**UTS**

**Understanding Data and Statistical Design (60117)**

**Assessment Task 2: Data Analysis Assignment**

**Autumn 2024**

This assessment task is marked from 60.

It is worth 40% of the marks for this subject.

Please submit via Canvas.

**Due by 11.59pm Sunday 5 May 2024.**

**You must use R to answer these questions.**

**REQUIREMENTS**

* **Include all R output that you refer to in answering the questions (use the Snipping Tool programme or similar to copy output produced by R).**
* **Make sure you define any symbols (other than those already defined) you use in answering these questions.**
* **Submit a single PDF file of your answers to Canvas.**

**Marks will be deducted for not adhering to these requirements.**

**You must also submit a signed copy of the coversheet with your answers.**

**Q1 & Q2 DATA**

The data for Q1 and Q2 is contained in the npk dataset built into R. The variables we will consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
|  | blocking factor | plot groups 1-6 |
|  | experimental factor | addition of nitrogen: “0” (no), “1” (yes) |
|  | response | yield of peas (lbs/plot) |

The data records the results of a study on the effect of nitrogen (and other elements) on

the yield of peas grown in an agricultural setting. Because the peas were grown in six different plot groups, the blocking factor has also been included. Within each plot group a total of four plots (two with nitrogen, two without) were used for a total of twenty-four plots.

The code below will create the data frame npk with the sample data necessary for Q1 and Q2.

data(npk)

**QUESTION 1. Single factor experiment [15 marks]**

1. Construct a histogram of the full sample of and on the same plot superimpose a normal density curve fitted to this sample. Use this to determine if the sample appears to be normally distributed **[3 marks]**.

A graph of a graph with a red line

Description automatically generated

Analysing the density plot together with its fitted line, it appears that the data have a normal distribution, but a hypothesis test is needed to check this.

1. Using significance level , perform a normality test on the full sample of . Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.

A black background with white text

Description automatically generated

**Ho:** The data is normally distributed.

**Ha:** The data is not normally distributed.

**Statistic:** 0.97883

**P\_value:** 0.8735

**Decision:** Since the p\_value > α, we don’t reject the null hypothesis.

**Conclusion:** We don’t have evidence to reject the null hypothesis. The yield data come from a normally distributed population.

1. Without using the t.test (or similar) function in R, construct a 95% one-sided confidence interval for population mean of that could be used to perform an upper-tail -test **[3 marks]**.

A black screen with text

Description automatically generated

1. Using significance level , perform a test to determine if the median yield of peas grown with nitrogen is more than 55 lbs/plot. Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.

A black background with white text

Description automatically generated

**Ho:** The median yield of peas grown with nitrogen is 55 lbs/plot or less.

**Ha:** The median yield of peas grown with nitrogen is greater than 55 lbs/plot.

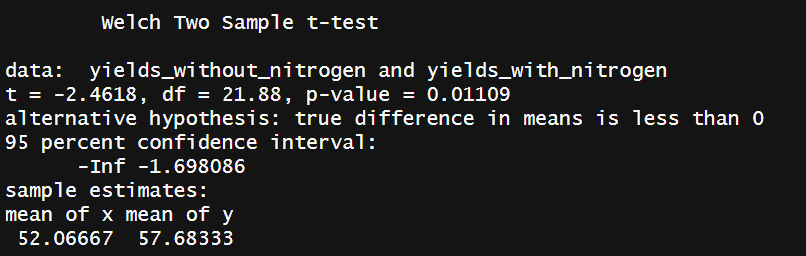
**Statistic:** V = 57

**P\_value:** 0.08813

**Decision:** Since the p\_value > α, we do not reject the null hypothesis.

**Conclusion:** There is insufficient evidence to conclude that the median yield of peas grown with nitrogen is greater than 55 lbs/plot.

1. Using significance level , perform a test to determine if the mean yield of peas grown without nitrogen is less than the mean yield of peas grown with nitrogen. Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.



**Ho:** The mean yield of peas grown without nitrogen is equal to or greater than the mean yield of peas grown with nitrogen.

**Ha:** The mean yield of peas grown without nitrogen is less than the mean yield of peas grown with nitrogen.

**Statistic:** t = -2.4618

**P\_value:** 0.01109

**Decision:** We reject the null hypothesis since the p\_value < α (0.05).

**Conclusion:** Evidence suggests that the mean yield of peas grown without nitrogen is less than that of peas grown with nitrogen.

**QUESTION 2. Two factor experiment [15 marks]**

1. Construct a single chart that displays twelve boxplots, one for each combination of factor levels **[2 marks]**. Of the three assumptions the -tests from two-way ANOVA rely on, which is most called into question by these plots **[1 mark]**?

A chart with green and blue rectangles

Description automatically generated

Homogeneity of variances: The variance among the groups should be approximately equal.

