**MATH42715: Introduction to Statistics for Data Science** **Assignment 1**
**Dr. Tahani Coolen-Maturi** **Michaelmas 2022**

Under the "Assignments" section on Ultra, please click on **Online Assignment 1**. Please follow the tasks described on the task sheet below, and answer questions **Q1** to **Q13**.

---

(1) **Hypothesis testing**

Consider the built-in data set in R called the `ToothGrowth`, which contains data from a study evaluating the effect of vitamin C on tooth growth in Guinea pigs. The experiment was performed on 60 pigs, where each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods (supplement types), orange juice '`OJ`' or ascorbic acid '`VC`'.

First, load and preview the data:

```
# load the data ToothGrowth
data(ToothGrowth)
?ToothGrowth
# preview the structure of the data
str(ToothGrowth)
```

You should get the following output:

```
> # preview the structure of the data
> str(ToothGrowth)
'data.frame': 60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

From now on consider only two variables: the Tooth length '`len`', and the Supplement type '`supp`' ( '`OJ`' or '`VC`').

(a) Which chart type is most appropriate to compare tooth length '`len`' by supplement type '`supp`'? (Q1)

(b) Considering tooth length '`len`' by supplement type '`supp`', what type of experiment design do we have? (Q2)

(c) Provide the means and standard deviations of the tooth length by supplement type (Q3).

(d) Use hypothesis testing to compare tooth length by supplement type. State the null and alternative hypothesis, and provide the value of the test statistic and the p-value for this problem, using the significance level 1%. State your conclusions and the assumptions needed for your conclusions. (Q4).

(2) **Regression analysis**

From the `ggplot2` package in R, we consider the `mpg` dataset on fuel efficiency for vehicles from 1999- 2008. A researcher wants to study whether there is a relationship between Engine Displacement in litres '`displ`' and City Highway Miles Per Gallon '`hwy`'. So here, the predictor variable is '`displ`' and the response variable is '`hwy`'.

The package can be installed by typing into the console:

```
install.packages("ggplot2")
```

Now, load and preview the data:

```
library(ggplot2)
data(mpg)
?mpg
str(mpg)
head(mpg)
```

(a) Consider the following linear model

$$\texttt{hwy} = \beta_0 + \beta_1\texttt{displ} + \epsilon \tag{1}$$

Fit the linear model (1) and save it into object `fit1`. Read from the summary output the standard error of $\hat{\beta}_1$ (Q5). Suppose now that `displ = 4.5`, find the predicted value of `hwy` (Q6).

(b) Provide the F-statistic value (Q7) and the value of the adjusted R-squared (Q8). Interpret the value of the adjusted R-Squared in the context of the problem (Q9).

(c) Create a 2 x 2 grid containing the following four residual diagnostic plots for `fit1`: (i) Residual vs Fitted, (ii) Normal QQ, (iii) Scale-Location, and (iv) Residuals vs Leverage. Using these plots, comment on whether the linear regression assumptions hold for the model `fit1` (Q10).

(d) It has also been suggested to use the model

$$\texttt{hwy} = \beta_0 + \beta_1\texttt{displ} + \beta_2(\texttt{displ})^2 + \epsilon \tag{2}$$

Fit the linear model (2), and save it into object `fit2`. For the same specific covariate value as given in part (a), find the predicted value of `hwy` (Q11).

(e) Provide the value of the adjusted R-squared for the model `fit2`, explain how it compared to the adjusted R-squared value obtained in part (b), and comment on your answer (Q12).

(f) Give the value of the residual standard error $s$ for both models. Explain the difference in the degrees of freedom and residual standard error for both models. Do these results suggest that one of the models should be preferred over the other one? Explain your answer (Q13).