**MGS 8040 Assignment 1: Regression**

Do the analyses in Excel, but **answer all questions in a Word document** **with the appropriate tables and charts copied from Excel.** Turn in both the Excel and the Word files online.

A researcher is trying to determine if exercise affects different people differently. Specifically, when people exercise, there is an increase in a protein called BDNF (Brain derived neurotrophic factor). Is this **Increase in BDNF** different among males and females, and is it different among different ethnicities?

The dataset provided shows data on **Increase in BDNF** in the body due to **Exercise**, for people in the age range of 18-25, along with their **Gender** and **Ethnicity**.

1. Create a Histogram of the variable **Exercise** to look at its distribution. Compute the Mean and Standard Deviation of the variable as well.
2. Create a Pivot table in Excel to analyze the **Increase in BDNF** by **Ethnicity** and **Gender** simultaneously. The table should look as follows, with the Average, Standard Deviation, and the Count values for **Increase in BDNF** in each of the cells:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Column Labels** |  |  |
| **Row Labels** | Female | Male | **Grand Total** |
| African |  |  |  |
| Asian |  |  |  |
| Caucasian |  |  |  |
| **Grand Total** |  |  |  |

Interpret the table – what can you say about the impact of **Gender** and **Ethnicity** overall? Does there seem to be an interaction between **Gender** and **Ethnicity**? In other words, is the impact of **Gender** on **Increase in** **BDNF** different for different Ethnicities?

1. Create a scatterplot of **Increase in BDNF** and **Exercise**. Make sure Increase in BDNF is on the Y axis. Interpret.
2. Perform a multiple regression analysis using the dataset provided (use all the Xs available for the regression), to predict Increase in BDNF in the body. For Gender, code Male as 1 and Female as 0 (baseline). For Ethnicity, create two dummies, one for Asian and one for African (keep Caucasian as the baseline).

**Note that in Excel, all the Xs that you will use must be in contiguous columns.**

1. Is the regression significant overall? What does that mean?
2. Is each of the variables in the model significant at the 5% level? Interpret.
3. What is the R-squared value? What does it mean?

**Now run the regression again after removing variables that are not significant at the 5% level.**

**(remember that you may have to move data in the worksheet so that the variables left in the regression are all in contiguous columns).**

1. What is the final equation to predict BDNF increase? What would your model predict as the increase in BDNF for a person who exercises 20 minutes, and is a Caucasian Female?
2. Interpret the coefficients for Asian and African.
3. If you want to improve the R-square value of the model, is increasing the sample size the best way? Why or why not?