* The lengths of the boxes, IQR, vary among the different groups.
* Some groups show a wide spread of data, while others have a much tighter spread.

1. Write down the statistical model for a completely randomised block design consistent with the sample data, excluding interaction between the factors **[2 marks]**. Identify the treatments **[1 marks]**.

Y: The yield

µ: The global mean level

α: The effect of the i-th level of Nitrogen

ρ: The effect of the j-th level of phosphorus

ω: The effect of the k-th of potassium

β: The effect of the l-th block

ε: The random error associated to each observation

**Treatments:**

* **Nitrogen:** 0=not applied, 1= applied
* **Phosphorus:** 0=not applied, 1= applied
* **Potassium:** 0=not applied, 1= applied

Total treatments: 23=8: (N0, P0, K0), (N0, P0, K1), (N0, P1, K0), (N0, P1, K1), (N1, P0, K0), (N1, P0, K1), (N1, P1, K0) and (N1, P1, K1).

1. Using significance level , perform two-way ANOVA (without interaction) and document the -test for the blocking factor only (marks will be deducted for documenting additional -tests other than this). Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.

A black screen with white text

Description automatically generated

**Ho:** The blocking factor has no significant effect on the mean yield of peas.

**Ha:** The blocking factor has a significant effect on the mean yield of peas.

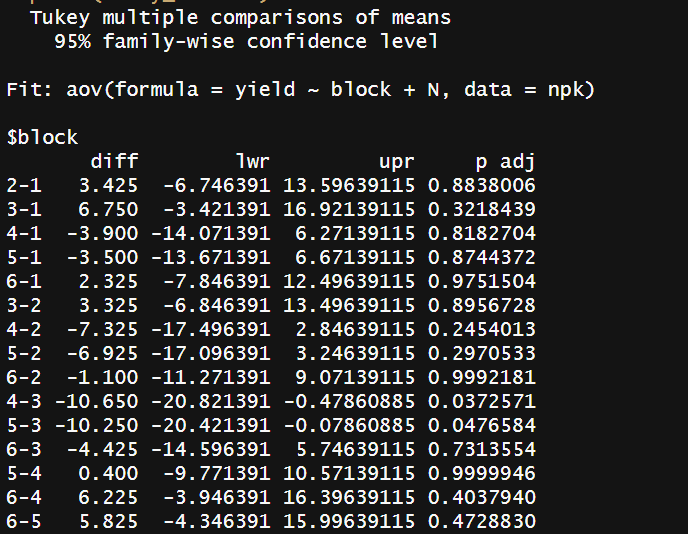
**Statistic:** t = 3.395

**P\_value:** 0.0262

**Decision:** We reject the null hypothesis since the p\_value < α (0.05).

**Conclusion:** Evidence suggests that the blocking factor has an effect on the mean yield of peas

1. Using significance level , perform Tukey post-hoc analysis to determine which pairs of plot groups are associated with different mean yield of peas **[3 marks]**.



There are statistically significant differences in mean yields between block 4 and block 3, and between block 5 and block 3.

1. Using diagnostic plots of the residuals, assess whether the assumptions underpinning the -test conducted above have been met **[3 marks]**.

* The assumption of equal variances (homoscedasticity) is questionable, as the Scale-Location plot indicates, where the variance increases with the fitted values.
* The assumption of normality of residuals needs more analysis because slight deviations observed in the tails of the Normal Q-Q plot could be a concern.

A graph of different values

Description automatically generated with medium confidence

**Q3 & Q4 DATA**

The data for Q3 and Q4 is contained in the mtcars dataset built into R. The variables we will consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
|  | continuous response | standing quarter mile time (s) |
|  | continuous predictor | vehicle fuel economy (mpg) |
|  | continuous predictor | engine horsepower (hp) |
|  | categorical predictor | engine configuration: “0” (V or boxer), “1” (other) |

The data records attributes of 32 motor vehicles.

The code below will create the data frame mtcars with the sample data necessary for Q3 and Q4.

data(mtcars)

mtcars$vs <- factor(x=mtcars$vs, levels=c("0","1"))

**QUESTION 3. Simple linear regression [12 marks]**

In this question we build a simple linear regression to model the relationship between standing quarter mile time ( and vehicle fuel economy . We consider the population model

where .

1. Fit the model described above, write down the regression equation **[1 mark]** and use the model to calculate the vehicle fuel economy associated with a predicted average quarter mile time of 17.5s **[2 marks]**.

A black screen with white text

Description automatically generated

Predicted vehicle fuel economy (mpg) for an average quarter mile time of 17.5s: 17.28122

1. Using significance level , test whether average quarter mile time changes by an amount different to 0.05s for each additional mile per gallon in vehicle fuel economy. Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.

