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
block	blocking factor	plot groups 1-6
N	experimental factor	addition of nitrogen: "0" (no), "1" (yes)
yield	response	yield of peas (lbs/plot)

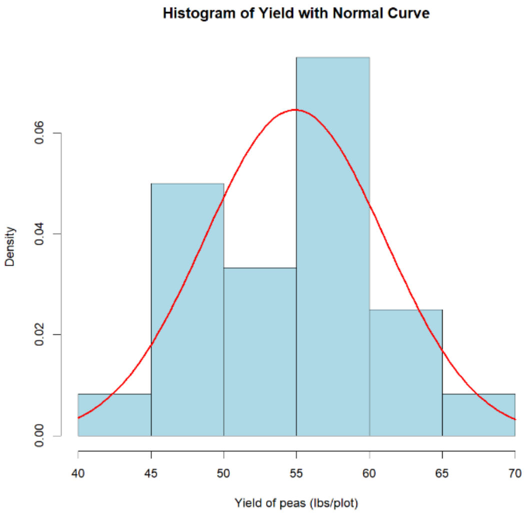
The data records the results of a study on the effect of nitrogen (and other elements) on the yield of peas grown in an agricultural setting. Because the peas were grown in six different plot groups, the blocking factor *block* has also been included. Within each plot group a total of four plots (two 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)
```

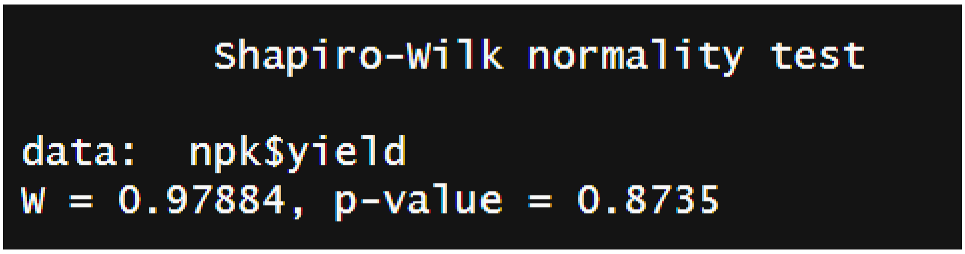
QUESTION 1. Single factor experiment [15 marks]

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



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.

(b) Using significance level $\alpha = 0.05$, perform a normality test on the full sample of *yield*. 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]**.



```
        Shapiro-Wilk normality test

data:  npk$yield
W = 0.97884, p-value = 0.8735
```

Ho: The data is normally distributed.

Ha: The data is not normally distributed.

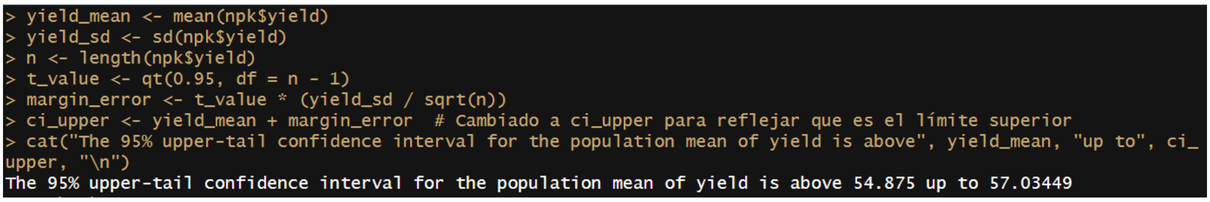
Statistic: 0.97883

P_value: 0.8735

Decision: Since the p_value > α, we don't reject the null hypothesis.

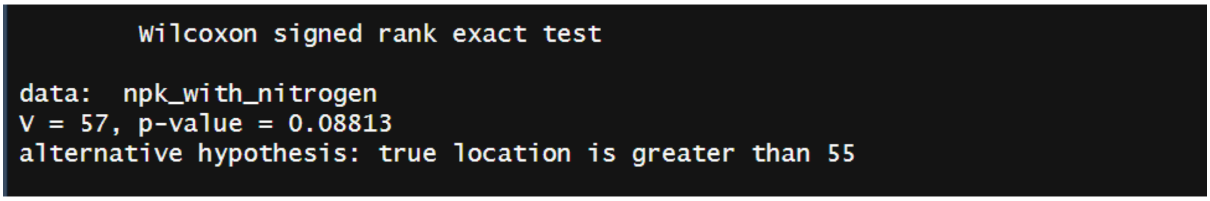
Conclusion: We don't have evidence to reject the null hypothesis. The yield data come from a normally distributed population.

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



```
> yield_mean <- mean(npk$yield)
> yield_sd <- sd(npk$yield)
> n <- length(npk$yield)
> t_value <- qt(0.95, df = n - 1)
> margin_error <- t_value * (yield_sd / sqrt(n))
> ci_upper <- yield_mean + margin_error  # Cambiado a ci_upper para reflejar que es el límite superior
> cat("The 95% upper-tail confidence interval for the population mean of yield is above", yield_mean, "up to", ci_
upper, "\n")
The 95% upper-tail confidence interval for the population mean of yield is above 54.875 up to 57.03449
```

(d) Using significance level $\alpha = 0.05$, 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]**.



