Econometrics 2020, Homework 1
Deadline: 6 April 2020

Please choose *one* of the following problems. Write your solution in a file named `Problem_1.R` or `Problem_2.R` depending on the number of the problem that you have picked. When you are finished, upload your code to your team's homework repository by pasting it in one of the files there: `Problem_1.R` or `Problem_2.R`.

1. In the wake of the COVID-19 outbreak data analysis can yield important insights that can help to contain the disease and ultimately to save lives. The dataset `COVID19_2020_open_line_list` contains data on patients with confirmed COVID19 infections in the United States, Japan and China (outside Hubei province).

   Note: This is an *exercise* dataset. Cases with missing or incomplete data on age and sex were removed. When age was given as an interval (e.g. 0-10 years) it was replaced with the midpoint of the interval. *Do not* draw any real-life conclusions based on the analysis here! For the full dataset refer to [Xu et al., 2020].

   ID (numeric): Case id.

   date_confirmation (date): Date when the infection was confirmed.

   sex (character): Patients' sex: male/female

   age (numeric): Patients' age in years.

   province (character): Province where the infection was confirmed.

   country (character): Country

   (a) (0 points) Download and read the dataset and store it in an object called `patients`.

   (b) (2 points) How many patients are there in the dataset?

   (c) (2 points) What are the earliest and latest date of infection confirmation?

   (d) (2 points) What is the average age of the patients?

   (e) (2 points) How many men and how many women are there in the sample?

   (f) (2 points) Plot the frequency distribution of sex using a bar-chart (see e).

   (g) (2 points) What was the age of the youngest woman?

   (h) (2 points) What was the age of the oldest man?

   (i) (2 points) Is there a difference between the average age of male and female patients?

   (j) (2 points) Compare the distribution of age between the countries using a box-plot. Interpret the plot (write your answer as a comment in the code).

2. Orley Ashenfelter, an Economics Professor at Princeton University claimed to have found a method to predict the quality of Bordeaux wine[1]. In this problem we will borrow data

[1] http://www.liquidasset.com/orley.htm

from `http://www.liquidasset.com/winedata.htm`. The dataset contains information about the prices of Bordeaux wines produced between 1952 and 1980 organised in the following columns:

Year : Year in which the wine was produced (unique).

LogPrice : Logarithm of the price of the wine.

WinterRain : Winter rain in the Bordeaux region (October to March, in ml).

Temperature : Average temperature in the region (April to September, in degrees Celsius).

HarvestRain : Harvest rain in the region (August and September, in ml).

TimeYears : Time since vintage in years.

For a short, entertaining, story about Ashenfelter's analysis and his predictions of wine prices, read the first few pages of Ayres (2008), freely available on books.google.com.

(a) (0 points) Download and read the dataset and store in a variable called `wines`.

(b) (0 points) Create a new variable (in the `data.frame`) `Price` by converting `LogPrice` back to its original scale. *Hint*: use the `exp` function.

(c) (2 points) How many years are recorded in the dataset?

(d) (2 points) What was the average temperature in 1953? Write your answer as a comment in the code.

(e) (2 points) Which was the coldest year recorded?

(f) (2 points) Compute the average wine price for hot and cold years. Define a cold year to be a year with below average temperature.

(g) (2 points) Are wines produced during cold years more valuable (on average) than wines produced during hot years?

(h) (2 points) How many years had below-average temperature? *Hint:* use the `table` function.

(i) (2 points) Compare the distribution of prices between hot and cold years using a box-plot. Interpret the plot.

(j) (2 points) Create a scatterplot for wine price and the rainfall level during harvest . Do you see any association pattern? Write your answer as a comment in the code.

References

[Xu et al., 2020] Xu, B., Kraemer, M. U. G., Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Loskill, A., Wang, L., Cohn, E., Hill, S., Zarebski, A., Li, S., Wu, C.-H., Hulland, E., Morgan, J., Scarpino, S., Brownstein, J., Pybus, O., Pigott, D., and Kraemer, M. (2020). Open access epidemiological data from the COVID-19 outbreak. *The Lancet Infectious Diseases.*