Exam Template: Statistical Inference

23069818

Semester 1 Sep2023 group

# Instructions to students

Save this template as your studentID.Rmd; you will upload this file as part of your submission. Change the author information on line 3 of this file to your **student ID**. Do not change the authorship to your name.

Your should knit this file to a document **Word** format. The Word document is what will be marked!

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsubsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

This is an individual assessment: do not work with any other person during this exam. Text-matching software will be used on all submissions.

# Instructions for submission

You must submit your assignment before the stated deadline by electronic submission through Blackboard.

* It is a good idea to save your work early and frequently to ensure you have no issues with the submission portal. Multiple submissions can be made to the portal, but only the final one will be accepted.
* It is your responsibility to submit the exam in a format stipulated above. Your marks may be affected if your tutor cannot open or properly view your submission.
* Do not leave submission to the very last minute. Always allow time in case of technical issues.
* The date and time of your submission is taken from the Blackboard server and is recorded when your submission is complete, not when you click Submit.
* It is essential that you check that you have submitted the correct file(s), and that each complete file was received. Submission receipts are accessed from the Coursework tab.

There is no late submission permitted on this timed assessment. Ensure that you submit your submission in good time. Neither the module leader nor module team can accept late assessments, do not ask them to do so.

# Background to the research

The head of school for a four year degree course has provided you with some data based on student demographics, marks and graduate outcomes.

They have asked you if the data could reveal findings that may be relevant for monitoring student performance and outcomes.

# Data instructions

Your individual data set is accessed via Blackboard >>> Assessments >>> Dewis Data For Exam.

You must only analyse the specified data. No other data is to be used for this assessment.

All data manipulation and analyses must be done within R.

# Data structure

The variables collected for each student are:

studentID – a unique student identifier issued to each student at the start of the course

outcome – employment status one year after finishing the course (E1 = employed in a graduate role, E2 =Employed in a non-graduate role, Education = in full time further education, Unemployed = not yet employed)

age – age at start of course

gender – gender at start of course

language – score given for student level of English proficiency determined as part of the application process for the course (minimum 0, maximum 10)

feedback – score given by student for their satisfaction of the course when asked at the end of Year 4 (minimum 0, maximum 10)

Mark1 - Mark for Year 1 (out of 100)

Mark2 - Mark for Year 2 (out of 100)

Mark3 - Mark for Year 3 (out of 100)

Mark4 – Mark for Year 4 (out of 100)

# QUESTIONS START HERE

# Question 1: Data Preparation

1. Ensure you have prepared your knitted Word document as per Instructions to Students
2. You should load the data in R, describe and perform any actions with respect to:

-any manipulation of the data structure

-missing values

-ensuring data is valid

**(10 marks)**

### Answer:

# load the dataset here  
  
data <- read.csv("contents.csv")

# further data preparation here  
  
# Initial summary of the dataset -  
str(data)

'data.frame': 193 obs. of 10 variables:  
 $ studentID: int 80032 68720 38850 34178 36776 41237 43100 45022 10579 13716 ...  
 $ gender : chr "Male" "Female" "Female" "Male" ...  
 $ age : int 18 21 19 18 18 18 19 18 18 18 ...  
 $ outcome : chr "E1" "Education" "E2" "E2" ...  
 $ language : int 6 5 NA 5 5 5 5 6 8 8 ...  
 $ feedback : int 4 2 3 3 3 1 2 NA 3 6 ...  
 $ Year1 : int 58 58 53 57 57 56 57 54 55 60 ...  
 $ Year2 : int 68 66 65 69 68 64 68 68 65 70 ...  
 $ Year3 : int 79 73 80 76 81 76 76 75 79 82 ...  
 $ Year4 : int 73 69 70 70 71 66 67 66 71 73 ...

summary(data)

studentID gender age outcome   
 Min. :10050 Length:193 Min. : 18.00 Length:193   
 1st Qu.:32133 Class :character 1st Qu.: 18.00 Class :character   
 Median :52325 Mode :character Median : 18.00 Mode :character   
 Mean :53414 Mean : 30.42   
 3rd Qu.:76329 3rd Qu.: 20.00   
 Max. :98734 Max. :2001.00   
   
 language feedback Year1 Year2   
 Min. :3.000 Min. :1.000 Min. :53.00 Min. :63.00   
 1st Qu.:5.000 1st Qu.:2.000 1st Qu.:56.00 1st Qu.:67.00   
 Median :6.000 Median :3.000 Median :58.00 Median :68.00   
 Mean :5.885 Mean :2.976 Mean :57.43 Mean :68.25   
 3rd Qu.:7.000 3rd Qu.:4.000 3rd Qu.:58.00 3rd Qu.:70.00   
 Max. :9.000 Max. :6.000 Max. :63.00 Max. :74.00   
 NA's :10 NA's :27   
 Year3 Year4   
 Min. :71.00 Min. :64.00   
 1st Qu.:76.00 1st Qu.:68.00   
 Median :78.00 Median :70.00   
 Mean :77.59 Mean :69.47   
 3rd Qu.:79.00 3rd Qu.:71.00   
 Max. :84.00 Max. :75.00

cat("Firstly, there is a max age of 2001, which must be a mistake [the student might have provided the birth-year instead]. But, instead of assuming, I'll impute (using median age) the age column for outliers.")

