Multiple Regression & Model Diagnostics: A Simulated Case Study

Benson Nyakundi Onyiego

A university research team aims to examine the determinants of students’ course performance in a quantitative subject. The dataset contains the following variables:

1. Course Assessment Score (Y) – Dependent Variable
2. Average Weekly Study Time (X_1) – Independent Variable
3. Class Participation Rate (X_2) – Independent Variable
4. Cumulative GPA Prior to the Course (X_3) – Independent Variable
5. Use of Online Learning Resources (Hours per Week) (X_4) – Independent Variable

REQUIRED TASKS

1. DATA PREPARATION

a. examine missing values, outliers, and coding errors.
b. Compute descriptive statistics (mean, median, standard deviation).
c. Use appropriate graphs to describe the distribution of variables and the relationships between the dependent variable and each predictor.

2. SIMPLE LINEAR REGRESSION

a. Estimate a simple linear regression model to explain Course Assessment Score (Y) using Average Weekly Study Time (X_1) as the explanatory variable.
b. Write down the estimated regression equation.
c. Interpret the slope coefficient of Average Weekly Study Time and comment on its statistical significance.
d. Explain the meaning of the R-squared obtained from the model.

3. MULTIPLE LINEAR REGRESSION

a. Estimate a multiple linear regression model using the four independent variables (X_1, X_2, X_3, X_4) to explain Course Assessment Score (Y).
b. Present the estimated regression equation.
c. Interpret the estimated coefficients of each independent variable.
d. Examine the presence of multicollinearity using the Variance Inflation Factor (VIF) and explain its implications for the model.
e. Assess the overall significance of the model using the ANOVA table and the R-squared value.

4. MODEL DIAGNOSTICS AND ASSUMPTIONS

a. Examine the normality of residuals using histograms and Normal Q–Q plots.
b. Test for homoscedasticity using a plot of standardized residuals against predicted values.
c. Detect any influential observations or outliers using Cook’s Distance and/or Leverage statistics, and comment on their impact on the model.

5. Conclusion and Recommendations • Summarize key findings from your regression analysis. • Provide recommendations based on the model results. • Discuss limitations of the study and potential improvements.

The dataset used for this analysis was generated via R with N=100

```
set.seed(42)
n <- 100

study_time <- pmax(2, pmin(25, rnorm(n, mean = 12, sd = 4)))
participation <- pmax(30, pmin(100, rnorm(n, mean = 75, sd = 12)))
gpa <- pmax(2.0, pmin(4.0, rnorm(n, mean = 3.1, sd = 0.4)))
online_resources <- pmax(0, pmin(15, rnorm(n, mean = 6, sd = 2.5)))

noise <- rnorm(n, mean = 0, sd = 5)
score <- 15 + 1.2*study_time + 0.25*participation + 8.0*gpa + 0.8*online_resources + noise
score <- pmax(0, pmin(100, score))

df <- data.frame(
  StudentID = 1001:(1000 + n),
  StudyTime_X1 = round(study_time, 1),
  Participation_X2 = round(participation, 1),
  GPA_X3 = round(gpa, 2),
  OnlineResources_X4 = round(online_resources, 1),
  CourseScore_Y = round(score, 1)
)

str(df)
```

```
## 'data.frame':    100 obs. of  6 variables:
##  $ StudentID         : int  1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 ...
##  $ StudyTime_X1      : num  17.5 9.7 13.5 14.5 13.6 11.6 18 11.6 20.1 11.7 ...
##  $ Participation_X2  : num  89.4 87.5 63 97.2 67 76.3 69.9 73.5 77.3 76.4 ...
##  $ GPA_X3            : num  2.3 3.23 3.57 3.92 2.55 2.64 2.82 2.68 2.84 3.03 ...
##  $ OnlineResources_X4: num  6 7.9 6.1 7.8 5.6 5.9 7.2 8.5 2.9 5.9 ...
##  $ CourseScore_Y     : num  88.2 76.4 80.6 94.6 70.1 68.8 82.4 78.8 90.8 67.6 ...
```

