Assignment 04

Programming and Estimation

In this assignment you will exercise some of your R programming skills. Please submit your responses to each of the questions below. 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 18 points. Each question is worth 1 point unless otherwise noted.

Randomization Test

In this set of exercises, you are going to write a function to carry out a randomization test as described in Statistics Without the Agonizing Pain.

1. Write a function that carries out a randomization test for the mean difference between two groups. This function should take inputs of (1) a vector containing the grouping variable (i.e., conditions), (2) a vector containing the outcome variable, and (3) the number of permutations/randomizations to perform. The output from this function should be a vector of randomized/permuted mean differences. Include the syntax for your function in your word-processed document. Be sure that your function includes comments. **(3pts.)**

Enter the data presented in Statistics Without the Agonizing Pain into a data frame so that it can be used in your function.

2. Use your function to carry out a randomization test (using 1000 permutations) on the data. Assign the output into an object. Show the results of running the `head()` function on this object.

3. Use the `t.test()` or the `lm()` function to carry out a parametric test to examine the mean difference in the number of mosquitos between the two conditions. Present all pertinent results (i.e., t-value, df, p-value) from this analysis. **(2pts.)**

4. Compute the p-value for the observed mean difference using the object containing your function's assigned output. Show the syntax you used to compute this.

5. How does that the p-value computed in the parametric analysis compare to that from the randomization analysis?

6. Compute the estimated standard error for the randomization analysis using the object containing your function's assigned output. Show the syntax you used to compute this.

7. How does that compare to the estimated standard error computed in the parametric analysis?

8. Plot the randomized/permuted mean differences collected in your function. Draw the appropriate t-distribution from the parameteric analysis as a line on this plot.

9. How do these two distributions compare? Make reference to the shape, center, and variation. **(2pts.)**

Regression Estimation

Least squares estimation optimizes the criterion of the sum of squared error in computing the estimates for a given set of regression parameters. In this set of exercises, you are going to examine the estimates we obtain for the set of parameters for the model $\text{Wage}_i = \beta_0 + \beta_1(\text{Age}_i) + \epsilon_i$ when we change the criterion for model fit (mis-fit). In this set of exercises, you will be computing the sum of the absolute errors, $\left|Y_i - \hat{Y}_i\right|$, as a measure of the model fit (mis-fit). Use the following data set to help you answer the following questions,

```
##    wage age
## 1 12.00  32
## 2  8.00  33
## 3 16.26  32
## 4 13.65  33
## 5  8.50  26
```

10. Write a function that computes the sum of the absolute errors given the estimated the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. Include the syntax for your function in your word-processed document. Be sure that your function includes comments. **(2pts.)**

11. Provide the output of your function (i.e., the sum of the absolute errors) for the parameter estimates of $\hat{\beta}_0 = 3$ and $\hat{\beta}_1 = 2$.

12. Carry out a grid search to estimate the coefficients (to the nearest hundredth) when we minimize the sum of the absolute errors. Report the parameter estimates. **(2pts.)**

13. Compare the estimated effect of age from this estimation to that from the OLS estimate. How different are the interpretations for the effect of age on wage when we use a different criterion for measuring the model fit (or misfit)?