STAT 420: Homework 01

Fall 2020, D. Unger

Due: Tuesday, September 15 by 11:30 PM CT

Contents

Directions

Students are encouraged to work together on homework. However, sharing, copying or providing any part of a homework solution or code is an infraction of the University's rules on Academic Integrity. Any violation will be punished as severely as possible.

Final submissions must be uploaded to our Compass 2g site on the Homework page. No email, hardcopy, or late submissions will be accepted.

- Your assignment must be submitted through the submission link on **Compass 2g.** You are required to attach one `.zip` file, named `hw01_yourNetID.zip`, which contains:
  - Your RMarkdown file which should be saved as `hw01_yourNetID.Rmd`. For example `hw01_dunger.Rmd`.
  - The result of knitting your RMarkdown file as `hw01_yourNetID.html`. For example `hw01_dunger.html`.
- Your resulting `.html` file will be considered a "report" which is the material that will determine the majority of your grade. Be sure to visibly include all `R` code and output that is relevant to answering the exercises. (You do not need to include irrelevant code you tried that resulted in error or did not answer the question correctly.)
- You are granted an unlimited number of submissions, but only the last submission *before* the deadline will be viewed and graded.
- If you use this `.Rmd` file as a template, be sure to remove the directions section. Consider removing `eval = FALSE` from any code chunks provided in the template, if you would like to run that code as part of your assignment.

- Your `.Rmd` file should be written such that, if it is placed in a folder with any data your are asked to import, it will knit properly without modification.
- Unless otherwise stated, you may use `R` for each of the exercises.
- Be sure to read each exercise carefully!
- Include your Name and NetID in the final document, not only in your filenames.

Assignment

Exercise 1 (Working with Vectors)

Recall the definitions of sample mean and sample standard deviation for data $x_1, x_2, \ldots, x_n$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2}$$

Consider the following vector of data.

```
x = c(1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144)
```

(a) Calculate the sample mean of `x` *without* the use of `mean()`, `sd()`, `sum()`, or `+`. Hint: Recall that vectors in `R` are column vectors. (Although when you display them, they often *look* like row vectors.) You may need to create a second vector, which is allowed, and should also look into the function `crossprod()`. Essentially, you need to use matrix algebra to recreate the formulas above. You may use `mean()` to check your answer.

```
mean(x)
```

```
## [1] 31.33333
```

The mean is 31.33

(b) Calculate the sample standard deviation of `x` *without* the use of `mean()`, `sd()`, `sum()`, or `+`. You may use `sd()` to check your answer.

Exercise 2 (Vectors and Matrices)

For this exercise, you will create several vectors and matrices, as well as perform various matrix operations.

(a) Create five vectors `x0`, `x1`, `x2`, `x3`, and `y`. Each should have a length of 30 and store the following:

- `x0`: Each element should be the value `1`.
- `x1`: The first 30 square numbers, starting from `1` (so `1`, `4`, `9`, etc.)
- `x2`: 30 evenly spaced numbers between `0` and `1`. (Including `0` and `1`. It may help to read the documentation for `seq()`.)
- `x3`: The natural log of the integers from `1` to `30`
- `y`: The result of running the following code, after creating the other four vectors:

```
set.seed(114)
y  = 5 * x0 + 1 * x1 + 6 * x2 + 3 * x3 + rnorm(n = 30, mean = 0 , sd = 1)
```

Report the value returned from `sum(y)`.

(b) Create a matrix `X` which stores `x0`, `x1`, `x2`, and `x3` as columns of the matrix. Report the value returned from `sum(X)`.

(c) Use matrix operations to create a new matrix `beta_hat` defined as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Report the values stored in the matrix `beta_hat`. Note that you can use the vector `y` as if it is a 30×1 matrix.

(d) Your `beta_hat` should store a 4×1 matrix. That is, a matrix with 4 rows and 1 column. Subset this matrix to exclude the first row, then square each element and report the sum of these results.

(e) Create a new variable `y_hat` which stores the result of the matrix operation,

$$\hat{y} = X\hat{\beta}.$$

The result will be a 30×1 matrix. Perform and report the result of the following operation,

$$\sum_{i=1}^{30} (y_i - \hat{y}_i)^2.$$

Here you can use the matrix `y_hat` as if it is a vector. Then, y_i is the ith element of y and $\hat{y}_i$ is the ith element of $\hat{y}$.