QUESTION 1

```
#question 1a

# 1. Check for Missing Values across all columns
missing_counts <- colSums(is.na(df))
print("--- Missing Values Count ---")
```

```
## [1] "--- Missing Values Count ---"
```

```
print(missing_counts)
```

```
##          StudentID       StudyTime_X1   Participation_X2             GPA_X3
##                  0                  0                  0                  0
## OnlineResources_X4      CourseScore_Y
##                  0                  0
```

```
# 2. Check structure and data types for coding errors
print("--- Data Structure and Summary ---")
```

```
## [1] "--- Data Structure and Summary ---"
```

```
summary(df)
```

```
##    StudentID     StudyTime_X1   Participation_X2     GPA_X3     
##  Min.   :1001   Min.   : 2.00   Min.   : 50.70   Min.   :2.020  
##  1st Qu.:1026   1st Qu.: 9.55   1st Qu.: 67.90   1st Qu.:2.817  
##  Median :1050   Median :12.40   Median : 74.20   Median :3.090  
##  Mean   :1050   Mean   :12.15   Mean   : 73.88   Mean   :3.095  
##  3rd Qu.:1075   3rd Qu.:14.62   3rd Qu.: 80.58   3rd Qu.:3.357  
##  Max.   :1100   Max.   :21.10   Max.   :100.00   Max.   :4.000  
##  OnlineResources_X4 CourseScore_Y  
##  Min.   : 1.800     Min.   :55.60  
##  1st Qu.: 4.675     1st Qu.:71.47  
##  Median : 5.900     Median :77.00  
##  Mean   : 6.082     Mean   :77.09  
##  3rd Qu.: 7.700     3rd Qu.:82.00  
##  Max.   :12.100     Max.   :96.30
```

```
# 3. Detect Outliers using Z-scores or IQR (Interquartile Range)
# Identify numeric variables excluding StudentID
analysis_vars <- c("StudyTime_X1", "Participation_X2", "GPA_X3", "OnlineResources_X4", "CourseScore_Y")

# Function to check outliers via Boxplot stats
outliers_list <- lapply(df[analysis_vars], function(x) boxplot.stats(x)$out)
print("--- Outliers Detected Per Variable ---")
```

```
## [1] "--- Outliers Detected Per Variable ---"
```

```
print(outliers_list)
```

```
## $StudyTime_X1
## numeric(0)
## 
## $Participation_X2
## [1] 100
## 
## $GPA_X3
## numeric(0)
## 
## $OnlineResources_X4
## numeric(0)
## 
## $CourseScore_Y
## [1] 55.6
```

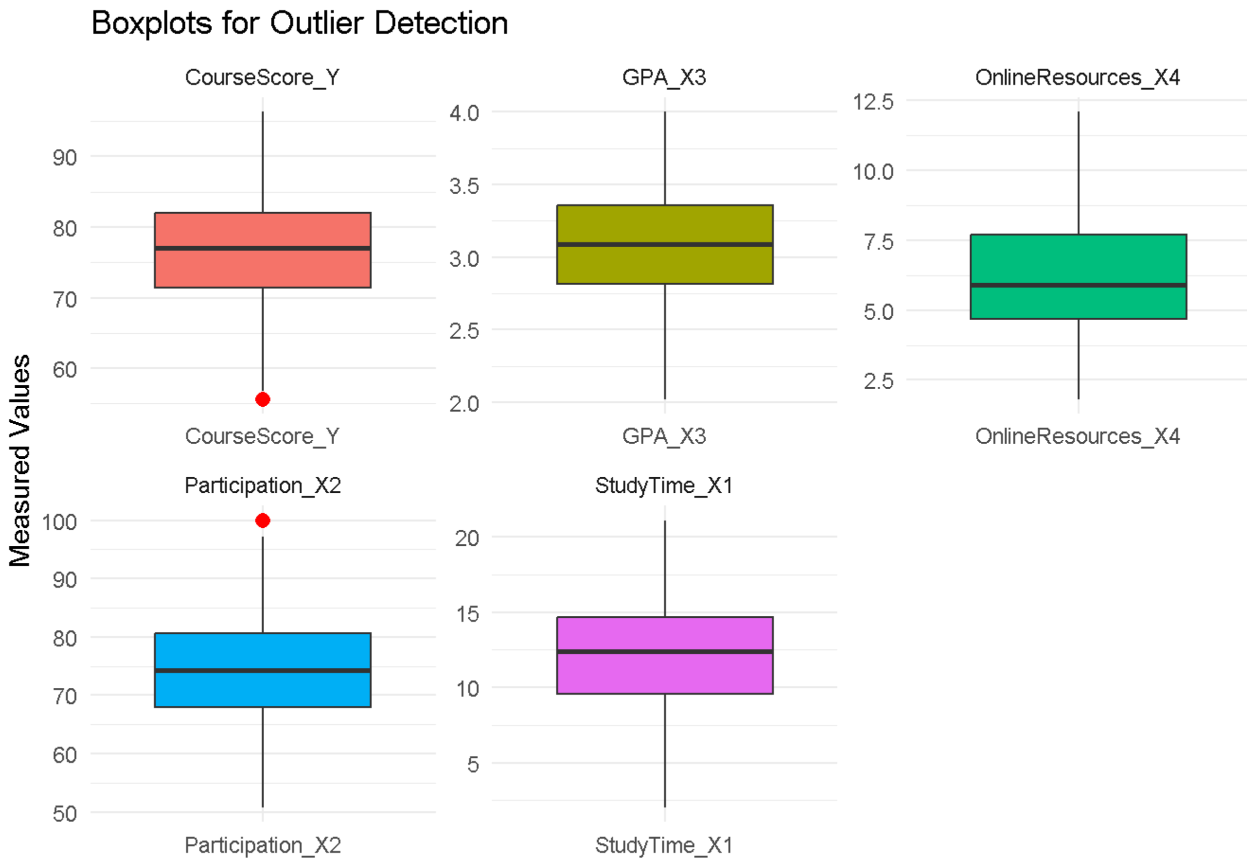
```
#visualization using boxplots
if (!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
if (!require(tidyr)) install.packages("tidyr")
```

```
## Loading required package: tidyr
```

```
library(ggplot2)
library(tidyr)

# Reshape data
df_long <- pivot_longer(
  df,
  cols = all_of(analysis_vars),
  names_to = "Variable",
  values_to = "Value"
)

# Plot faceted boxplots
ggplot(df_long, aes(x = Variable, y = Value, fill = Variable)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2.5) +
  facet_wrap(~ Variable, scales = "free") +
  theme_minimal() +
  labs(
    title = "Boxplots for Outlier Detection",
    x = "",
    y = "Measured Values"
  ) +
  theme(legend.position = "none")
```



