ISDS Online Assignment 1

March 2023

Total Marks: 93

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**).

Question 1 Which chart type is most appropriate to compare tooth length '**len**' by supplement type '**supp**'? Show answer choices (**B**)

A: Bar chart

B: Boxplot

C: Line chart

D: Pie chart

Question 2 Considering tooth length 'len' by supplement type 'supp', what type of experiment design do we have? Show answer choices (**A**)

A: Two independent samples

B: Paired samples

Question 3 To 3 decimal places, provide the means and standard deviations of the tooth length by supplement type.

tooth length by	**Mean**	**Standard deviation**
supplement type = OJ	**20.663**	**6.606**
supplement type = VC	**16.963**	**8.266**

Table 1: Caption

Question 4 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.

Your Answer

Marks:13/20

Firstly, we need to test the homogeneity of variance between these two samples (OJ or VC). The null hypothesis is: The variance between these two samples (OJ or VC) is equal. The alternative hypothesis is: The variance between these two samples (OJ or VC) is not equal. For homogeneity of variance test, leveneTest will be a good method. According to leveneTest results (centre = median), F = 1.2136, p-value = 0.2752 ¿ 0.05, the null hypothesis can't rejected, which means the variance is equal.

As the variance between two samples is equal, t-test with parameter var.equal=TRUE will be correct. The null hypothesis: tooth length by supplement type in means is equal The alternative hypothesis: tooth length by supplement type in means is not equal The result of t test shows t = 1.9153, df = 58, p-value = 0.06039 ¿ 0.05, with 99 percent confidence interval from -1.445056 to 8.845056.

Conclusion: we can reject the null hypothesis, which means we cannot draw a conclusion that different supplies(OJ and VC) have different effect on tooth-growth

2 Regression analysis

Question 5

Considering the model fit1, read from the summary output the standard error of $\hat{\beta}_1$

Your Answer: 0.1945

Question 6

Considering the model **fit1**, suppose that **displ** = 4.5, then the predicted value of **hwy** is

Your Answer: 19.81

Question 7 Considering the model **fit1**, the F-statistic value is

Your Answer: 329.5

Question 8

Considering the model **fit1**, the value of the adjusted R-squared is Your Answer: 0.585

Question 9

The value of the adjusted R-Squared tells us: Show answer choices (**C**)

A: that there is a positive relationship between *hwy* and *displ*.

B: that there is a negative relationship between *hwy* and *displ*.

C: the proportion of the variation in the response variable, *hwy*, explained by the predictor variable *displ*.

D: that there is no relationship between *hwy* and *displ*.

Question 10 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**.

Your Answer

(i)Residuals vs Fitted: The Residuals vs Fitted plot graph shows whether there are nonlinear patterns in the residual and the data or not. The line in this graph should be a relatively flat line if the data pattern fits a linear model. According to the graph above, the line is not flat, but bow-shaped, which means there will be some non-linear patterns that we fail to capture. This is the reason for the model fit2 in Question11.

(ii)Normal QQ: The normal QQ plot shows whether the residuals are distribute normally or not. If the residuals distribution is normal, the residuals will distribute on the line or close to the line. As we see the Q-Q plot above, there are a lot of data distributing far from the line in the upper right corner and the lower left corner. So, this graph indicate the residuals are not normally distributed and we should try other models.

(iii) Scale-Location: The Scale-Location plot show whether the residuals are distributed equally along the range of the predictors. An ideal Scale-Location plot will be a horizontal line with randomly point. According to the Scale-Location plot above, the line looks like a ^ with higher values on the upper left corner and upper right corner, which indicates there are non-linear patterns not captured.

(iv) Residuals vs Leverage: The Residuals vs Leverage plot shows whether there are some influential cases. Influential cases is not equal to extreme values. Some influential values are not extreme but have a bad damage on the model. The influential cases means the values outside of the Cook's distance. According to the Residuals vs Leverage plot above, there are no influential cases or influential case.

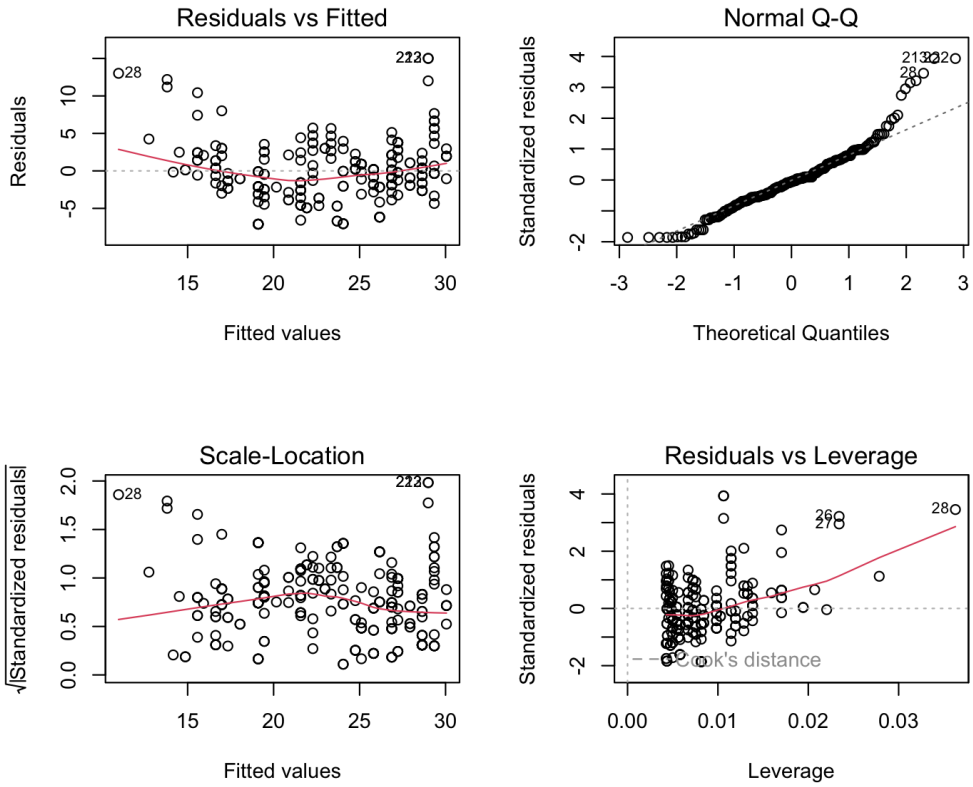

Figure 1: Q10 plot