```
        Wilcoxon signed rank exact test

data:  npk_with_nitrogen
V = 57, p-value = 0.08813
alternative hypothesis: true location is greater than 55
```

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.

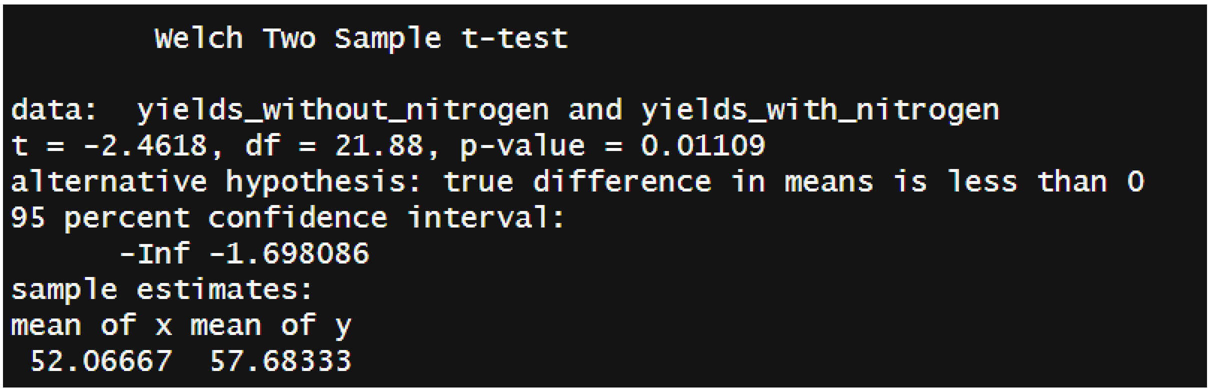
Statistic: V = 57

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.

(e) Using significance level $\alpha = 0.05$, 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]**.