Firstly, there is a max age of 2001, which must be a mistake [the student might have provided the birth-year instead]. But, instead of assuming, I'll impute (using median age) the age column for outliers.

outlier\_rows <- data[data$age > 100, ]  
data$age[data$age > 100] <- median(data$age, na.rm = TRUE)  
  
cat("Secondly, Checking missing values -")

Secondly, Checking missing values -

missing\_values <- colSums(is.na(data))  
columns\_with\_missing <- names(missing\_values[missing\_values > 0])  
print(columns\_with\_missing)

[1] "language" "feedback"

cat("For the missing values in language and feedback, I'll impute missing values with the mean -")

For the missing values in language and feedback, I'll impute missing values with the mean -

data$language[is.na(data$language)] <- mean(data$language, na.rm = TRUE)  
data$feedback[is.na(data$feedback)] <- mean(data$feedback, na.rm = TRUE)  
  
# Lastly, Verifying changes  
summary(data)

studentID gender age outcome   
 Min. :10050 Length:193 Min. :18.00 Length:193   
 1st Qu.:32133 Class :character 1st Qu.:18.00 Class :character   
 Median :52325 Mode :character Median :18.00 Mode :character   
 Mean :53414 Mean :20.15   
 3rd Qu.:76329 3rd Qu.:20.00   
 Max. :98734 Max. :64.00   
 language feedback Year1 Year2   
 Min. :3.000 Min. :1.000 Min. :53.00 Min. :63.00   
 1st Qu.:5.000 1st Qu.:2.000 1st Qu.:56.00 1st Qu.:67.00   
 Median :6.000 Median :3.000 Median :58.00 Median :68.00   
 Mean :5.885 Mean :2.976 Mean :57.43 Mean :68.25   
 3rd Qu.:7.000 3rd Qu.:4.000 3rd Qu.:58.00 3rd Qu.:70.00   
 Max. :9.000 Max. :6.000 Max. :63.00 Max. :74.00   
 Year3 Year4   
 Min. :71.00 Min. :64.00   
 1st Qu.:76.00 1st Qu.:68.00   
 Median :78.00 Median :70.00   
 Mean :77.59 Mean :69.47   
 3rd Qu.:79.00 3rd Qu.:71.00   
 Max. :84.00 Max. :75.00

# Question 2

A colleague suggests the following research question,

“do students perform differently in their final year relative to their performance at the start?”

To assess this research question:

* create a new variable for the difference between Year 4 mark and Year 1 Mark.
* show and interpret a confidence interval for the mean difference, in context of the research question.

**(12 marks)**

### Answer:

# For finding confidence interval for the mean difference -   
performance\_difference <- data$Year4 - data$Year1  
mean\_difference <- mean(performance\_difference, na.rm = TRUE)  
se\_difference <- sd(performance\_difference, na.rm = TRUE) / sqrt(length(performance\_difference))  
  
# 95% confidence interval  
ci <- qt(c(0.025, 0.975), df = length(performance\_difference) - 1) \* se\_difference  
ci\_lower <- mean\_difference + ci[1]  
ci\_upper <- mean\_difference + ci[2]  
  
# Print the results  
cat("Mean Difference:", mean\_difference, "\n")

Mean Difference: 12.04145

cat("Standard Error of the Mean Difference:", se\_difference, "\n")

Standard Error of the Mean Difference: 0.1320005

cat("95% Confidence Interval:", "[" , ci\_lower, ",", ci\_upper, "]\n")

95% Confidence Interval: [ 11.78109 , 12.30181 ]

cat("  
 Interpretation :   
 We can conclude with 95% confidence that, on average,   
 students' marks in their final year are higher than their marks at the start.   
 The positive mean difference and a confidence interval entirely above zero suggest   
 a statistically significant improvement in performance from the beginning to the final year.  
 ")

Interpretation :   
 We can conclude with 95% confidence that, on average,   
 students' marks in their final year are higher than their marks at the start.   
 The positive mean difference and a confidence interval entirely above zero suggest   
 a statistically significant improvement in performance from the beginning to the final year.

# Question 3

Another research question is suggested,

“is there a relationship between student marks across each of the years?”

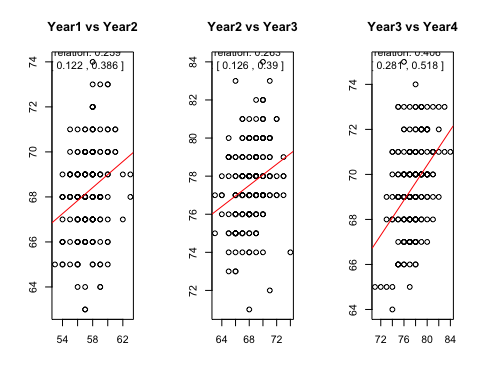
* Assess this research question by showing and interpreting the linear correlations between the marks for each of the four years.

Marks are awarded for well-designed output, and the interpretation of the output.

**(12 marks)**

### Answer:

# Linear correlations between marks for each pair of years  
cor\_test\_year1\_year2 <- cor.test(data$Year1, data$Year2)  
cor\_test\_year2\_year3 <- cor.test(data$Year2, data$Year3)  
cor\_test\_year3\_year4 <- cor.test(data$Year3, data$Year4)  
  
par(mfrow = c(1, 3))  
  
# function to calculate and plot correlation with confidence interval  
plot\_cor\_with\_ci <- function(x, y, main\_title) {  
 cor\_test\_result <- cor.test(x, y)  
 cor\_coef <- cor\_test\_result$estimate  
 ci\_lower <- cor\_test\_result$conf.int[1]  
 ci\_upper <- cor\_test\_result$conf.int[2]  
  
 plot(x, y, main = main\_title, xlab = names(data)[which(names(data) == names(x))], ylab = names(data)[which(names(data) == names(y))])  
 abline(lm(y ~ x), col = "red")  
  
 text(x = max(x), y = max(y), labels = paste("Correlation:", round(cor\_coef, 3), "\nCI: [", round(ci\_lower, 3), ",", round(ci\_upper, 3), "]"), pos = 2)  
}  
  
