ASSIGNMENT #2

EPsy 8252

This assignment is intended to review some of the skills you obtained in a regression course. The questions are adapted from Exercise #3 (Section 3.9), and Exercise #5 (Section 4.9) of (Gelman & Hill, 2007). Please submit your responses to each of the questions below in a printed document. Only provide your responses to the question asked. You do not need to include any R syntax and output unless it is specifically required in the question.

Any graphics you include should be resized so that they do not take up more room than necessary and all should have an appropriate caption. Any equations should be appropriately typeset within the document.There are 12 points possible for the assignment (each question is worth one point unless otherwise noted).

Adapted from Exercise #5

For a study of congressional elections, you would like a measure of the relative amount of money raised by each of the two major-party candidates in each district. Suppose you know the amount of money raised by each candidate; label these dollar values D_i (for the Democrats) and R_i (for the Republicans). You would like to combine these into a single variable that can be included as an input variable into a model predicting vote share for the Democrats.

Discuss the (1) advantages and (2) disadvantages of using each of the following measures:

1. The simple difference, $D_i - R_i$ **(2pts.)**
2. The ratio, D_i / R_i **(2pts.)**
3. The difference on the logarithmic scale, $\log D_i - \log R_i$ **(2pts.)**
4. The simple difference, $D_i - R_i$ **(2pts.)**

Adapted from Exercise #3

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 using the `rnorm()` function.

Generate 1000 observations from a normal distribution with mean 0 and standard deviation 1 using the syntax

```
y = rnorm(1000, mean = 0, sd = 1)
```

Generate another 1000 observations in the same way (call it `x`). Regress the variable `y` on the variable `x`.

5. Is the slope coefficient statistically significant? Explain.
6. Given that the two variables were created to be statistically independent, would we expect the slope coefficient to be statistically significant? Explain.

Now you will run a simulation where you will repeat this entire process 100 times. This can be automated using a loop. 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). If the absolute value of the z-score exceeds 2, it will be considered statistically significant. Below is the syntax to carry out this simulation.

```
library(arm)
z.scores = rep(NA, 5000)
for(i in 1:100){
  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]
  }
```

7. How many of these 5,000 z-scores are statistically significant?
8. Based on your previous answer, what is the empirical type I error rate?

References

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.