```
        Welch Two Sample t-test

data:  yields_without_nitrogen and yields_with_nitrogen
t = -2.4618, df = 21.88, p-value = 0.01109
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.698086
sample estimates:
mean of x mean of y
 52.06667  57.68333
```

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

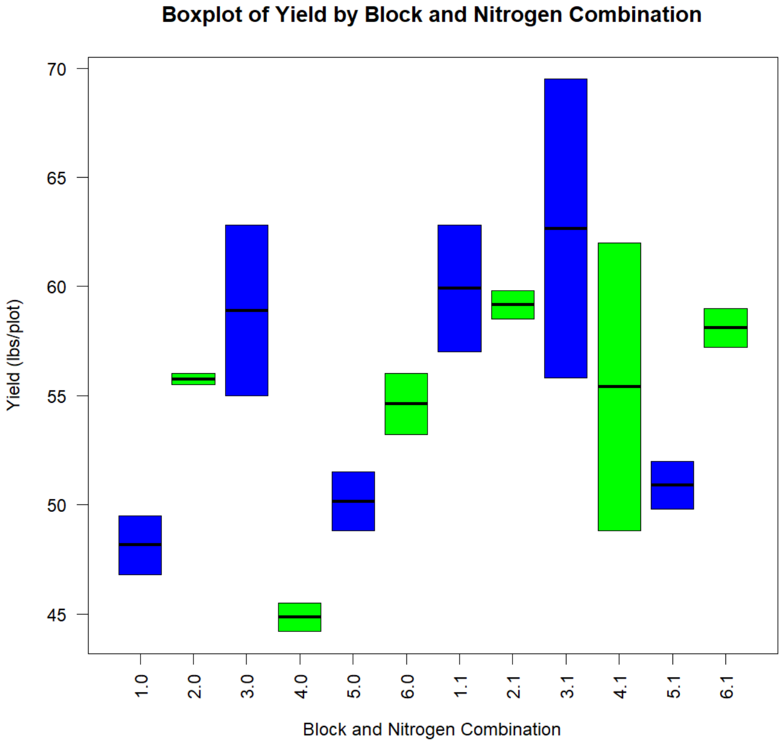
P_value: 0.01109

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]

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



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.

(b) Write down the statistical model for a 2 × 6 completely randomised block design consistent with the sample data, excluding interaction between the factors **[2 marks]**. Identify the treatments **[1 marks]**.

$$Y_{ijkl} = \mu + \alpha_i + \rho_j + \omega_k + \beta_l + \varepsilon_{ijkl}$$

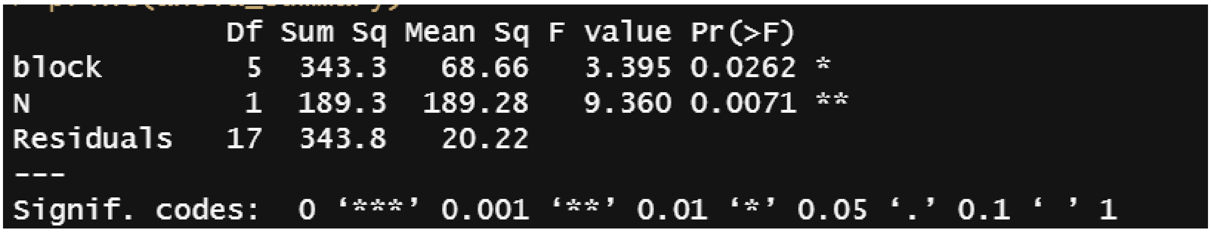
Y: The yield
μ: The global mean level
α: The effect of the i-th level of Nitrogen
ρ: The effect of the j-th level of phosphorus
ω: The effect of the k-th of potassium
β: The effect of the l-th block
ε: The random error associated to each observation

Treatments:

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

Total treatments: 2^3=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).

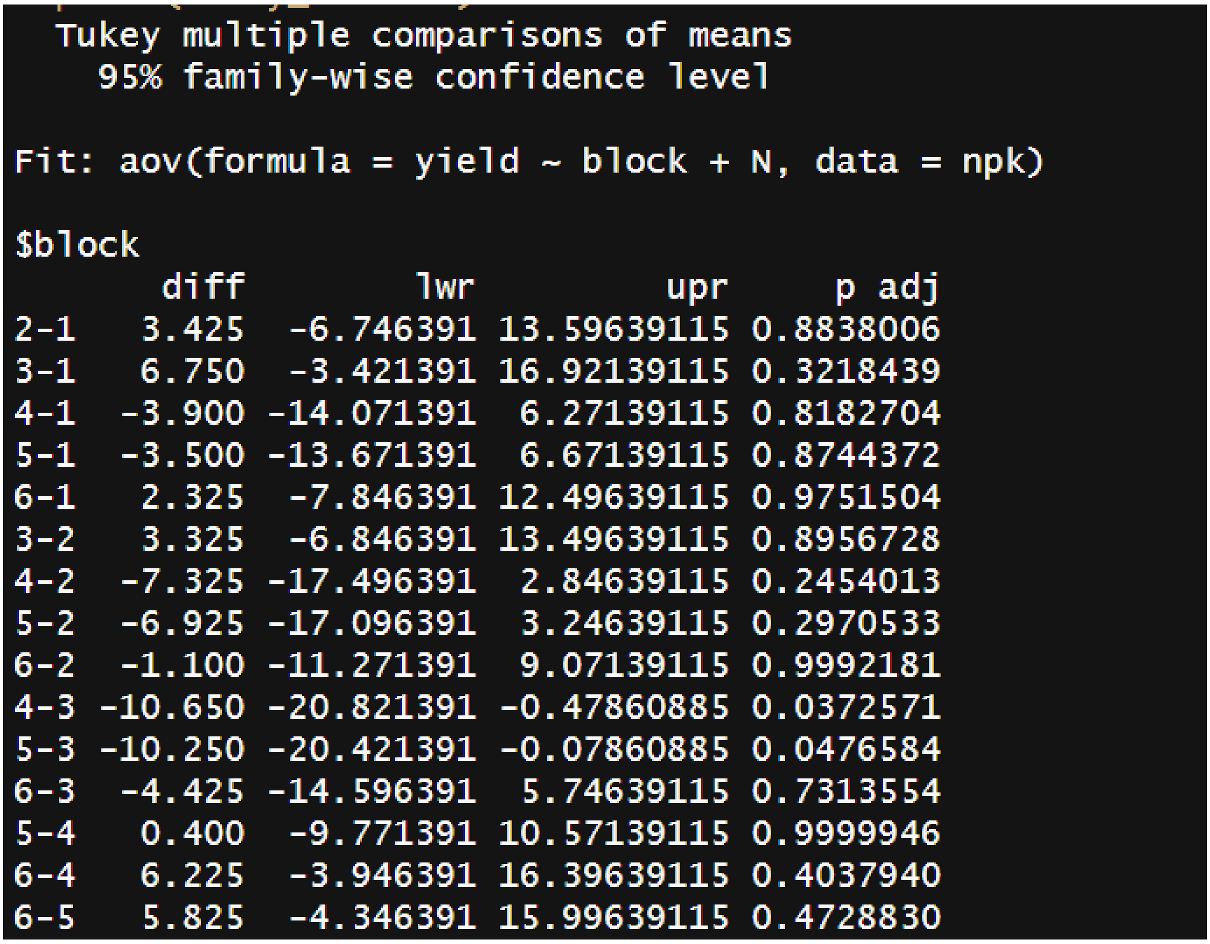
(c) Using significance level $\alpha = 0.05$, perform two-way ANOVA (without interaction) and document the F-test for the blocking factor only (marks will be deducted for documenting additional F-tests other than this). Write down the null and alternative hypotheses, the test statistic and p-value, the test decision with reason and a conclusion using a minimum of mathematical language **[3 marks]**.