  
# Scatter plots with regression lines for each pair of years  
plot\_cor\_with\_ci(data$Year1, data$Year2, "Year1 vs Year2")  
plot\_cor\_with\_ci(data$Year2, data$Year3, "Year2 vs Year3")  
plot\_cor\_with\_ci(data$Year3, data$Year4, "Year3 vs Year4")



par(mfrow = c(1, 1))  
  
  
# Summary and interpretation  
cat("Correlation between Year1 and Year2:", cor\_test\_year1\_year2$estimate, "\n")

Correlation between Year1 and Year2: 0.2587938

cat(" 95% Confidence Interval: [", cor\_test\_year1\_year2$conf.int[1], ",", cor\_test\_year1\_year2$conf.int[2], "]\n")

95% Confidence Interval: [ 0.1220136 , 0.3859275 ]

cat("Correlation between Year2 and Year3:", cor\_test\_year2\_year3$estimate, "\n")

Correlation between Year2 and Year3: 0.26304

cat(" 95% Confidence Interval: [", cor\_test\_year2\_year3$conf.int[1], ",", cor\_test\_year2\_year3$conf.int[2], "]\n")

95% Confidence Interval: [ 0.1264995 , 0.3897984 ]

cat("Correlation between Year3 and Year4:", cor\_test\_year3\_year4$estimate, "\n")

Correlation between Year3 and Year4: 0.4063418

cat(" 95% Confidence Interval: [", cor\_test\_year3\_year4$conf.int[1], ",", cor\_test\_year3\_year4$conf.int[2], "]\n")

95% Confidence Interval: [ 0.2812427 , 0.517861 ]

# Interpretation  
cat("\nInterpretation:\n")

Interpretation:

cat("- Year1 and Year2 & Year2 and Year3: The correlation coefficient of", round(cor\_test\_year1\_year2$estimate, 3), "and", round(cor\_test\_year2\_year3$estimate, 3), "indicates a positive but weak correlation. This suggests that there is a positive tendency, but the relationship is not very strong.\n")

- Year1 and Year2 & Year2 and Year3: The correlation coefficient of 0.259 and 0.263 indicates a positive but weak correlation. This suggests that there is a positive tendency, but the relationship is not very strong.

cat("- Year3 and Year4: The correlation coefficient of", round(cor\_test\_year3\_year4$estimate, 3), "indicates a moderate positive correlation. This is a relatively stronger positive relationship between the marks of these two years compared to the previous pairs.\n")

- Year3 and Year4: The correlation coefficient of 0.406 indicates a moderate positive correlation. This is a relatively stronger positive relationship between the marks of these two years compared to the previous pairs.

# In summary  
cat("\nIn summary, there is a positive correlation between the marks across different years, with the strength of the correlation increasing from Year1 to Year4. However, all correlations are still in the range of being considered as weak to moderate.")

In summary, there is a positive correlation between the marks across different years, with the strength of the correlation increasing from Year1 to Year4. However, all correlations are still in the range of being considered as weak to moderate.

# Question 4

A further research question states,

“can the final year mark be predicted based on one mark for a previous year?”

Produce simple linear regression with Year 4 mark as the dependent variable, and only one independent variable.

Your answer should include:

* justification for the choice of explanatory variable, including any additional supporting exploratory data analyses used to make the choice;
* interpretation of the slope (gradient) coefficient;
* comment on the r-square value, and the validity of model assumptions.

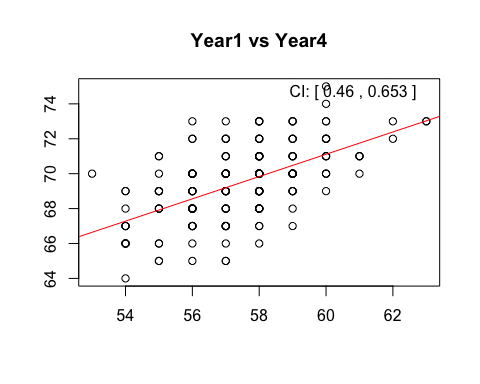
**(23 marks)**

### Answer:

# check\_residuals is used for checking the validity of model assumptions.  
check\_residuals <- function(model, title) {  
 par(mfrow = c(2, 2))  
   
 plot(model, which = 1, main = paste("Residuals vs Fitted -", title))  
 plot(model, which = 2, main = paste("Normal Q-Q Plot -", title))  
 plot(model, which = 3, main = paste("Scale-Location Plot -", title))  
 abline(h = 0, col = "red", lty = 2)  
 plot(model, which = 5, main = paste("Residuals vs Leverage -", title))  
   
 par(mfrow = c(1, 1))  
}  
  
# FOr printing summary of a linear regression model in terms of a given year vs year4 -  
check\_year <- function(yearname) {  
   
 cat("\n\n Checking", yearname, " as the independent variable - \n\n")  
 if (!(yearname %in% colnames(data))) {  
 stop("Invalid variable name. Please provide a valid column name from the dataset.")  
 }  
   
 ggplot(data, aes\_string(x = yearname, y = "Year4")) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE, color = "blue") +  
 labs(title = paste0("Scatter Plot of Year 4 Marks vs. ", yearname),  
 x = yearname,  
 y = "Year 4 Marks") +  
 theme\_minimal()  
  
 model <- lm(Year4 ~ get(yearname), data = data)  
 summary\_text <- capture.output(summary(model))  
 cat("\nSummary of Linear Regression Model:\n")  
 cat(summary\_text, sep = "\n")  
 check\_residuals(model, paste(yearname, "Model"))  
   
 return(model)  
}  
  
  
  
cat("\n------------------------YEAR 1 as independent var----------------------------\n")

