Assignment 03

R Markdown and Probability Simulation

In this assignment you will use probability simulation to empirically generate data. You will use R Markdown to produce an HTML document that includes your responses to each of the questions on this assignment.

When we ask you to *show your work*, you can either show your work like you would in a high-school mathematics class—although we ask that you typeset all mathematics using an equation editor—or you can provide the relevant part of your script file. Please submit your responses to each of the questions below in a printed document. Also, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be resized so that they do not take up more room than necessary and should have an appropriate caption. Both of these should be done using knitr syntax in Markdown.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document using Markdown's display equations. See here for some examples of how mathematics can be typeset in R Markdown.
- All syntax included should be included in an R Markdown code chunk and be appropriately commented. Follow the Data Camp Style Guide (http://docs.datacamp.com/teach/style-guide.html as close as you can.

This assignment is worth 13 points. Each question is worth 1 point unless otherwise noted.

Independence

In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other. The are no data to be loaded, you will randomly generate the data you need to answer the questions.

Generate 1000 random observations from the standard normal distribution and write it into an object called `y`. Generate another 1000 random observations in the same way; write these into an object called `x`. Regress the variable `y` on the variable `x`.

1. Write the regression equation.
2. Using a significance level of 0.05, would you reject the null hypothesis that the slope is equal to 0? Explain.
3. Explain why we would not expect to reject the null hypothesis by referring to how the variables were created (i.e., using the idea of statistical independence).

Now you will run a simulation where you will repeat this entire process 5,000 times. The process you will repeat includes: (1) randomly generating the variables `x` and `y`; (2) regressing `y` on `x`. From each of the 5,000 trials of the simulation you will save the z-score of the slope (the estimated slope coefficient divided by its standard error). This can be automated using a loop. Below is the skeleton of the syntax you can to carry out this simulation.

```r
#Load arm library; you may need to install this first
library(arm)

# Create a vector with 5000 elements of NA (empty vector)
z.scores = rep(NA, 5000)

# This is the syntax for the loop
for(i in 1:5000){
  y = rnorm(1000, mean = 0, sd = 1)
  x = rnorm(1000, mean = 0, sd = 1)
  fit = lm(y ~ x)
  z.scores[i] = coef(fit)[2] / se.coef(fit)[2]
  }
```

For the following questions, if the absolute value of the z-score exceeds 2, you can consider it to be statistically significant.

4. Use R to compute the empirical type I error rate. Show the syntax you used, and explain how you determined this.

Regression Simulation #1

The file *beauty.csv* contains data collected from student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. The variables are:

- `prof`: Professor ID number
- `avgeval`: Average course rating
- `btystdave`: Measure of the professor's beauty composed of the average score on six standardized beauty ratings
- `tenured`: 0 = non-tenured; 1 = tenured
- `nonenglish`: 0 = native English speaker; 1 = non-native English speaker
- `age`: Professor's age (in years)
- `female`: 0 = male; 1 = female
- `students`: Number of students enrolled in the course
- `percentevaluating`: Percentage of enrolled students who completed an evaluation

These source of these data is Hamermesh, D. S. & Parker, A. M. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. The data were made available by Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press. Use these data to answer each of the following questions.

5. Fit a regression model that includes the predictors of beauty, age, sex, and whether the instructor is a native English speaker to explain variation in course ratings. Write the fitted regression equation.

Now consider the following instructor:

- **Instructor A:** A 50-year old female who is a native English speaker and has a beauty score of -1

Generate 1000 random predictions of the course evaluation rating for Instructor A by summing (1) the appropriate fitted part of the regression model and (2) a random error (simulate this).

6. Create a plot of the precited course evaluations for the Instructor A. **(2pts.)**
7. Based on your simulation results, give the limits for the empirical 95% prediction interval. Explain how you obtained these limits.

Regression Simulation #2

In the previous simulation, you accounted for prediction uncertainty by allowing the error to vary. However, you did not account for sampling uncertainty (sampling error) because the regression coefficients were the same in every prediction. This time carry out a simulation that accounts for both the sampling uncertainty and the prediction uncertainty (again for Instructor A).

8. Create a plot of the precited course evaluations for the Instructor A. **(2pts.)**
9. Based on your simulation results, give the limits for the empirical 95% prediction interval. Explain how you obtained these limits.
10. Compute the *standard error* for each of the two intervals (that from Question 7 and Question 9). Why does the interval from Question 9 have a larger standard error?

Regression Simulation #3

Now carry out a similar simulation, accounting for both the sampling uncertainty and prediction uncertainty, but this time for Instructor B.

-**Instructor B:** A 60-year old male who is a native English speaker and has a beauty score of -0.5

11. Use the simulation results from both instructors to compute the probability that Instructor A will have a higher rating than Instructor B.