```
             Df Sum Sq Mean Sq F value Pr(>F)  
block         5  343.3   68.66   3.395 0.0262 *
N             1  189.3  189.28   9.360 0.0071 **
Residuals    17  343.8   20.22                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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

(d) Using significance level $\alpha = 0.05$, perform Tukey post-hoc analysis to determine which pairs of plot groups are associated with different mean yield of peas **[3 marks]**.

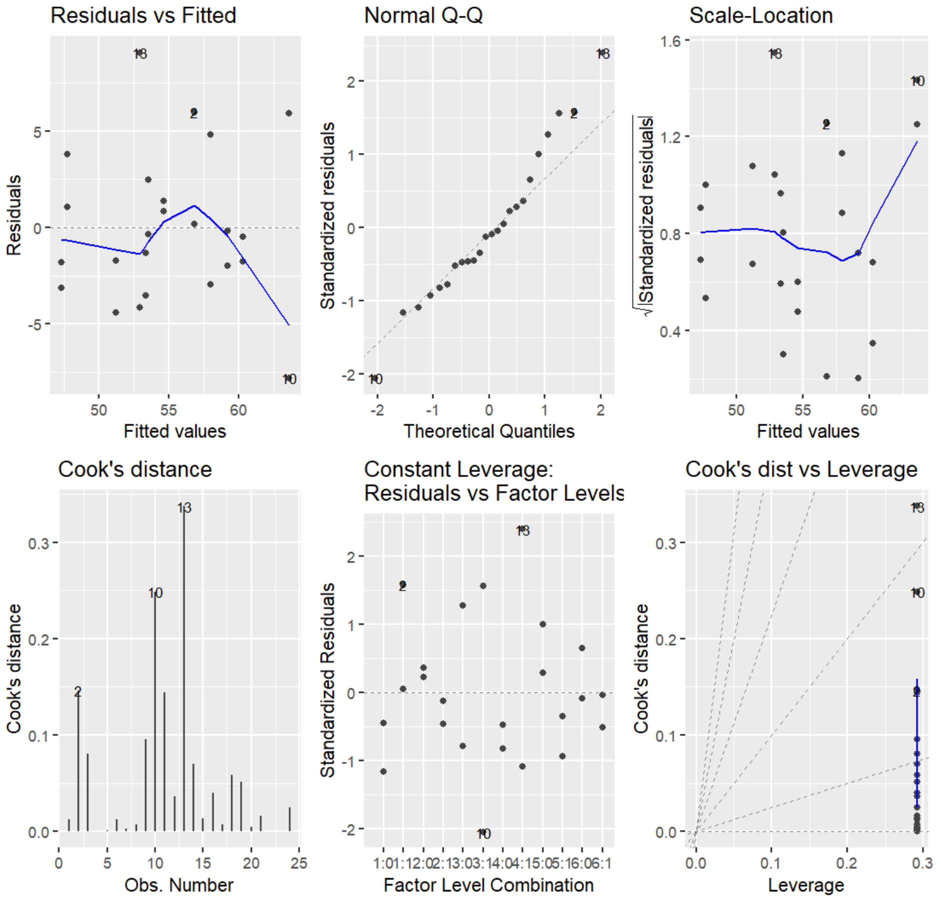


```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ block + N, data = npk)

$block
       diff        lwr         upr     p adj
2-1   3.425  -6.746391 13.59639115 0.8838006
3-1   6.750  -3.421391 16.92139115 0.3218439
4-1  -3.900 -14.071391  6.27139115 0.8182704
5-1  -3.500 -13.671391  6.67139115 0.8744372
6-1   2.325  -7.846391 12.49639115 0.9751504
3-2   3.325  -6.846391 13.49639115 0.8956728
4-2  -7.325 -17.496391  2.84639115 0.2454013
5-2  -6.925 -17.096391  3.24639115 0.2970533
6-2  -1.100 -11.271391  9.07139115 0.9992181
4-3 -10.650 -20.821391 -0.47860885 0.0372571
5-3 -10.250 -20.421391 -0.07860885 0.0476584
6-3  -4.425 -14.596391  5.74639115 0.7313554
5-4   0.400  -9.771391 10.57139115 0.9999946
6-4   6.225  -3.946391 16.39639115 0.4037940
6-5   5.825  -4.346391 15.99639115 0.4728830
```

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

(e) Using diagnostic plots of the residuals, assess whether the assumptions underpinning the F-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.