a. Data Cleaning, Missingness, and Outlier Diagnostics

Data integrity was verified across all 100 observations using R data validation routines.

Missing Value Analysis: System-wide evaluation via colSums(is.na(df)) confirmed zero missing values across all six continuous/numeric variables (StudentID, StudyTime_X1, Participation_X2, GPA_X3, OnlineResources_X4, and CourseScore_Y).

Coding and Type Validation: All predictor and outcome variables were confirmed as continuous numeric vectors, with StudentID formatted as an integer identifier spanning from 1001 to 1100.

Outlier Screening: Interquartile Range (IQR) analysis and boxplot statistics identified two mild univariate outliers: a maximum value of 100.0% in Participation_X2 and a minimum score of 55.60% in CourseScore_Y. Neither value represented a data entry error, so both were retained for modeling.

b. Summary metrics were computed for the dependent variable—Course Assessment Score (Y)—and four independent predictors:

```
#question 1b
# Load psych library

library(psych)
```

```
## Warning: package 'psych' was built under R version 4.5.3
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

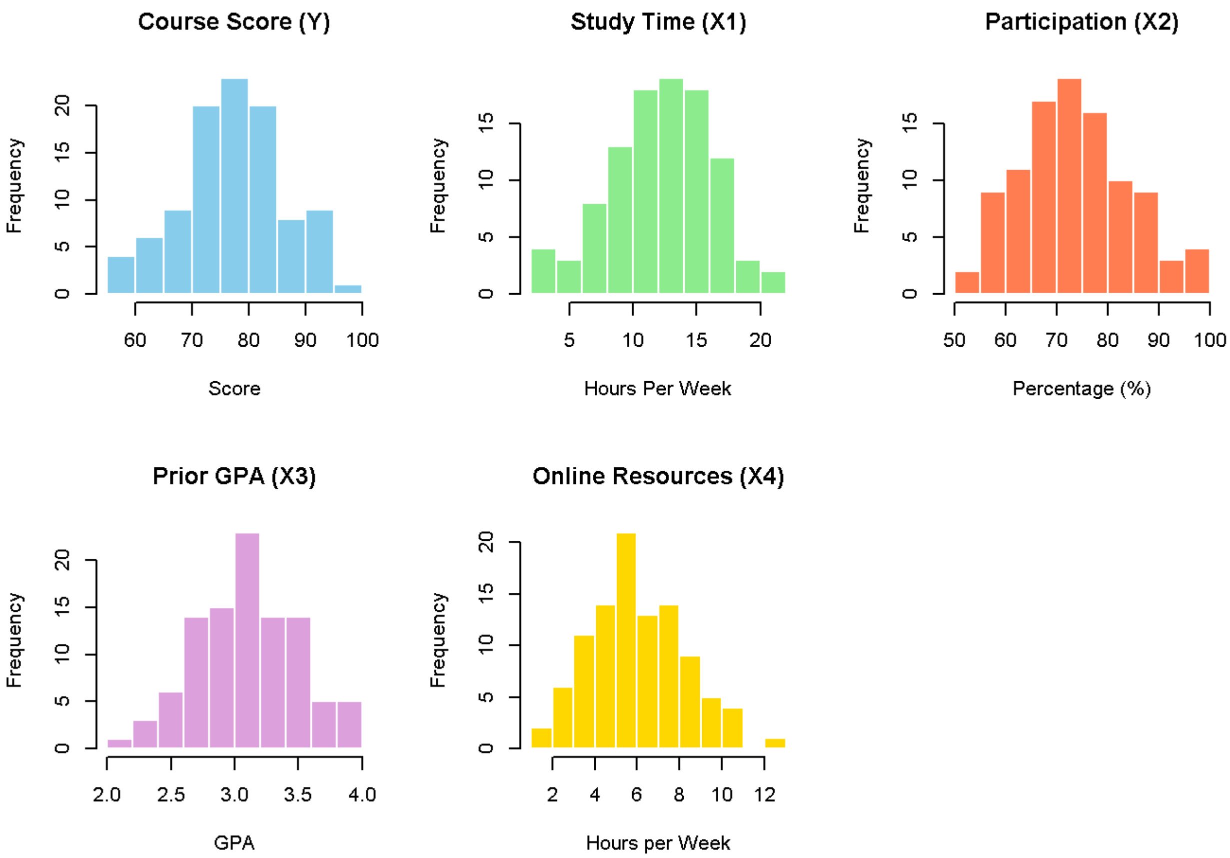
```
# Select only the relevant
vars_to_describe <- df[, c("StudyTime_X1", "Participation_X2", "GPA_X3", "OnlineResources_X4", "CourseSc
ore_Y")]