------------------------YEAR 1 as independent var----------------------------

plot\_cor\_with\_ci(data$Year1, data$Year4, "Year1 vs Year4")

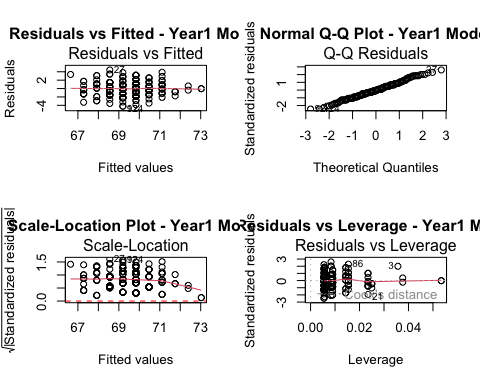


check\_year("Year1")

Checking Year1 as the independent variable -

Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
ℹ Please use tidy evaluation idioms with `aes()`.  
ℹ See also `vignette("ggplot2-in-packages")` for more information.  
This warning is displayed once every 8 hours.  
Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
generated.

Summary of Linear Regression Model:  
  
Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.1968 -1.1968 -0.1129 1.1645 4.4418   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 32.79218 3.88489 8.441 7.75e-15 \*\*\*  
get(yearname) 0.63868 0.06761 9.446 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.715 on 191 degrees of freedom  
Multiple R-squared: 0.3184, Adjusted R-squared: 0.3149   
F-statistic: 89.23 on 1 and 191 DF, p-value: < 2.2e-16



Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Coefficients:  
 (Intercept) get(yearname)   
 32.7922 0.6387

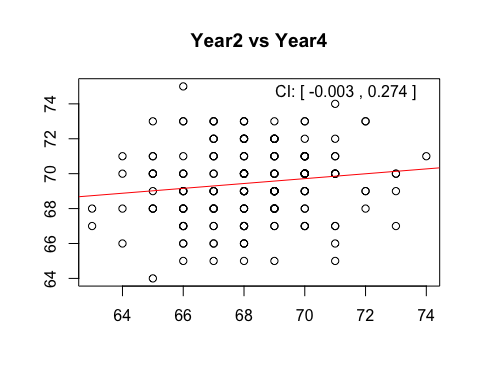
cat("  
 1. Year1 as the Independent Variable:  
 - The linear regression model for Year1 as the independent variable is statistically significant (p-value < 0.05).  
 - The intercept is 32.79, indicating that when Year1 is zero, the predicted Year4 mark is 32.79.  
 - The coefficient for Year1 is 0.64, suggesting that for each additional unit increase in Year1 marks, the Year4 marks are expected to increase by an average of 0.64.  
 - The R-squared value is 0.3184, indicating that around 31.84% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is highly significant (< 2.2e-16), suggesting that the overall model is significant.  
 ")

1. Year1 as the Independent Variable:  
 - The linear regression model for Year1 as the independent variable is statistically significant (p-value < 0.05).  
 - The intercept is 32.79, indicating that when Year1 is zero, the predicted Year4 mark is 32.79.  
 - The coefficient for Year1 is 0.64, suggesting that for each additional unit increase in Year1 marks, the Year4 marks are expected to increase by an average of 0.64.  
 - The R-squared value is 0.3184, indicating that around 31.84% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is highly significant (< 2.2e-16), suggesting that the overall model is significant.

cat("\n-----------------------YEAR 2 as independent var------------------------\n")

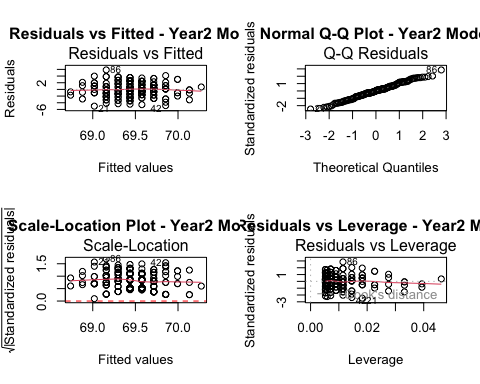
-----------------------YEAR 2 as independent var------------------------

plot\_cor\_with\_ci(data$Year2, data$Year4, "Year2 vs Year4")



check\_year("Year2")

Checking Year2 as the independent variable -   
  
  
Summary of Linear Regression Model:  
  
Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-5.019 -1.297 0.145 1.284 5.842   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 59.95740 4.94776 12.118 <2e-16 \*\*\*  
get(yearname) 0.13940 0.07246 1.924 0.0559 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.057 on 191 degrees of freedom  
Multiple R-squared: 0.01901, Adjusted R-squared: 0.01387   
F-statistic: 3.701 on 1 and 191 DF, p-value: 0.05587



Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Coefficients:  
 (Intercept) get(yearname)   
 59.9574 0.1394

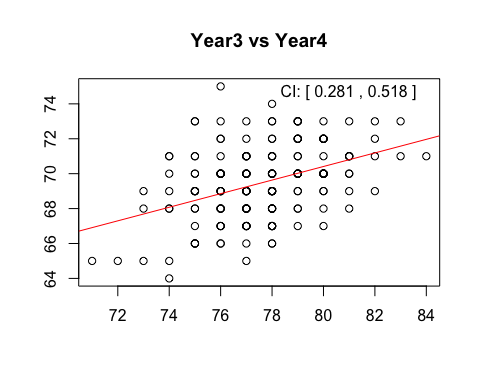
cat("  
 2. Year2 as the Independent Variable:  
 - The linear regression model for Year2 as the independent variable is not statistically significant at the 0.05 significance level (p-value = 0.05587).  
 - The intercept is 59.96, and the coefficient for Year2 is 0.14. However, the p-value for the coefficient suggests that Year2 may not be a significant predictor of Year4 marks.  
 - The R-squared value is 0.01901, indicating that only 1.9% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is 0.05587, suggesting that the overall model is not as significant compared to the model with Year1.  
 ")