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. Number
Constant Leverage:
Residuals vs Factor Levels
Standardized Residuals
Factor Level Combination
Cook's dist vs Leverage
Cook's distance
Leverage
13
10
2

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
$qsec$	continuous response	standing quarter mile time (s)
mpg	continuous predictor	vehicle fuel economy (mpg)
hp	continuous predictor	engine horsepower (hp)
vs	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"))
```

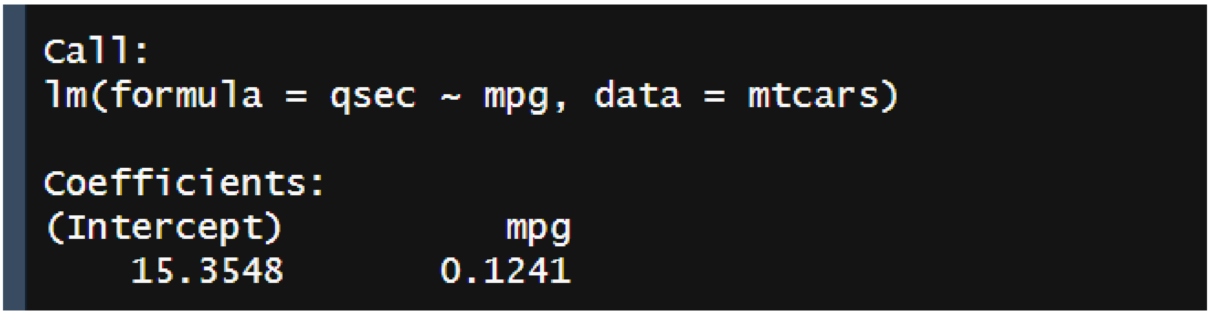
QUESTION 3. Simple linear regression [12 marks]

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

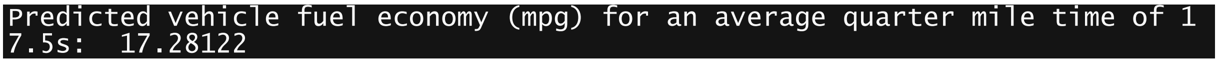
$$qsec = \beta_0 + \beta_m * mpg + \epsilon$$

where $\mathrm{var}(\epsilon) = \sigma^2$.

(a) 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]**.

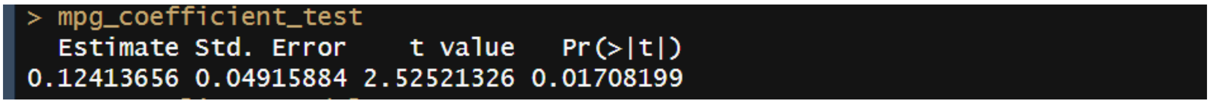


```
Call:
lm(formula = qsec ~ mpg, data = mtcars)

Coefficients:
(Intercept)          mpg
    15.3548       0.1241
```



```
Predicted vehicle fuel economy (mpg) for an average quarter mile time of 1
7.5s:  17.28122
```

(b) Using significance level $\alpha = 0.05$, 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].**



```
> mpg_coefficient_test
  Estimate Std. Error    t value   Pr(>|t|)
0.12413656 0.04915884 2.52521326 0.01708199
```

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.

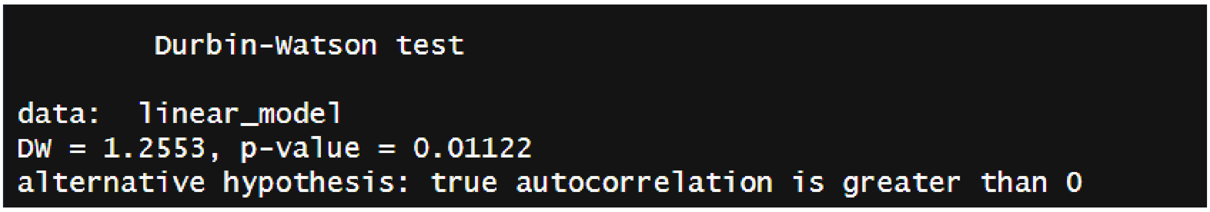
Statistic: t = 2.525

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.

(c) Is there any statistical evidence against the assumption of independent errors **[3 marks]**?