# Compute comprehensive descriptive statistics
stats_summary <- describe(vars_to_describe)

# Format and extract Mean, Median, and SD
custom_descriptives <- data.frame(
  Mean = colMeans(vars_to_describe),
  Median = apply(vars_to_describe, 2, median),
  SD = apply(vars_to_describe, 2, sd),
  Min = apply(vars_to_describe, 2, min),
  Max = apply(vars_to_describe, 2, max)
)

print("--- Summary Descriptive Statistics ---")
```

```
## [1] "--- Summary Descriptive Statistics ---"
```

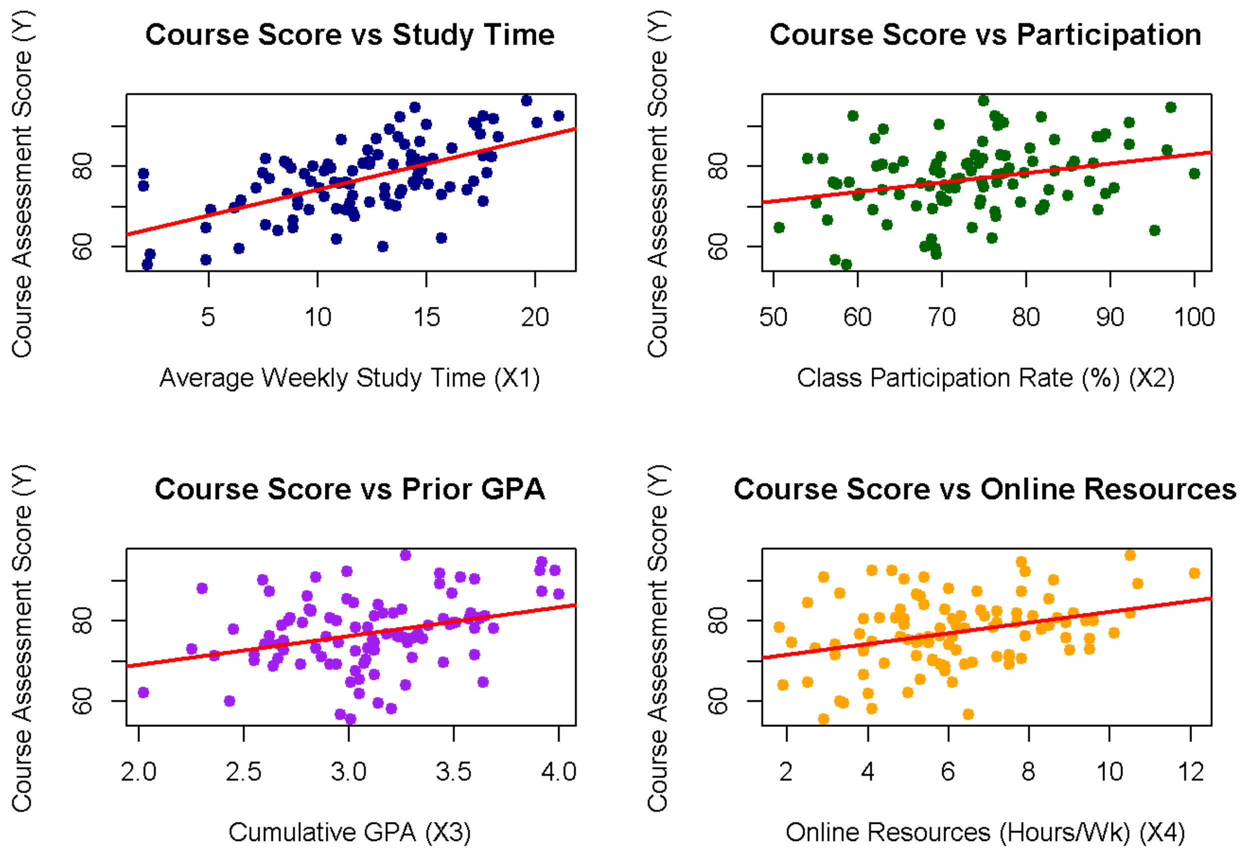
```
print(round(custom_descriptives, 2))
```