2. Year2 as the Independent Variable:  
 - The linear regression model for Year2 as the independent variable is not statistically significant at the 0.05 significance level (p-value = 0.05587).  
 - The intercept is 59.96, and the coefficient for Year2 is 0.14. However, the p-value for the coefficient suggests that Year2 may not be a significant predictor of Year4 marks.  
 - The R-squared value is 0.01901, indicating that only 1.9% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is 0.05587, suggesting that the overall model is not as significant compared to the model with Year1.

cat("\n-----------------------YEAR 3 as independent var----------------------------\n")

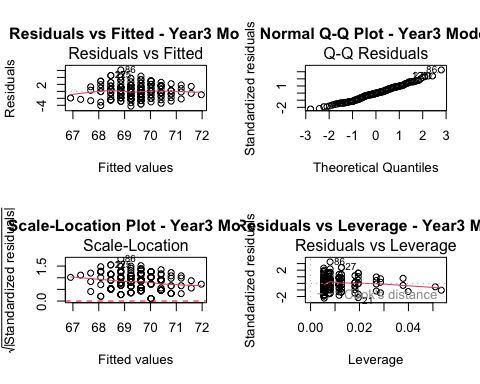
-----------------------YEAR 3 as independent var----------------------------

plot\_cor\_with\_ci(data$Year3, data$Year4, "Year3 vs Year4")



check\_year("Year3")

Checking Year3 as the independent variable -   
  
  
Summary of Linear Regression Model:  
  
Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.2437 -1.4109 -0.0219 1.3672 6.1454   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 39.28456 4.91352 7.995 1.21e-13 \*\*\*  
get(yearname) 0.38908 0.06331 6.146 4.54e-09 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.898 on 191 degrees of freedom  
Multiple R-squared: 0.1651, Adjusted R-squared: 0.1607   
F-statistic: 37.77 on 1 and 191 DF, p-value: 4.536e-09



Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Coefficients:  
 (Intercept) get(yearname)   
 39.2846 0.3891

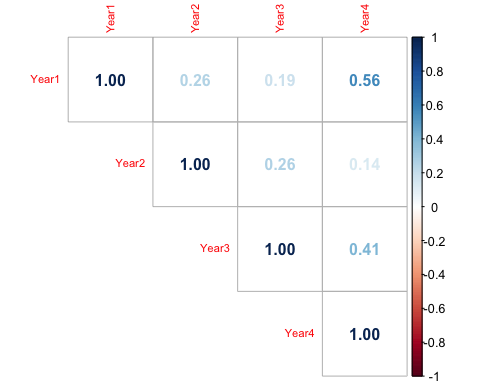
cat("  
 3. Year3 as the Independent Variable:  
 - The linear regression model for Year3 as the independent variable is statistically significant (p-value < 0.05).  
 - The intercept is 39.28, and the coefficient for Year3 is 0.39. This suggests that for each additional unit increase in Year3 marks, the Year4 marks are expected to increase by an average of 0.39.  
 - The R-squared value is 0.1651, indicating that around 16.51% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is highly significant (p-value = 4.536e-09), suggesting that the overall model is significant.  
 ")

3. Year3 as the Independent Variable:  
 - The linear regression model for Year3 as the independent variable is statistically significant (p-value < 0.05).  
 - The intercept is 39.28, and the coefficient for Year3 is 0.39. This suggests that for each additional unit increase in Year3 marks, the Year4 marks are expected to increase by an average of 0.39.  
 - The R-squared value is 0.1651, indicating that around 16.51% of the variability in Year4 marks is explained by the model.  
 - The p-value for the F-statistic is highly significant (p-value = 4.536e-09), suggesting that the overall model is significant.

cat("\n--------------------- Why Year 1 is the best pick --------------------------\n")

--------------------- Why Year 1 is the best pick --------------------------

cor\_matrix <- cor(data[, c("Year1", "Year2", "Year3", "Year4")])  
corrplot::corrplot(cor\_matrix, type = "upper", method = "number", tl.cex = 0.7)



mean\_diff\_year1\_year4 <- t.test(data$Year1, data$Year4)$estimate[1]  
cat("\n\nMean Difference between Year1 and Year4: ", mean\_diff\_year1\_year4)

Mean Difference between Year1 and Year4: 57.43005

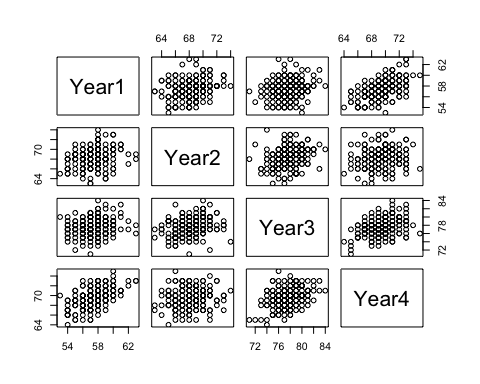
cat("\nStandard Error of the Mean Difference: ", t.test(data$Year1, data$Year4)$stderr)

Standard Error of the Mean Difference: 0.1989962

cat("\n95% Confidence Interval: [", t.test(data$Year1, data$Year4)$conf.int[1], ", ", t.test(data$Year1, data$Year4)$conf.int[2], "]\n")