A screenshot of a computer program

Description automatically generated

**Ho:** There is no change in quarter-mile time other than by 0.05 seconds for each additional mile per gallon.

**Ha:** There is a change in quarter-mile time other than by 0.05 seconds for each additional mile per gallon.

**Statistic:** t = 2.525

**P\_value =** 0.0171

**Decision:** We reject the null hypothesis since the p\_value < α (0.05).

**Conclusion:** Evidence suggests that the change in average quarter mile time per additional mile per gallon in vehicle fuel economy is significantly different from 0.05 seconds.

1. Is there any statistical evidence against the assumption of independent errors **[3 marks]**?

A black screen with white text

Description automatically generated

**Ho:** There is no autocorrelation in the residuals.

**Ha:** There is autocorrelation in the residuals.

**Statistic:** t = 1.2553

**P\_value:** 0.01122

**Decision:** We reject the null hypothesis since the p\_value < α (0.05).

**Conclusion:** There is statistical evidence to suggest that there is autocorrelation in the residuals of the model, indicating a violation of the assumption of independent errors.

1. Provide an estimate of **[3 marks]**.

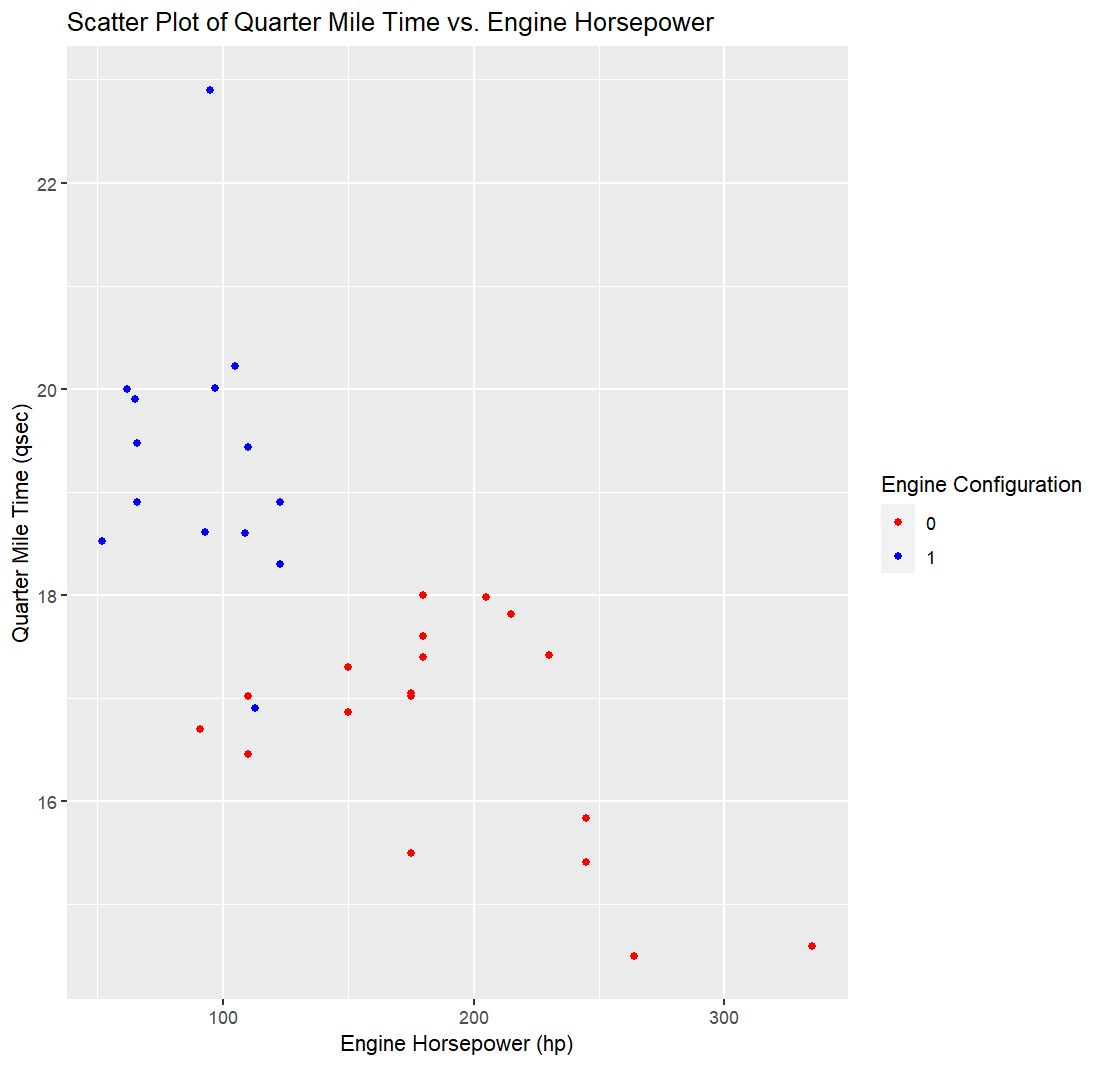
A black background with white text

Description automatically generated

**QUESTION 4. Multiple linear regression [18 marks]**

In this question we extend the model from Q3 into a multiple linear regression.

