

# Homework

STATS/CME 195, Spring 2019

*Assigned: April 7, 2019*

*Due: April 19, 2018 at 11:59pm*

## Instructions

This assignment is due on Friday, April 19, 2018 at 11:59pm. You must submit your homework as a PDF on Gradescope. Your submission should be a report including all relevant code, the generated outputs, and text explaining your results. You should learn how to use R Markdown to generate your report. You can refer to chapter 3 of R Markdown: The Definitive Guide to review details on generating reports using R Markdown.

Remarks:

- This homework roughly covers the material of lectures 1-6, so you are encouraged to start working on it early and continue gradually as we advance into the class.
- If you generate random numbers, set a seed and report it, so that all your work can be reproduced.
- Do not print an entire `data.frame` or a whole vector with more than 20 entries in your homework, since it would produce an unnecessarily large file, when you render your document.
- Label your final answer to the questions asked instead of only producing output of your code. Your answers should be as concise as possible and your code should be easily understandable.
- You are welcome to work with other students, but you must write your own code, and prepare your own write up separately. If you choose to collaborate, indicate who you worked with on your submission.
- You are free to use the web or any books to look up the documentation of relevant R functions and debug errors, but you cannot look for the solution to the specific homework problems.
- If you have any questions regarding the homework, please ask on piazza or at office hours.

## Exercise 1: R Basics [20 pt]

**Material: Lecture 1.**

### a. Arithmetic operations [5pt]

Compute the following using R:

- $5.78 \log_5(7!)$ , where  $x!$  is factorial of  $x$
- $4.02 \sqrt[7]{7^2 + e^8}$
- $20 \cos(3\pi + 0.25) + 37 \sin\left(\frac{3\pi}{4}\right)$
- $\lfloor \frac{34.011\pi}{3} \binom{5}{2} \rfloor$ , where  $\lfloor x \rfloor$  means rounding to the largest integer not greater than  $x$ . where  $\binom{x}{y}$  is the notation for combination
- $8.99 \sum_{i=1}^{90} \frac{1}{i}$

## b. Matrix operations [5pt]

Generate a matrix  $A$  with 20 rows and 5 columns with entries being random uniform numbers on an interval  $[0, 1]$ . Then generate a matrix  $B$  with 5 rows and 7 columns where entries are drawn from a Gaussian distribution with mean 0 and variance 10. Use `set.seed()` function with a chosen seed (record the seed) for reproducibility. Type in `?set.seed` in the R console to learn more about the function. With the two matrices compute:

- $AB$  (a matrix product)
- multiply the 3rd row of  $A$  by the 4th column of  $B$  and compute the sum of entries in the resulting vector, then check that agrees with a corresponding term in the matrix product you computed in the previous part.
- obtain a vector which is a product of matrix multiplication between matrix  $A$  and the 4th column of  $B$ .

## c. Data frames [10pt]

Create a data frame, ‘birthdays’, which stores information on the birthdays of 5 people, either real or fictional. The data table should have columns:

1. ‘first’: first name
2. ‘last’: last name
3. ‘birthday’: the person’s birthday in format YYYY-MM-DD (“%Y-%m-%d”)
4. ‘city’: city where the person lives

Convert the birthdays to date objects using `as.Date()` function. Compute the difference (in days) between your birthday and the birthday of each of the people and append that information as a new column ‘bday\_diff’ of the data-frame.

## Exercise 2: Programming [20pt]

Material: Lecture 2.

### a. Parametric function [5pt].

- Write a function in R that evaluates the following:

$$f(\theta) = 7 - 0.5 \sin(\theta) + 2.5 \sin(3\theta) + 2 \sin(5\theta) + 3 \cos(2\theta) - 2 \cos(4\theta)$$

- Generate a vector, `theta`, equal to a sequence from 0 to  $2\pi$  with increments of 0.01
- Compute a vector  $x = f(\theta) \cdot \cos(\theta)$  and  $y = f(\theta) \cdot \sin(\theta)$  for  $\theta$  you just created.
- Plot a scatter plot of  $(x, y)$  with two vectors computed.

### b. Multiple arguments [10pt]

Write a function `time_diff()` that takes two dates as inputs and returns the difference between them in units of “years” or “days”, defined by an optional argument ‘units’, set by default to “days”. Use the function to compute time left to your next birthday separately in units of years and days.

### c. Control flow: Fibonacci numbers [5pt].

The Fibonacci sequence starts with the numbers 1 and 2, and each subsequent term is generated by adding the previous two terms. The first 10 terms of the Fibonacci sequence are thus: 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, ... . Find the total sum of **even numbers** in the Fibonacci sequence, each not exceeding one million.