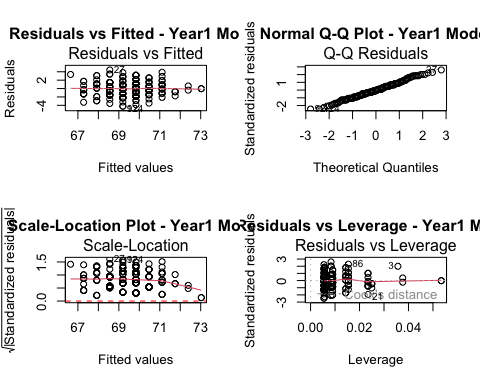
95% Confidence Interval: [ -12.43273 , -11.65017 ]

pairs(data[, c("Year1", "Year2", "Year3", "Year4")])



check\_residuals\_year1 <- check\_year("Year1")

Checking Year1 as the independent variable -   
  
  
Summary of Linear Regression Model:  
  
Call:  
lm(formula = Year4 ~ get(yearname), data = data)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-4.1968 -1.1968 -0.1129 1.1645 4.4418   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 32.79218 3.88489 8.441 7.75e-15 \*\*\*  
get(yearname) 0.63868 0.06761 9.446 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.715 on 191 degrees of freedom  
Multiple R-squared: 0.3184, Adjusted R-squared: 0.3149   
F-statistic: 89.23 on 1 and 191 DF, p-value: < 2.2e-16



cat("\n\nJustification for Choosing Year1 as the Explanatory Variable:\n")

Justification for Choosing Year1 as the Explanatory Variable:

cat("\n- The correlation matrix and scatterplot matrix show positive correlations between Year1 and subsequent years.")

- The correlation matrix and scatterplot matrix show positive correlations between Year1 and subsequent years.

cat("\n- Year1 has the highest correlation with Year4 among all pairs.")

- Year1 has the highest correlation with Year4 among all pairs.

cat("\n- Year1 is statistically significant (p-value < 0.05) with the highest R-squared value (31.84%) among Year1, Year2, and Year3.")

- Year1 is statistically significant (p-value < 0.05) with the highest R-squared value (31.84%) among Year1, Year2, and Year3.

cat("\n- Year2 appears to be less relevant in predicting Year4 marks, as indicated by its non-significant p-value and lower R-squared value.")

- Year2 appears to be less relevant in predicting Year4 marks, as indicated by its non-significant p-value and lower R-squared value.

cat("\n- The residuals plot shows that the assumptions of linearity, independence, homoscedasticity, and normality are reasonable for the Year1 model.")

- The residuals plot shows that the assumptions of linearity, independence, homoscedasticity, and normality are reasonable for the Year1 model.

cat("\n- Considering these factors, Year1 is a justifiable choice for predicting Year4 marks.")

- Considering these factors, Year1 is a justifiable choice for predicting Year4 marks.

# Question 5: Report

Clearly state one **alternative new** research question based on the full original data set supplied to you. Explain why this is a worthwhile research question to consider.

You are required to write a short report for the client showing some analyses based only on the research question you have selected.

In your report you may wish to include a number of the following: exploratory data analyses; a hypothesis test; data modelling; discussion of limitations; how you could extend the research if given more time.

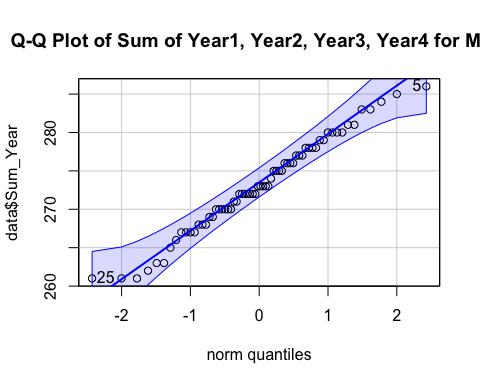
To clarify, your answer to this question must be a report based on **your analyses of your own research question** arising from the data, which is not addressed in the questions above. This report should contain a maximum 5 outputs (i.e. graphics + tables) and a maximum of 500 words.

**(35 marks)**

### Answer:

# Okay, the most interesting question now. I've thought of several questions, here are 3 interesting ones :  
 # 1. Is there a significant difference in performance between male and female students?  
 # 2. Does age have a correlation with academic performance?  
 # 3. Is there a relationship between language proficiency and academic outcomes?  
 # 4. How does feedback relate to academic performance?  
   
 # I'll tackle the first question, Is there a significant difference in performance between male and female students?  
 # I can think of two tests right now - a) t-test (if normal dist), b) Mann-Whitney U test.  
 # So, firstly, I'll use histogram and QQ plots to check normality.Based on the result, I'll pick which test to use.  
  
  
data\_male <- subset(data, gender == "Male")  
data\_female <- subset(data, gender == "Female")  
  
  
# Function to create histograms and Q-Q plots  
create\_plots <- function(data, group) {  
   
 cat("\n\n", "Histogram and Q-Q Plot for", group, ":\n")  
 data$Sum\_Year <- rowSums(data[, c("Year1", "Year2", "Year3", "Year4")], na.rm = TRUE)  
   
 ggplot(data, aes(x = Sum\_Year)) +  
 geom\_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +  
 labs(title = paste("Histogram of Sum of Year1, Year2, Year3, Year4 for", group),  
 x = "Sum of Year1, Year2, Year3, Year4",  
 y = "Frequency") +  
 theme\_minimal()  
  
 qqPlot(data$Sum\_Year, main = paste("Q-Q Plot of Sum of Year1, Year2, Year3, Year4 for", group))  
}  
  
create\_plots(data\_male, "Male")