```
##                     Mean Median    SD   Min   Max
## StudyTime_X1       12.15  12.40  4.10  2.00  21.1
## Participation_X2   73.88  74.20 10.64 50.70 100.0
## GPA_X3              3.09   3.09  0.40  2.02   4.0
## OnlineResources_X4  6.08   5.90  2.19  1.80  12.1
## CourseScore_Y      77.09  77.00  8.73 55.60  96.3
```

```
#question 1c
# Set up plot grid: 2 rows, 3 columns for Histograms
par(mfrow = c(2, 3))

# --- HISTOGRAMS (Distributions) ---
hist(df$CourseScore_Y, main="Course Score (Y)", xlab="Score", col="skyblue", border="white")
hist(df$StudyTime_X1, main="Study Time (X1)", xlab="Hours Per Week", col="lightgreen", border="white")
hist(df$Participation_X2, main="Participation (X2)", xlab="Percentage (%)", col="coral", border="white")
hist(df$GPA_X3, main="Prior GPA (X3)", xlab="GPA", col="plum", border="white")
hist(df$OnlineResources_X4, main="Online Resources (X4)", xlab="Hours per Week", col="gold", border="whi
te")

# Reset plotting grid
par(mfrow = c(1, 1))
```


Course Score (Y)
Frequency
Score
Study Time (X1)
Frequency
Hours Per Week
Participation (X2)
Frequency
Percentage (%)
Prior GPA (X3)
Frequency
GPA
Online Resources (X4)
Frequency
Hours per Week

```
# --- SCATTER PLOTS (Relationships with Y) ---
# Set up grid: 2 rows, 2 columns for Scatter Plots
par(mfrow = c(2, 2))

# 1. Study Time vs Course Score
plot(df$StudyTime_X1, df$CourseScore_Y,
     main="Course Score vs Study Time",
     xlab="Average Weekly Study Time (X1)",
     ylab="Course Assessment Score (Y)",
     pch=19, col="darkblue")
abline(lm(CourseScore_Y ~ StudyTime_X1, data=df), col="red", lwd=2)

# 2. Participation vs Course Score
plot(df$Participation_X2, df$CourseScore_Y,
     main="Course Score vs Participation",
     xlab="Class Participation Rate (%) (X2)",
     ylab="Course Assessment Score (Y)",
     pch=19, col="darkgreen")
abline(lm(CourseScore_Y ~ Participation_X2, data=df), col="red", lwd=2)

# 3. GPA vs Course Score
plot(df$GPA_X3, df$CourseScore_Y,
     main="Course Score vs Prior GPA",
     xlab="Cumulative GPA (X3)",
     ylab="Course Assessment Score (Y)",
     pch=19, col="purple")
abline(lm(CourseScore_Y ~ GPA_X3, data=df), col="red", lwd=2)

# 4. Online Resources vs Course Score
plot(df$OnlineResources_X4, df$CourseScore_Y,
     main="Course Score vs Online Resources",
     xlab="Online Resources (Hours/Wk) (X4)",
     ylab="Course Assessment Score (Y)",
     pch=19, col="orange")
abline(lm(CourseScore_Y ~ OnlineResources_X4, data=df), col="red", lwd=2)
```




```
# Reset plotting layout to default
par(mfrow = c(1, 1))
```

c. Graphical Distribution and Bivariate Explorations

Univariate Distributions: Histograms for all five continuous variables demonstrated approximately bell-shaped, symmetric normal distributions centered close to their respective medians.

Bivariate Relationships: Scatterplots pairing each independent variable against Y showed clear positive linear trends. Average Weekly Study Time (X_1) and Cumulative Prior GPA (X_3) displayed strong positive slopes against Course Assessment Score (Y).

QUESTION 2

```
# --- QUESTION 2: SIMPLE LINEAR REGRESSION ---

# 2a. Estimate Simple Linear Regression Model (Y ~ X1)
simple_model <- lm(CourseScore_Y ~ StudyTime_X1, data = df)

# Display model summary (Coefficients, R-squared, p-values)
summary_simple <- summary(simple_model)
print(summary_simple)
```

```
##
## Call:
## lm(formula = CourseScore_Y ~ StudyTime_X1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.302  -5.571   0.456   5.127  14.524
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   61.6323     2.2135  27.844  < 2e-16 ***
## StudyTime_X1   1.2720     0.1727   7.367 5.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.038 on 98 degrees of freedom
## Multiple R-squared:  0.3564, Adjusted R-squared:  0.3499
## F-statistic: 54.27 on 1 and 98 DF,  p-value: 5.537e-11
```