## Exercise 3: Data Import/Export and transformation [20pt]

**Material:** Lecture 3 and 5.

### a. Import data [5pt]

Visit the following URL: <https://raw.githubusercontent.com/cme195/cme195.github.io/master/assets/data/share-of-people-who-say-they-are-happy.txt>

Observe the structure and format of the data. Then, use a function from the `readr` package to read the data in the URL into R. Then, find the country with the highest share of happy people in 2014.

### b. Export data [5pt]

Filter observations from the data set on happiness that correspond to years after 2000. Export the subset of the data as a tab-delimited text file to a chosen location on your computer.

### c. dplyr functions [10pt]

In this exercise we use the package, `nycflights13`, storing datasets on flights and airports in the city of New York in 2013. Install the package with `install.packages("nycflights13")` if you have not done so already, then load it with `library(nycflights13)`.

The dataset 'flights' is a tibble with 336,776 observations! To learn about the details about this dataset, type `?flights` in your R console.

Use `dplyr` functions (and the `%>%` operator) to compute, for each combination of departure airport, 'origin', destination airport, 'dest', and 'carrier': the average 'dep\_delay', the average 'air\_time' and the average ratio 'dep\_delay'/'air\_time'

For each 'carrier' report the route ('origin'-'dest') with the highest mean ratio of departure delay over air time. Now, you know which flights not to take with a given carrier.

**Note:** Since, the dataset contains missing values, when computing the averages, remember to exclude the missing values (use `'na.rm = TRUE'` in `mean()` function).

## Exercise 4: Data Visualization [20pt]

**Material:** Lecture 3-4.

The following url contains data on fossil fuel emissions for different countries between 1751 and 2014: "[http://cdiac.ess-dive.lbl.gov/ftp/ndp030/CSV-FILES/nation.1751\\_2014.csv](http://cdiac.ess-dive.lbl.gov/ftp/ndp030/CSV-FILES/nation.1751_2014.csv)"

### a. Import data with readr [5pt]

This dataset is messy, and you will need to fix the warning messages `read_csv()` returns.

- Rows 1-3 in contain information on the dataset itself, and not the variables; so after reading the data in, we need to delete these rows.
- The datasets contains characters “.” which needs to be replaced with NA for missing data.
- Rename column Total CO2 emissions from fossil-fuels and cement production (thousand metric tons of C) to something shorter, e.g. ‘Total\_CO2’

### b. Summarize data [5pt]

Compute the total yearly  $CO_2$  emissions (column `Total.CO2.emissions.from.fossil.fuels.and.cement.production..thousand.of.C.`) summed over all countries (the world total  $CO_2$  emission) using `dplyr` or an alternative function. Use the dataset to plot the World’s yearly  $CO_2$  emission in Gt.

### c. Line plots [5pt]

Find the top 10 countries with highest emission after year 2000 (including 2000).

Plot the yearly total CO2 emissions of these top 10 countries with a different color for each country. Use billion tonnes (Gt) units, i.e. divide the total emissions by  $10^6$ .

### d. Stacked plots [5pt]

Use `geom_area()` or an alternative function to generate a plot similar to the one you generated above but with emission levels stacked on top of each other (summing to the total for the ten countries) with areas colored by countries.

## Exercise 5: Linear Models [20pt]

**Material: Lecture 4 and 6.**

In this exercise we will use a dataset containing information on sales of a product and the amount spent on advertising using different media channels. The data are available from: <http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv>.

### a. Import and plot the data [5pt]

Read the data from “<http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv>”.

Then, generate a scatterplot of sales against the amount of TV advertising. Color the points by the amount of ‘radio’ advertising. Then, add a linear fit line.

### b. Simple Linear Regression [5pt]

The dataset has 200 rows. Divide it into a training set with 150 observations and a test set with 50 observations, i.e. use `sample()` without replacement to randomly choose row indices of the advertising dataset to include in the train set. The remaining indices should be used for the test set.

Fit a linear model to the training set, where the sales values are predicted by the amount of TV advertising. Print the summary of the fitted model. Then, predict the sales values for the test set and evaluate the test model accuracy in terms of root mean squared error (RMSE), which measures the average level of error between the prediction and the true response.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

**c. Multiple linear regression [5pt]**

Fit a multiple linear regression model including all the variables 'TV', 'radio', 'newspaper' to model the 'sales' in the training set. Then, compute the predicted sales for the test set with the new model and evaluate the RMSE.

Did the error decrease from the one corresponding to the previous model?

**d. Evaluate the model [5pt]**

Look at the summary output for the multiple regression model and note which of the coefficient in the model is significant. Are all of them significant? If not refit the model including only the features found significant. Which of the models should you choose?