Assignment 2 STA4164, Dr. Alexander V. Mantzaris

Oct 21, 2021, Due: Nov, 05 2021

Put all working out and code into a single pdf file or as a single notebook.

Please use either SAS, Python, Julia or R. Work is to be done individually.

1 Question 1 (20 points)

Data from: `https://www.kaggle.com/sudalairajkumar/undata-country-profiles`

1. Load the dataset 'country_profile_variables.csv' which is UN data on countries. (2 points)

2. Subset the data for just the regions 'EasternAsia' and 'WesternAsia'. Then perform a t-test for the 'Population in thousands (2017)' and report the results. (4 points)

3. Subset the data for just the regions 'EasternAsia', 'WesternAsia', 'SouthAmerica', and 'MiddleAfrica'. Using the 'Population in thousands (2017)' column as the independent variable, perform a 1 way fixed effects ANOVA to determine the independence of the Population variable ($Population = \mu + Region_i$) (5 points)

4. Perform regression between the variables 'Population in thousands (2017)', 'Population density' as the independents and 'GDP: Gross domestic product (million current US$)' and comment on the results. (3 points)

5. Perform regression between the variables where one is a factor 'Region', 'Population in thousands (2017)', 'Population density' as the independents and 'GDP: Gross domestic product (million current US$)' and comment on the results. (4 points)

6. R code to produce a qq plot is 'qqnorm(datavector, pch = 1, frame = FALSE)' and can help determine is the data is normally distributed or not. Use this function on the data of the columns 'Population density' and 'GDP growth' and comment on whethe the data is normally distributed or not. (2 points)