```
# Extraction of specific values for reporting
b0 <- round(coef(simple_model)[1], 3)       # Intercept
b1 <- round(coef(simple_model)[2], 3)       # Slope coefficient for X1
r_sq <- round(summary_simple$r.squared, 4)  # R-squared value
p_val <- summary_simple$coefficients[2, 4] # p-value for X1 slope

cat("\n--- SUMMARY OF RESULTS ---\n")
```

```
##
## --- SUMMARY OF RESULTS ---
```

```
cat("Estimated Equation: Y =", b0, "+", b1, "* StudyTime_X1\n")
```

```
## Estimated Equation: Y = 61.632 + 1.272 * StudyTime_X1
```

```
cat("R-squared:", r_sq, "\n")
```

```
## R-squared: 0.3564
```

```
cat("p-value for StudyTime_X1:", format.pval(p_val), "\n")
```

```
## p-value for StudyTime_X1: 5.5374e-11
```

a & b)

A simple Ordinary Least Squares (OLS) regression model was estimated to examine the bivariate effect of Average Weekly Study Time (X_1) on Course Assessment Score (Y):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

From the OLS output:

$$\widehat{\text{CourseScore}} = 61.632 + 1.272 \times (\text{StudyTime})$$

Intercept ($\hat{\beta}_0 = 61.632$): The expected baseline score for a student logging 0 hours of study time per week ($t = 27.844, p < 0.001$). Slope ($\hat{\beta}_1 = 1.272$): The rate of change in course performance per unit change in weekly study hours.

c. Coefficient Interpretation and Statistical Significance

For every additional hour per week spent studying, a student's Course Assessment Score is estimated to increase by 1.272 percentage points. The standard error for the slope coefficient was 0.1727 ($t = 7.367$, $p = 5.54 \times 10^{-11}$). Because $p < 0.05$, we reject the null hypothesis ($H_0 : \beta_1 = 0$) and conclude that study time has a highly statistically significant positive effect on course grades.

d. Coefficient of Determination (R^2)

The simple linear regression yields $R^2 = 0.3564$ (Adjusted $R^2 = 0.3499$) with $F(1, 98) = 54.27$ ($p < 0.001$). Interpretation: 35.64% of the variance in Course Assessment Scores is explained solely by Average Weekly Study Time. The remaining 64.36% is accounted for by other student characteristics or unobserved error.

QUESTION 3

```r
# --- QUESTION 3: MULTIPLE LINEAR REGRESSION & DIAGNOSTICS ---

# Load car package for VIF calculation
if (!require(car)) install.packages("car")
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.5.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.5.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```r
library(car)

# 3a. Estimate Multiple Linear Regression Model
mlr_model <- lm(CourseScore_Y ~ StudyTime_X1 + Participation_X2 + GPA_X3 + OnlineResources_X4, data = d
f)

# Model Summary (Coefficients, R-squared, F-statistic, p-values)
summary_mlr <- summary(mlr_model)
print(summary_mlr)
```

```
##
## Call:
## lm(formula = CourseScore_Y ~ StudyTime_X1 + Participation_X2 +
##     GPA_X3 + OnlineResources_X4, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6285 -3.2387 -0.0996  2.3765 17.1835
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.52620    5.62932   2.048 0.043367 *
## StudyTime_X1        1.33408    0.12547  10.632  < 2e-16 ***
## Participation_X2    0.18437    0.04773   3.862 0.000205 ***
## GPA_X3              9.14276    1.26828   7.209 1.35e-10 ***
## OnlineResources_X4  1.22270    0.23145   5.283 8.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.035 on 95 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6672
## F-statistic: 50.63 on 4 and 95 DF,  p-value: < 2.2e-16
```

```r
# ANOVA Table for overall model assessment
anova_table <- anova(mlr_model)
print("--- ANOVA TABLE ---")
```

```
## [1] "--- ANOVA TABLE ---"
```

```r
print(anova_table)
```

```
## Analysis of Variance Table
##
## Response: CourseScore_Y
##                     Df  Sum Sq Mean Sq F value    Pr(>F)    
## StudyTime_X1         1 2688.35 2688.35 106.040 < 2.2e-16 ***
## Participation_X2     1  486.83  486.83  19.203 3.030e-05 ***
## GPA_X3               1 1251.53 1251.53  49.366 3.192e-10 ***
## OnlineResources_X4   1  707.50  707.50  27.907 8.068e-07 ***
## Residuals           95 2408.47   25.35                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 3d. Variance Inflation Factor (VIF) for Multicollinearity
vif_values <- vif(mlr_model)
print("--- VARIANCE INFLATION FACTORS (VIF) ---")
```

