Prof. Yiping Lu
IEMS 304: Statistical Learning for Data Analysis
April 1, 2025

Homework 1

This homework is to give a brief reminder of R, RStudio, and statistical topics covered in IEMS 303.

Note: The homework is scored out of 100 points. The problems add up to 90 points, while the remaining ten points will be graded according to a writing rubric, given at the end of the assignment.

R/RStudio installation If you have not installed R and RStudio, follow the installation instructions outlined in https://posit.co/download/rstudio-desktop/. You are strongly encouraged to use R Markdown to integrate text, code, images and mathematics or you can you use the latex code we provide.

Question 1. Data manipulation using R The miles.csv file contains information about several cities and the total mileages people drive in each city every day. The variables in this dataset are:

- `City`: Name of city
- `Population`: Total population in the city
- `Roads`: Amount of roads in the city
- `Mileage`: Total mileages per day in the city
- `Area`: Size of the city

Answer the following questions:

1) Give the command you would use to load the file, and to check the number of rows and columns it has. How many rows and columns does it have?
2) Which row of the file contains information for `Ann Arbor, MI`? How did you find out?
3) How many miles are driven per person per day in `Ann Arbor, MI`? Give the R commands you use to calculate this, and report the answer to the nearest mile.
4) Calculate the number of miles driven per person per day for every city. Store the answer as a new column called `PerCapitaMiles`. Check that the new column contains the right number for Ann Arbor. Give all the commands you use to do this.
5) Make a histogram of the population of cities. Label your axes appropriately.
6) Provide an appropriate scatterplot to investigate the relationship between population and the per-capita miles driven. If necessary, consider using log scale for one or more axes. Label your axes.

Question 2. Statistical concepts and computation in R Using the same data, answer the following questions:

1) Provide summary statistics for the miles driven per capita, including but not limited to mean, standard deviation, selected quantiles, etc. Give a brief description.
2) Is the area of the cities normally distributed? How did you find out?
3) Conduct a t-test for the following hypotheses: The average mileage driven per capita is higher than 40. Choose an appropriate α value and report your conclusion. For each hypothesis test above, explain precisely what the p-value represents in your own words.
4) Create a new column `PopSL` that takes value S if the population is below or equal to the median, otherwise takes value L. Conduct a t-test for the difference in average mileage driven per capita, between S and L populations. Choose an appropriate α value and report your conclusion.
5) The Bureau of Transportation is interested in the average mileage driven by a resident. Report a 99% two-sided interval for the average mileage driven per capita. Write out the equation you have used, and define all variables.
6) Use `lm()` to fit a regression line for `PerCapitaMiles`$=\beta_0+\beta_1$`Population`$+\epsilon$. Report the best fitted line. Use the `summary()` command to report the results from the regression model. What does the R-squared value mean? What does the p-value from the F-statistic mean? What does the column of p-values `Pr(>|t|)` mean?