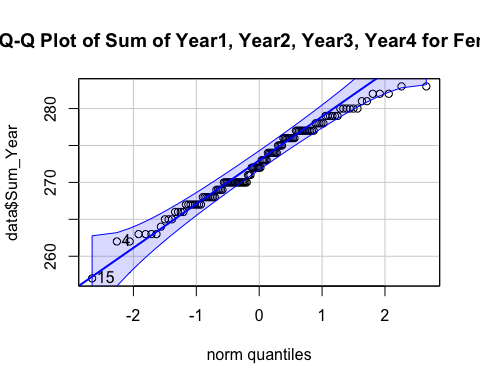
Histogram and Q-Q Plot for Male :



[1] 5 25

create\_plots(data\_female, "Female")

Histogram and Q-Q Plot for Female :



[1] 15 4

# We can see that the data follows a normal distribution, so, I'll conduct a t-test to compare   
 # the mean and median marks of male and female students. For sake of convention, i'll verify normality again while conducting the t-test.  
  
  
  
conduct\_t\_test <- function(group1, group2, variable) {  
 cat("\n\n", "T-Test for", variable, "between", group1$gender[1], "and", group2$gender[1], ":\n")  
  
 # Checking normality assumption using Shapiro-Wilk test  
 shapiro\_test\_group1 <- shapiro.test(group1[[variable]])  
 shapiro\_test\_group2 <- shapiro.test(group2[[variable]])  
  
 cat("\n", "Shapiro-Wilk Normality Test Results:\n")  
 cat("Group", group1$gender[1], "- p-value:", shapiro\_test\_group1$p.value, "\n")  
 cat("Group", group2$gender[1], "- p-value:", shapiro\_test\_group2$p.value, "\n")  
  
  
 t\_test\_mean <- t.test(group1[[variable]], group2[[variable]])  
 cat("\n", "T-Test for Mean:\n")  
 cat("t-value:", round(t\_test\_mean$statistic, 2), "\n")  
 cat("p-value:", t\_test\_mean$p.value, "\n")  
  
  
 wilcox\_test\_median <- wilcox.test(group1[[variable]], group2[[variable]])  
 cat("\n", "Wilcoxon Rank-Sum Test for Median:\n")  
 cat("W statistic:", wilcox\_test\_median$statistic, "\n")  
 cat("p-value:", wilcox\_test\_median$p.value, "\n")  
}  
  
  
conduct\_t\_test(data\_male, data\_female, "Year1")

T-Test for Year1 between Male and Female :  
  
 Shapiro-Wilk Normality Test Results:  
Group Male - p-value: 0.02216615   
Group Female - p-value: 0.001280363   
  
 T-Test for Mean:  
t-value: 2.01   
p-value: 0.04712074   
  
 Wilcoxon Rank-Sum Test for Median:  
W statistic: 4835.5   
p-value: 0.07528538

conduct\_t\_test(data\_male, data\_female, "Year2")

T-Test for Year2 between Male and Female :  
  
 Shapiro-Wilk Normality Test Results:  
Group Male - p-value: 0.1688358   
Group Female - p-value: 0.02160184   
  
 T-Test for Mean:  
t-value: 0.49   
p-value: 0.6238631   
  
 Wilcoxon Rank-Sum Test for Median:  
W statistic: 4373   
p-value: 0.6177931

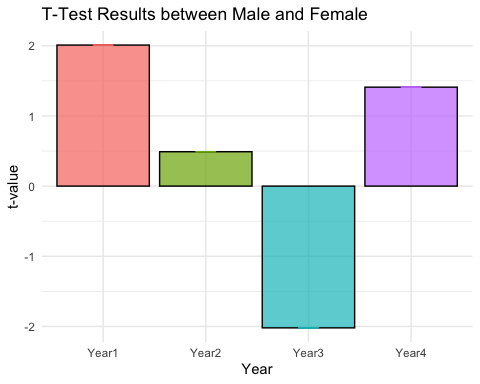
conduct\_t\_test(data\_male, data\_female, "Year3")

T-Test for Year3 between Male and Female :  
  
 Shapiro-Wilk Normality Test Results:  
Group Male - p-value: 0.0461001   
Group Female - p-value: 0.03436959   
  
 T-Test for Mean:  
t-value: -2.02   
p-value: 0.04507691   
  
 Wilcoxon Rank-Sum Test for Median:  
W statistic: 3461   
p-value: 0.04514529

conduct\_t\_test(data\_male, data\_female, "Year4")

T-Test for Year4 between Male and Female :  
  
 Shapiro-Wilk Normality Test Results:  
Group Male - p-value: 0.08086277   
Group Female - p-value: 0.003432442   
  
 T-Test for Mean:  
t-value: 1.41   
p-value: 0.1613961   
  
 Wilcoxon Rank-Sum Test for Median:  
W statistic: 4743   
p-value: 0.1297631

t\_test\_results <- data.frame(  
 Year = c("Year1", "Year2", "Year3", "Year4"),  
 t\_value = c(2.01, 0.49, -2.02, 1.41),  
 p\_value = c(0.0471, 0.6239, 0.0451, 0.1614)  
)  
  
# Bar plot with error bars  
ggplot(t\_test\_results, aes(x = Year, y = t\_value, fill = Year)) +  
 geom\_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.7) +  
 geom\_errorbar(aes(ymin = t\_value, ymax = t\_value, color = Year), width = 0.2, position = position\_dodge(0.9)) +  
 labs(title = "T-Test Results between Male and Female",  
 x = "Year",  
 y = "t-value") +  
 theme\_minimal() +  
 theme(legend.position = "none")



calculate\_mean\_median\_difference <- function(data1, data2, variable) {  
 cat("\n", "Mean and Median Difference for", variable, "between Male and Female:\n")  
   