1. Construct a scatter plot of standing quarter mile time (vertical axis) against engine horsepower (horizontal axis), colouring the data points blue for V or boxer engine configuration and red otherwise **[3 marks]**.

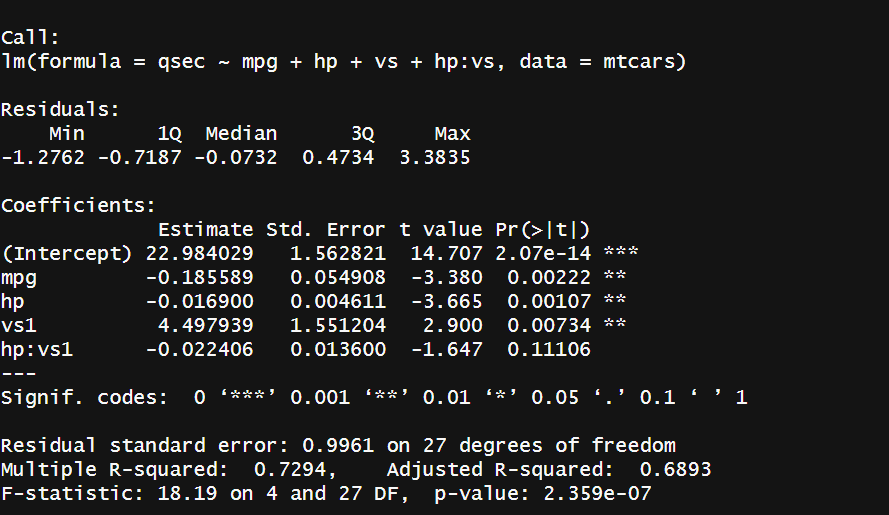


We now consider the population model

where

Note that R will create the dummy variable automatically ( won’t be used directly as it is a factor).

1. Fit the model described above, write down the regression equation that applies for vehicles with engine configuration that is not a V or boxer **[1 mark]** and provide interpretations of the estimated coefficients and **[2 marks]**.



Equation:

qsec=(22.984029+4.497939)+(−0.185589)⋅mpg+(−0.016900−0.022406)⋅hp

qsec=27.481968 −0.185589⋅mpg−0.039306⋅hp

qsec=27.481968−0.185589⋅mpg−0.039306⋅hp

Interpretation  
: Represents the estimated quarter mile time for a hypothetical vehicle with 0 mpg and 0 hp; this is the baseline o the model.

: Indicates that the effect of horsepower on quarter mile time for non-V/boxer engines is different compared to V/boxer engines by this value seconds per hp

1. Using significance level , determine if the regression is significant. Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.

A screen shot of a computer code

Description automatically generated

**Ho:** None of the predictor variables (mpg, hp, vs, and hp:vs) significantly predict the response variable.

**Ha:** The predictor variables (mpg, hp, vs, and hp:vs) are significant in predicting the response variable.

**Statistic:** t = 18.19

**P\_value=** 2.359e-07

**Decision:** We reject the null hypothesis since the p\_value < α (0.05).

**Conclusion:** Strong evidence suggests that at least one of the predictors (mpg, hp, vs, and hp:vs) provides significant information in predicting the response variable.

1. Compute the 95% individual prediction interval for quarter mile time of a vehicle with a V or boxer engine, 26.5mpg fuel economy and 300hp engine **[3 marks].**

A screenshot of a computer

Description automatically generated

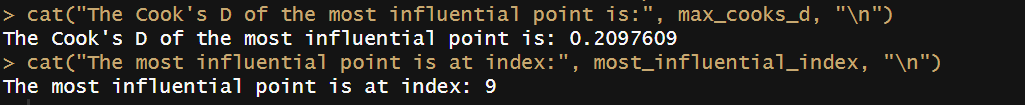
1. Without using vif (or similar) function in R and using only a regression model involving , , and , calculate the variance inflation figure (VIF) for the predictor in the regression model fitted in part (b) **[3 marks]**.

A screen shot of a computer code

Description automatically generated

1. Provide the Cook’s D of the most influential point **[1 mark]**, refit the regression model on sample data excluding this point and write down the fitted equation **[2 marks]**.

Most influential point:



Model excluding that point:

A screenshot of a computer program

Description automatically generated