<u>Tutorial 4 STQD6214</u>

Answer the following questions using R. Note that although several assumptions are required for different hypothesis tests or linear regression, we are not too concerned about that in this tutorial and will just apply these methods to the datasets. Additionally, you may use $\alpha = 0.05$ for making decisions in hypothesis tests.

1. Import the file “TestScore.csv” into R. The data contains the score, IQ, weekly study hours and GPAs of students from two classes.

   a) The following lines can be used to separate the data for the two classes.

   ```
   class_A <- subset(TestScore, Class=="A")
   class_B <- subset(TestScore, Class=="B")
   ```

   Run the lines above and test whether the mean score for class A students is greater than 70. State the null and alternative hypotheses.

   b) Test whether the variance of the scores of students in the two classes are equal. State the null and alternative hypothesis.

   c) Test whether the mean GPA of students in Class A is equal to the mean GPA of students in Class B, assuming the variances are equal. State the null and alternative hypothesis.

   d) Using one-way ANOVA, test whether the whether the mean GPA of students in Class A is equal to the mean GPA of students in Class B. Compare the *p*-value of your test here with the *p*-value found in (c).

   e) From the scatterplot, it is hypothesized that IQ can be used as a predictor for test score using the linear regression model. Fit the linear regression model using test score as the dependent variable and IQ as the independent variable.

      i) Write down the fitted regression line.

      ii) Do the two variables test score and IQ have a significant linear relationship? Explain your answer.

2. A researcher measured 13 batches of cement and would like to predict heat evolved in calories per gram of cement given the percentage of ingredients:

   - Response y: heat evolved in calories during hardening of cement on a per gram basis
   - Predictor x_1: % of tricalcium aluminate
   - Predictor x_2: % of tricalcium silicate
   - Predictor x_3: % of tetracalcium alumino ferrite
   - Predictor x_4: % of dicalcium silicate

   The file “cement.csv” contains this dataset.

a) Using x_1, x_2, x_3 and x_4 as the predictors, fit a linear regression model with y as the response variable. Based on your results here, which predictors are significant to the model? Use 10% significance level for your decision.

b) Now, using only x_1 and x_2 as the predictors, fit a linear regression model with y as the response variable. Based on your results here, which predictors are significant to the model?

c) Write down the fitted regression model using the result in (b).

d) Using the model in (b) what is the estimated heat evolved in calories per gram of cement when using 8% of tricalcium aluminate and 50% of tricalcium silicate?

3. Import the file "property_sales.csv" into R, which we used in previous tutorial. The dataset shows the sales for residential properties in four neighbourhoods together with their land and improvement values.

a) The following lines give the sales value for properties in Cheval and Hyde Park.

```
cheval_sale <- subset(property_sales,
                      Neighbourhood=="Cheval")$Sales
hydepark_sale <- subset(property_sales,
                        Neighbourhood=="HydePark")$Sales
```

Run the lines above and test whether the sales for properties in Cheval and Hyde Park have the same variance. State the null and alternative hypotheses.

b) Test whether the mean sales for properties in Cheval is smaller than the mean sales for properties in Hyde Park, assuming that the variances for the two populations are not equal.

c) Using one-way ANOVA, test whether the mean land values are equal for all four neighbourhoods. State the null and alternative hypotheses.

d) Using one-way ANOVA, test whether the mean improvement values are equal for all four neighbourhoods.

e) Using land and improvement values as the regressors or independent variables, fit a linear regression model to predict the sales using the two regressors.

i) Write down the fitted regression model.

ii) Based on the results, is land value significant in the linear model? Explain your answer.

iii) What is the estimated sales price of a property with $100 000 land value (or equivalently `Land.value=100`) and $200 000 improvement value (or equivalently `Improvement.value=200`)?