```
        Durbin-Watson test

data:  linear_model
DW = 1.2553, p-value = 0.01122
alternative hypothesis: true autocorrelation is greater than 0
```

Ho: There is no autocorrelation in the residuals.

Ha: There is autocorrelation in the residuals.

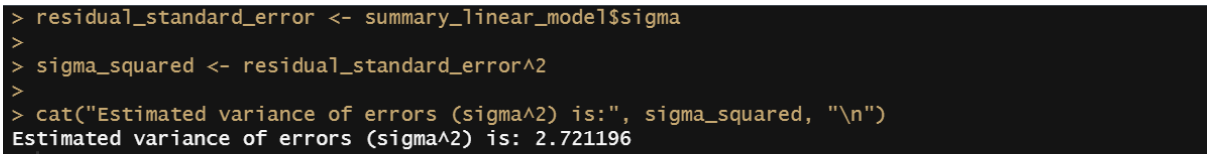
Statistic: t = 1.2553

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.

(d) Provide an estimate of σ^2 **[3 marks].**

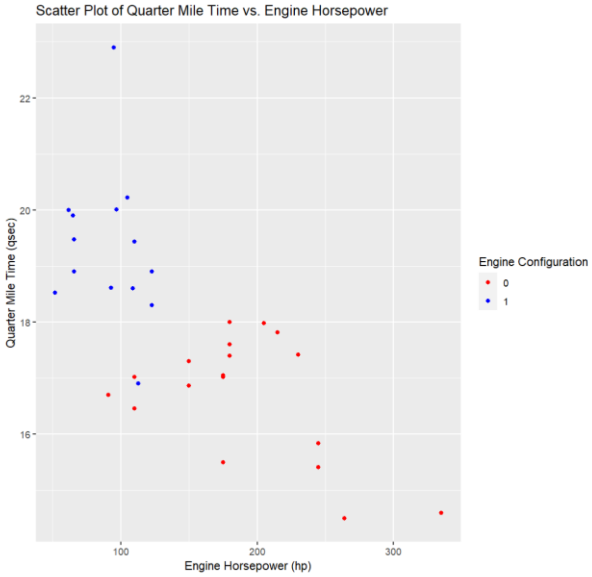


```
> residual_standard_error <- summary_linear_model$sigma
>
> sigma_squared <- residual_standard_error^2
>
> cat("Estimated variance of errors (sigma^2) is:", sigma_squared, "\n")
Estimated variance of errors (sigma^2) is: 2.721196
```

QUESTION 4. Multiple linear regression [18 marks]

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

(a) 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

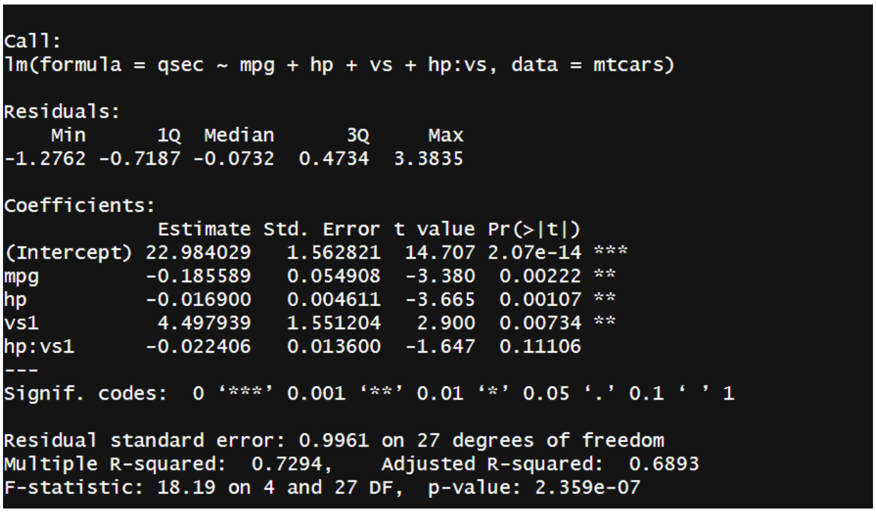
$$qsec = \beta_0 + \beta_m * mpg + \beta_h * hp + \gamma * vs1 + \delta * vs1 * hp + \epsilon$$

where

$$vs1 = \begin{cases} 0, & vs = \text{"0"} \;\; \text{(V or boxer)} \\ 1, & vs = \text{"1"} \;\text{(otherwise)} \end{cases}.$$

Note that R will create the dummy variable $vs1$ automatically (vs won't be used directly as it is a factor).

(b) 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 $\hat{\beta}_0$ and $\hat{\delta}$ **[2 marks]**.