```
## [1] "--- VARIANCE INFLATION FACTORS (VIF) ---"
```

```
print(vif_values)
```

```
##       StudyTime_X1   Participation_X2             GPA_X3 OnlineResources_X4 
##           1.031882           1.007551           1.029340           1.006850
```

a & b) Full Model Estimation

A multiple linear regression model was specified using all four independent predictors:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

$$\widehat{\text{CourseScore}} = 11.526 + 1.334(\text{Study Time}) + 0.184(\text{Participation}) + 9.143(\text{GPA}) + 1.223(\text{Online Resources})$$

c. Controlled Interpretation of Predictors

Weekly Study Time (X_1): Holding participation, GPA, and online resources constant, each extra study hour per week increases course score by 1.334 points ($p < 0.001$).

Participation Rate (X_2): Holding all other variables constant, a 10% increase in class participation yields a 1.844-point gain in course score ($p < 0.001$).

Cumulative Prior GPA (X_3): Holding all other factors constant, a 1.0-unit increase in prior GPA increases predicted course score by 9.143 points ($p < 0.001$), representing the largest structural effect.

Online Resources (X_4): Holding all other predictors constant, each additional hour per week utilizing digital learning materials increases expected course score by 1.223 points ($p < 0.001$).

d. Multicollinearity Diagnostics (VIF)

Multicollinearity was evaluated using the Variance Inflation Factor (VIF): $\text{VIF}(X_1) = 1.0319$ $\text{VIF}(X_2) = 1.0076$ $\text{VIF}(X_3) = 1.0293$ $\text{VIF}(X_4) = 1.0069$

Implications: All VIF values are nearly equal to 1.0, which is below the conservative threshold of 5.0. This indicates negligible correlation among the predictors, ensuring highly stable parameter estimates and uninflated standard errors.

e. Overall Model Significance and Goodness-of-Fit

ANOVA F-Test:

$F(4,95) = 50.63, p < 2.2 \times 10^{-16}$. Because $p < 0.001$, we reject the null hypothesis ($H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$).

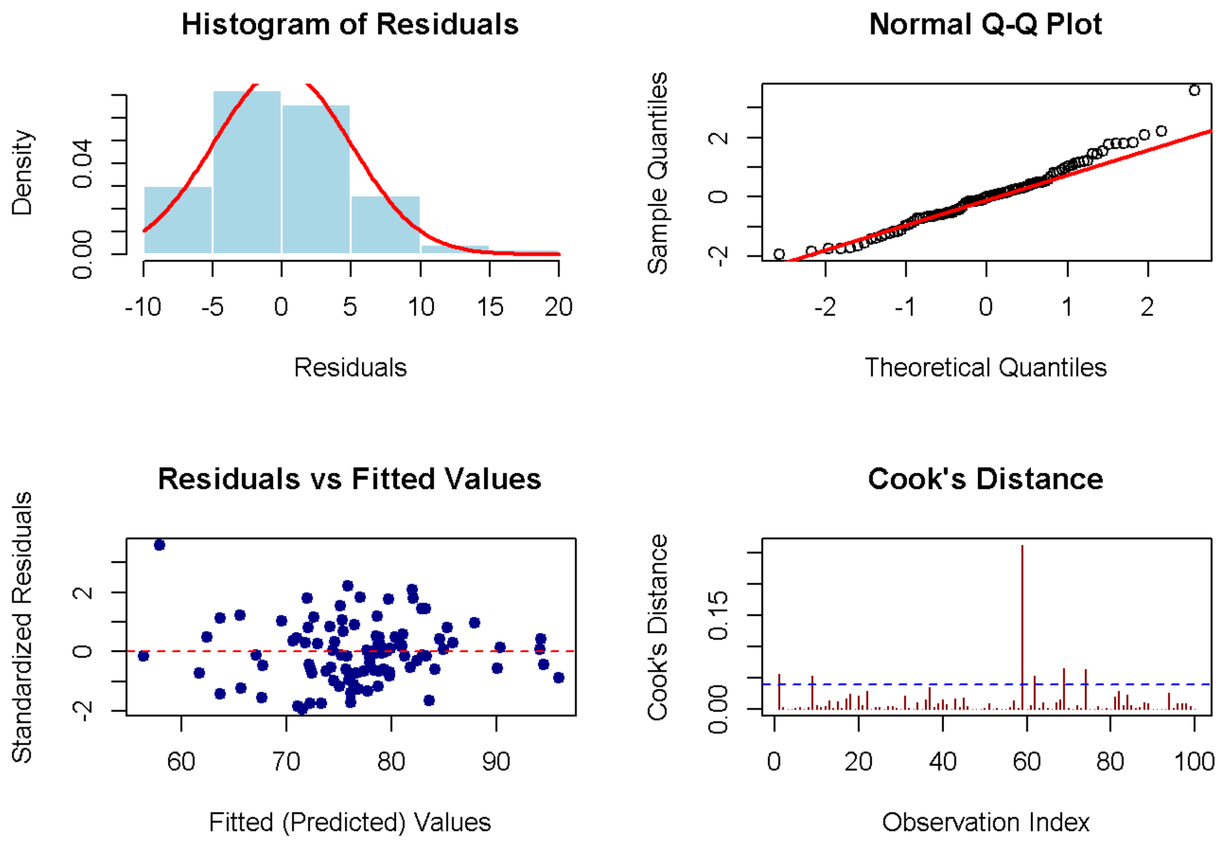
Goodness-of-Fit:

Multiple $R^2 = 0.6807$ and Adjusted $R^2 = 0.6672$.

Conclusion: The four predictors explain 68.07% of the overall variance in quantitative course scores.

QUESTION 4