Exercise 3 (Discrete Probability)

The 2016 United States presidential election has been an interesting contest. A Fox News National Poll from June which surveyed 1004 registered voters placed former New Mexico Governor Gary Johnson in third place, which is expected, however polling in the double digits at 12%. (Third party candidates haven't performed well in recent years. The last time a third party managed to get votes in the Electoral College was 1968.)

Suppose the true proportion of registered voters that support Johnson is 12% and we obtain our own random sample of 50 registered voters. Answer the follows questions as they relate to this sample of 50 voters.

(a) What is the probability that exactly 5 of the voters are Johnson supporters?

(b) What is the probability that 10 or fewer voters are Johnson supporters?

(c) What is the probability that 37 or more voters are **not** Johnson supporters?

(d) What is the probability that between 3 and 9 (inclusive) voters are Johnson supporters?

Exercise 4 (Continuous Probability)

For this exercise, consider a random variable X which is normally distributed with a mean of 120 and a standard deviation of 15. That is,

$$X \sim N\left(\mu = 120, \sigma^2 = 225\right).$$

(a) Calculate $P(X < 95)$.

(b) Calculate $P(X > 140)$.

(c) Calculate $P(95 < X < 120)$.

(d) Find q such that $P(X < q) = 0.05$.

(e) Find q such that $P(X > q) = 0.10$.

Exercise 5 (Packages, Plotting)

For this exercise, we will use the `diabetes` dataset from the `faraway` package.

(a) Install and load the `faraway` package. **Do not** include the install command in your .Rmd file. (If you do it will install the package every time you knit your file.) **Do** include the command to load the package into your environment.

(b) How many observations are in this dataset? How many variables?

(c) What are the names of the variables in this dataset?

(d) What is the mean HDL level (High Density Lipoprotein) of individuals in this sample?

(e) What is the standard deviation total cholesterol of individuals in this sample?

(f) What is the range of ages of individuals in this sample?

(g) What is the mean HDL of females in this sample?

(h) Create a scatterplot of HDL (y-axis) vs weight (x-axis). Use a non-default color for the points. (Also, be sure to give the plot a title and label the axes appropriately.) Based on the scatterplot, does there seem to be a relationship between the two variables? Briefly explain.

(i) Create a scatterplot of total cholesterol (y-axis) vs weight (x-axis). Use a non-default color for the points. (Also, be sure to give the plot a title and label the axes appropriately.) Based on the scatterplot, does there seem to be a relationship between the two variables? Briefly explain.

Exercise 6 (Importing Data, Plotting)

For this exercise we will use the data stored in `nutrition.csv`. It contains the nutritional values per serving size for a large variety of foods as calculated by the USDA. It is a cleaned version totaling 5138 observations and is current as of September 2015.

The variables in the dataset are:

- `ID`
- `Desc` - Short description of food
- `Water` - in grams
- `Calories` - in kcal
- `Protein` - in grams
- `Fat` - in grams

- `Carbs` - Carbohydrates, in grams
- `Fiber` - in grams
- `Sugar` - in grams
- `Calcium` - in milligrams
- `Potassium` - in milligrams
- `Sodium` - in milligrams
- `VitaminC` - Vitamin C, in milligrams
- `Chol` - Cholesterol, in milligrams
- `Portion` - Description of standard serving size used in analysis

(a) Create a histogram of `Calories`. Do not modify `R`'s default bin selection. Make the plot presentable. Describe the shape of the histogram. Do you notice anything unusual?

(b) Create a scatterplot of calories (y-axis) vs protein (x-axis). Make the plot presentable. Do you notice any trends? Do you think that knowing only the protein content of a food, you could make a good prediction of the calories in the food?

(c) Create a scatterplot of `Calories` (y-axis) vs `4 * Protein + 4 * Carbs + 9 * Fat + 2 * Fiber` (x-axis). Make the plot presentable. You will either need to add a new variable to the data frame, or, use the `I()` function in your formula in the call to `plot()`. If you are at all familiar with nutrition, you may realize that this formula calculates the calorie count based on the protein, carbohydrate, and fat values. You'd expect then that the result here is a straight line. Is it? If not, can you think of any reasons why it is not?