\top}. Hint: $(\mathbf{A}+\mathbf{B})(\mathbf{C}+\mathbf{D}) = \mathbf{A}(\mathbf{C}+\mathbf{D})+\mathbf{B}(\mathbf{C}+\mathbf{D}) = \mathbf{AC}+\mathbf{AD}+\mathbf{BC}+\mathbf{BD}$. You can also write it out with a pen, take a picture with your phone, and include the picture in your knitted document by putting something like

```
![inverse of matrices](pic.png)
```

in your RMarkdown file

Stratifying

Put the following line into a chunk in your RMarkdown file

```
cdc <- read.csv("https://www.openintro.org/stat/data/cdc.csv")
```

to bring a `data.frame` called `cdc` into R whose variables are described here. Use the **dplyr** package in conjunction with the conditioning variables `gender` and `hlthplan` to calculate the mean and median of the difference between `wtdesire` and `weight` for each of the subgroups defined by the intersection of these two conditioning variables. What do you conclude from the results?

Apartment Prices

```
apts <- readRDS(url('https://courseworks.columbia.edu/x/pJdP39'))
```

contains 109 randomly selected observations on apartments for purchase in a Western European city in 2005. The variable of interest is `totalprice`, which is the purchase price of the apartment in Euros. The other variables are:

1. `area` the number of square meters in the apartment
2. `zone` an unordered factor indicating what neighborhood the apartment is in

3. `category` an ordered factor indicating the condition of the apartment
4. `age` number of years since the apartment was built
5. `floor` the floor of the building where the apartment is located
6. `rooms` the total number of rooms in the apartment
7. `out` an ordered factor indicating what percentage of the apartment's exterior is exposed to the outside
8. `conservation` an ordered factor indicating how well the apartment is conserved
9. `toilets` a count
10. `garage` a count, i.e. some apartments have two garages
11. `elevator` a binary variable
12. `streetcategory` an ordered factor that captures the quality of the street the apartment building is on
13. `heating` an unordered factor indicating something about the presence or absence of (possibly central) heating for the apartment
14. `storage` a count of the number of storage rooms for the apartment

Use the functions in the **ggplot2** package to create a scatterplot between `totalprice` and one other numeric variable (possibly with transformations, rescalings, etc.) using relevant discrete variables to distinguish points by size, color, and / or symbol that is explained in a legend. What do you discover by visualizing the data this way?

Making plots

Look at help(iris) for information on the iris data.frame. Produce a box-and-whiskers plot of Petal.Length for each of the three types of Species using the ggplot2 package. Interpret your plots.

Histograms

Load the MASS package by executing

```
library(MASS)
```

The MASS package contains a `data.frame` called `Cars93`, which you can see more about by executing

```
help(Cars93)
```

Create a separate histogram using the functions in the ggplot2 package for each of the following four variables in the `Cars93 data.frame`: `Min.Price`, `Max.Price`, `Weight`, `Length`. Then, create a histogram of the `Price` variable conditioning (stratifying) on each level of `DriveTrain`, which indicates which axle(s) of the car turn the wheels. What would you conclude from these plots?