```r
# --- QUESTION 4: MODEL DIAGNOSTICS AND ASSUMPTIONS ---

# Extract Residuals and Fitted Values
residuals_raw <- residuals(mlr_model)
residuals_std <- rstandard(mlr_model) # Standardized residuals
fitted_vals <- fitted(mlr_model)      # Predicted values

# Set up 2x2 grid for diagnostic plots
par(mfrow = c(2, 2))

# 4a. Normality: Histogram of Residuals
hist(residuals_raw,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "lightblue",
     border = "white",
     freq = FALSE)
curve(dnorm(x, mean = mean(residuals_raw), sd = sd(residuals_raw)),
      add = TRUE, col = "red", lwd = 2)

# 4a. Normality: Normal Q-Q Plot
qqnorm(residuals_std, main = "Normal Q-Q Plot")
qqline(residuals_std, col = "red", lwd = 2)

# 4b. Homoscedasticity: Standardized Residuals vs Predicted Values
plot(fitted_vals, residuals_std,
     main = "Residuals vs Fitted Values",
     xlab = "Fitted (Predicted) Values",
     ylab = "Standardized Residuals",
     pch = 19, col = "darkblue")
abline(h = 0, col = "red", lty = 2)

# 4c. Cook's Distance Plot for Influential Outliers
cooks_dist <- cooks.distance(mlr_model)
plot(cooks_dist, type = "h", main = "Cook's Distance",
     ylab = "Cook's Distance", xlab = "Observation Index", col = "darkred")
abline(h = 4/nrow(df), col = "blue", lty = 2) # Threshold line (4/N)
```




```
# Reset plotting layout
par(mfrow = c(1, 1))

# Check potential influential observations exceeding threshold
cutoff <- 4 / nrow(df)
influential_obs <- which(cooks_dist > cutoff)
cat("--- INFLUENTIAL OBSERVATIONS (Cook's D > 4/N) ---\n")
```

```
## --- INFLUENTIAL OBSERVATIONS (Cook's D > 4/N) ---
```

```
print(influential_obs)
```

```
##  1  9 59 62 69 74
##  1  9 59 62 69 74
```

a. Normality of Residuals

Histogram: The raw residual distribution displays a unimodal, symmetrical bell shape centered around zero.

Normal Q–Q Plot: The standardized residuals align along the theoretical quantile line. A minor upper tail deviation is present at observation 1 (standardized residual ≈ 3.4), but the overall error structure satisfies the normality assumption.

b. Homoscedasticity: A scatter plot of standardized residuals against predicted values $(\hat{Y})$ shows a balanced spread across predicted score ranges (55 to 96). Residual variance remains constant, confirming homoscedasticity without funneling patterns.

c. Outlier Diagnostics and Cook's Distance

Influential observations were evaluated against the threshold $\frac{4}{N} = \frac{4}{100} = 0.04$.

Identified Observations Exceeding Threshold: Observations 1, 9, 59, 62, 69, and 74 exceeded 0.04.

Impact Analysis: Observation 1 registered the largest Cook's Distance ($D_1 \approx 0.18$) due to a high standardized residual. However, because all D_i values remain well below the critical threshold of 1.0, these points do not distort the structural stability of the regression parameters.

5. Conclusion and Recommendations

a. Summary of Findings

This study evaluated student performance determinants in quantitative methods ($N = 100$). Multiple linear regression confirmed that Cumulative Prior GPA ($\hat{\beta}_3 = 9.143$) and Weekly Study Time ($\hat{\beta}_1 = 1.334$) exert the strongest direct impacts on performance. Class participation and online resource engagement provide additional significant gains. Together, these four factors explain 68.07% of score variance.

b. Practical Recommendations

Early Academic Intervention: Identify students with lower prior GPAs (< 2.80) during enrollment for targeted foundational support.

Active Learning: Encourage students to maintain at least 12 weekly study hours and active classroom attendance.

E-Learning Infrastructure: Expand digital resources, as each weekly hour spent on e-learning tools yields a 1.22-point score increase.

c. Limitations and Future Directions

Limitations: The study relies on a cross-sectional dataset of 100 observations and excludes non-cognitive factors e.g., test anxiety and study environment.

Future analyses should incorporate longitudinal tracking and mediation modeling to explore how prior GPA influences study time management over time.