```
Call:
lm(formula = qsec ~ mpg + hp + vs + hp:vs, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2762 -0.7187 -0.0732  0.4734  3.3835 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.984029   1.562821  14.707 2.07e-14 ***
mpg         -0.185589   0.054908  -3.380  0.00222 ** 
hp          -0.016900   0.004611  -3.665  0.00107 ** 
vs1          4.497939   1.551204   2.900  0.00734 ** 
hp:vs1      -0.022406   0.013600  -1.647  0.11106    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9961 on 27 degrees of freedom
Multiple R-squared:  0.7294,	Adjusted R-squared:  0.6893 
F-statistic: 18.19 on 4 and 27 DF,  p-value: 2.359e-07
```

Equation:

$qsec=(22.984029+4.497939)+(-0.185589)\cdot mpg+(-0.016900-0.022406)\cdot hp$

$qsec=27.481968 -0.185589\cdot mpg-0.039306\cdot hp$

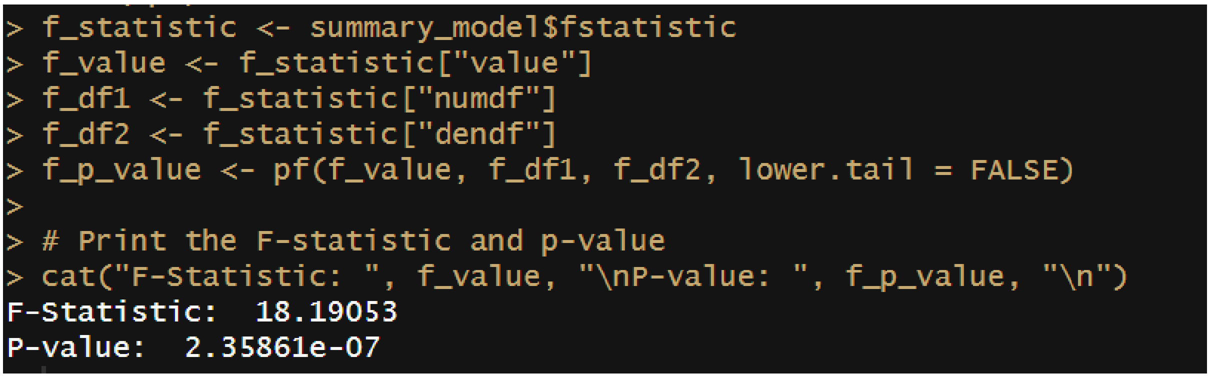
$qsec=27.481968-0.185589\cdot mpg-0.039306\cdot hp$

Interpretation

$\hat{\beta}_0 = 27.481968$: Represents the estimated quarter mile time for a hypothetical vehicle with 0 mpg and 0 hp; this is the baseline o the model.

$\delta = -0.022406$: 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

(c) Using significance level $\alpha = 0.05$, 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]**.

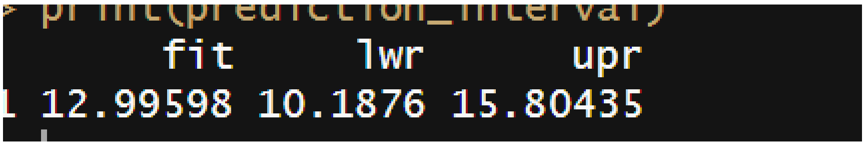


```
> f_statistic <- summary_model$fstatistic
> f_value <- f_statistic["value"]
> f_df1 <- f_statistic["numdf"]
> f_df2 <- f_statistic["dendf"]
> f_p_value <- pf(f_value, f_df1, f_df2, lower.tail = FALSE)
>
> # Print the F-statistic and p-value
> cat("F-Statistic: ", f_value, "\nP-value: ", f_p_value, "\n")
F-Statistic:  18.19053 
P-value:  2.35861e-07 
```

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

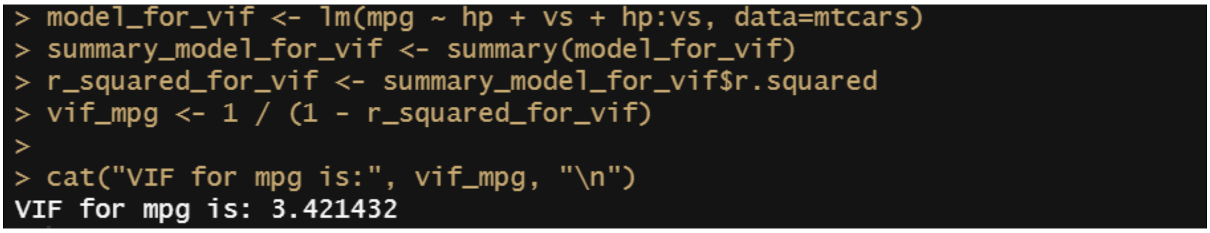
Ha: The predictor variables (mpg, hp, vs, and hp:vs) are significant in predicting the response variable.
Statistic: t = 18.19
P_value= 2.359e-07
Decision: We reject the null hypothesis since the p_value < α (0.05).
Conclusion: Strong evidence suggests that at least one of the predictors (mpg, hp, vs, and hp:vs) provides significant information in predicting the response variable.

(d) 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].**



```
> print(prediction_interval)
       fit     lwr      upr
1 12.99598 10.1876 15.80435
```

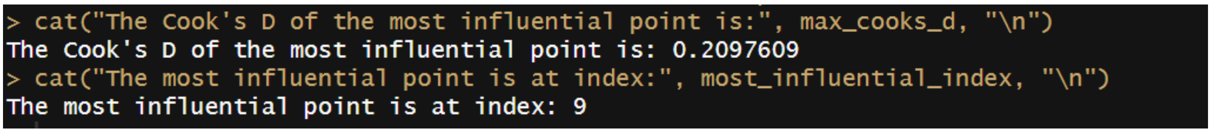
(e) Without using `vif` (or similar) function in R and using only a regression model involving *mpg*, *hp*, *vs* and $vs * hp$, calculate the variance inflation figure (VIF) for the predictor *mpg* in the regression model fitted in part (b) **[3 marks]**.