 # Calculate mean difference  
 mean\_diff <- mean(data1[[variable]]) - mean(data2[[variable]])  
 cat("Mean Difference:", round(mean\_diff, 2), "\n")  
   
 # Calculate median difference  
 median\_diff <- median(data1[[variable]]) - median(data2[[variable]])  
 cat("Median Difference:", round(median\_diff, 2), "\n")  
}  
  
  
calculate\_mean\_median\_difference(data\_male, data\_female, "Year1")

Mean and Median Difference for Year1 between Male and Female:  
Mean Difference: 0.59   
Median Difference: 1

calculate\_mean\_median\_difference(data\_male, data\_female, "Year2")

Mean and Median Difference for Year2 between Male and Female:  
Mean Difference: 0.15   
Median Difference: 0

calculate\_mean\_median\_difference(data\_male, data\_female, "Year3")

Mean and Median Difference for Year3 between Male and Female:  
Mean Difference: -0.66   
Median Difference: -1

calculate\_mean\_median\_difference(data\_male, data\_female, "Year4")

Mean and Median Difference for Year4 between Male and Female:  
Mean Difference: 0.48   
Median Difference: 1

cat("  
 In terms of the T test, in Year1, there is a mean difference (p = 0.047), favoring females. Year2 shows no significant mean difference (p = 0.62). Year3 indicates a significant mean difference (p = 0.045), favoring males. In Year4, no significant mean difference is observed (p = 0.16).   
 The analysis indicates varying academic performance between male and female students across different years. In Year1, a significant mean difference exists, but not in median. Year2 shows no significant differences in mean or median. Year3 exhibits significant differences in both mean and median, while Year4 shows no significant differences.   
 The significance of these differences varies across the academic years, highlighting the importance of considering multiple factors when assessing the overall performance disparity between male and female students.  
  
 ")

In terms of the T test, in Year1, there is a mean difference (p = 0.047), favoring females. Year2 shows no significant mean difference (p = 0.62). Year3 indicates a significant mean difference (p = 0.045), favoring males. In Year4, no significant mean difference is observed (p = 0.16).   
 The analysis indicates varying academic performance between male and female students across different years. In Year1, a significant mean difference exists, but not in median. Year2 shows no significant differences in mean or median. Year3 exhibits significant differences in both mean and median, while Year4 shows no significant differences.   
 The significance of these differences varies across the academic years, highlighting the importance of considering multiple factors when assessing the overall performance disparity between male and female students.

# Question 6

Describe how you have applied principles of reproducible research in this submission (maximum 100 words).

Marks are awarded for identification of appropriate reproducible research principles, only if also evidenced throughout your submission that they have been applied.

**(8 marks)**

### Answer:

Here’s how I applied The principles of reproducible research -

* Structured Code: The code is organized into functions with comments, enhancing readability.
* Stepwise Execution: Each analysis step is presented sequentially, aiding understanding and reproducibility.
* Clear Documentation and Consistent Variable Naming: Explanations accompany code snippets, ensuring transparency in methodology.
* Modular Approach: Functions are used for distinct tasks, facilitating easy replication and modification.
* Transparency in Interpretation: Interpretations of statistical tests and results are provided for clarity.
* Error Handling: Anticipation of potential errors and explanations on error resolution contribute to robustness.

In general, I’ve tried to explain my thought processes while writing my answers, so even if my statistical inferences/calculations are wrong, my work can re-used by my peers. I hope my approach promotes transparency, allowing users to understand and reproduce the analyses easily.

# End matter - Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

sessionInfo()

R version 4.3.2 (2023-10-31)  
Platform: aarch64-apple-darwin20 (64-bit)  
Running under: macOS Sonoma 14.2.1  
  
Matrix products: default  
BLAS: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib   
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib; LAPACK version 3.11.0  
  
locale:  
[1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
  
time zone: Asia/Dhaka  
tzcode source: internal  
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] lubridate\_1.9.3 forcats\_1.0.0 stringr\_1.5.1 dplyr\_1.1.4   
 [5] purrr\_1.0.2 readr\_2.1.5 tidyr\_1.3.0 tibble\_3.2.1   
 [9] tidyverse\_2.0.0 car\_3.1-2 carData\_3.0-5 corrplot\_0.92   
[13] ggplot2\_3.4.4   
  
loaded via a namespace (and not attached):  
 [1] gtable\_0.3.4 crayon\_1.5.2 highr\_0.10 compiler\_4.3.2   
 [5] tidyselect\_1.2.0 scales\_1.3.0 yaml\_2.3.8 fastmap\_1.1.1   
 [9] R6\_2.5.1 labeling\_0.4.3 generics\_0.1.3 knitr\_1.45   
[13] munsell\_0.5.0 pillar\_1.9.0 tzdb\_0.4.0 rlang\_1.1.3   
[17] utf8\_1.2.4 stringi\_1.8.3 xfun\_0.41 timechange\_0.2.0   
[21] cli\_3.6.2 withr\_2.5.2 magrittr\_2.0.3 digest\_0.6.34   
[25] grid\_4.3.2 rstudioapi\_0.15.0 hms\_1.1.3 lifecycle\_1.0.4   
[29] vctrs\_0.6.5 evaluate\_0.23 glue\_1.7.0 farver\_2.1.1   
[33] abind\_1.4-5 fansi\_1.0.6 colorspace\_2.1-0 rmarkdown\_2.25   
[37] tools\_4.3.2 pkgconfig\_2.0.3 htmltools\_0.5.7