```
> model_for_vif <- lm(mpg ~ hp + vs + hp:vs, data=mtcars)
> summary_model_for_vif <- summary(model_for_vif)
> r_squared_for_vif <- summary_model_for_vif$r.squared
> vif_mpg <- 1 / (1 - r_squared_for_vif)
>
> cat("VIF for mpg is:", vif_mpg, "\n")
VIF for mpg is: 3.421432
```

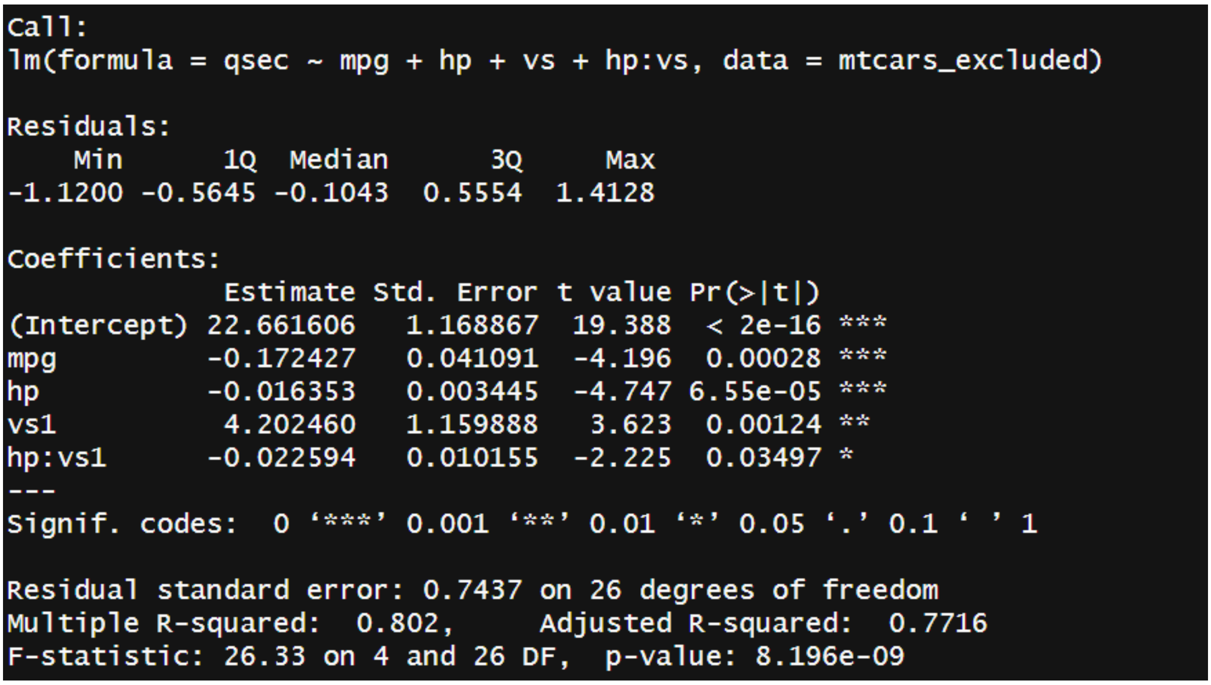
(f) Provide the Cook's D of the most influential point **[1 mark]**, refit the regression model on sample data excluding this point and write down the fitted equation **[2 marks]**.

Most influential point:



```
> cat("The Cook's D of the most influential point is:", max_cooks_d, "\n")
The Cook's D of the most influential point is: 0.2097609
> cat("The most influential point is at index:", most_influential_index, "\n")
The most influential point is at index: 9
```

Model excluding that point:



```
Call:
lm(formula = qsec ~ mpg + hp + vs + hp:vs, data = mtcars_excluded)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1200 -0.5645 -0.1043  0.5554  1.4128 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.661606   1.168867  19.388  < 2e-16 ***
mpg         -0.172427   0.041091  -4.196  0.00028 ***
hp          -0.016353   0.003445  -4.747 6.55e-05 ***
vs1          4.202460   1.159888   3.623  0.00124 ** 
hp:vs1      -0.022594   0.010155  -2.225  0.03497 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7437 on 26 degrees of freedom
Multiple R-squared:  0.802,	Adjusted R-squared:  0.7716 
F-statistic: 26.33 on 4 and 26 DF,  p-value: 8.196e-09
```