RSM8512_Assignment2_1006759189_Linear_Regression

Question 1

The p-values in Table 3.4 correspond to hypothesis tests for each predictor (TV, radio, and newspaper) in a multiple linear regression model that predicts sales. In this context, the null hypotheses for each predictor are as follows:

Null Hypothesis (H_0) for TV: There is no relationship between TV advertising and sales (i.e., the coefficient for TV is zero). Null Hypothesis (H_0) for radio: There is no relationship between radio advertising and sales (i.e., the coefficient for radio is zero). Null Hypothesis (H_0) for newspaper: There is no relationship between newspaper advertising and sales (i.e., the coefficient for newspaper is zero). The corresponding p-values tell us whether we can reject these null hypotheses at a certain significance level (commonly $\alpha = 0.05$).

Interpretation of p-values:

TV p-value (< 0.0001): The p-value is extremely small, far below 0.05. This suggests that we can strongly reject the null hypothesis for TV. There is a statistically significant relationship between TV advertising and sales.

Radio p-value (< 0.0001): Like TV, the p-value for radio is also very small, indicating that we can reject the null hypothesis. There is a statistically significant relationship between radio advertising and sales.

Newspaper p-value (0.8599): The p-value for newspaper is very high (much higher than 0.05). Therefore, we fail to reject the null hypothesis for newspaper. This suggests that there is no statistically significant relationship between newspaper advertising and sales.

Conclusion:

Based on the p-values, we can conclude that TV and radio advertising have significant effects on sales, while newspaper advertising does not appear to have a significant effect. In practical terms, this suggests that spending on TV and radio advertising is likely to increase sales, whereas spending on newspaper advertising might not have a meaningful impact on sales.

Question 2

Part (a): Which answer is correct, and why?

The correct answer is **iii**: "For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough."

This is because the coefficient for the interaction between GPA and Level ($\beta_5 = -10$) reduces the benefit that college graduates receive from the positive coefficient for Level ($\beta_3 = 35$) as GPA increases. When GPA is high enough, the negative interaction outweighs the positive advantage of being a college graduate, meaning high school graduates can earn more.

Part (b): Predict the salary of a college graduate with an IQ of 110 and GPA of 4.0

We use the following linear model to predict the salary:

$$\hat{Salary} = \beta_0 + \beta_1 \cdot GPA + \beta_2 \cdot IQ + \beta_3 \cdot Level + \beta_4 \cdot (GPA \cdot IQ) + \beta_5 \cdot (GPA \cdot Level)$$

Where: - $\beta_0 = 50$ - $\beta_1 = 20$ - $\beta_2 = 0.07$ - $\beta_3 = 35$ - $\beta_4 = 0.01$ - $\beta_5 = -10$

Let's calculate the predicted salary for a college graduate ($Level = 1$) with an IQ of 110 and GPA of 4.0.

```
beta_0 <- 50
beta_1 <- 20
beta_2 <- 0.07
beta_3 <- 35
beta_4 <- 0.01
beta_5 <- -10


GPA <- 4.0
IQ <- 110
Level <- 1


predicted_salary <- beta_0 + beta_1 * GPA + beta_2 * IQ + beta_3 * Level +
  beta_4 * (GPA * IQ) + beta_5 * (GPA * Level)

predicted_salary
```

```
## [1] 137.1
```

c)This statement is false. The size of the coefficient β_4=0.01 is indeed small, but this does not automatically imply there is no significant interaction effect. The statistical significance of the interaction term would depend on its p-value. Even a small coefficient can have a statistically significant effect, depending on the variability and the scale of the data. Therefore, we cannot conclude there is little evidence of interaction without considering the p-value.

Question 3

Simple Linear Regression Model

In simple linear regression, the equation of the least squares line is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where: - $\hat{\beta}_1$ is the slope of the regression line. - $\hat{\beta}_1$ is the intercept of the regression line.

The formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$ are:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof that the Line Passes Through $(\bar{x}, \bar{y})$

We will show that the least squares regression line passes through the point $(\bar{x}, \bar{y})$.

Step 1: Substitute $\bar{x}$ into the Regression Equation

Substitute $\bar{x}$ into the equation of the regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Step 2: Substitute the Expression for $\hat{\beta}_0$

Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, substitute this expression into the equation:

$$\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x}$$

Step 3: Simplify the Equation

Simplify the equation by canceling out the terms $-\hat{\beta}_1 \bar{x}$ and $+\hat{\beta}_1 \bar{x}$:

$$\hat{y} = \bar{y}$$

Conclusion

Thus, when $x = \bar{x}$, we have $\hat{y} = \bar{y}$. This shows that the least squares regression line always passes through the point $(\bar{x}, \bar{y})$, meaning the regression line goes through the means of both the predictor $(\bar{x})$ and the response variable $(\bar{y})$.

Example in R

Let's demonstrate this with a simple linear regression example using R.

```
set.seed(123)
x <- rnorm(100)
y <- 3 + 2 * x + rnorm(100)


model <- lm(y ~ x)


mean_x <- mean(x)
mean_y <- mean(y)


predicted_y_at_mean_x <- predict(model, newdata = data.frame(x = mean_x))


cat("Mean of x:", mean_x, "\n")
```

```
## Mean of x: 0.09040591
```

```
cat("Mean of y:", mean_y, "\n")
```

```
## Mean of y: 3.073265
```

```
cat("Predicted y at mean(x):", predicted_y_at_mean_x, "\n")
```

```
## Predicted y at mean(x): 3.073265
```

```
all.equal(mean_y, predicted_y_at_mean_x)
```

```
## [1] "names for current but not for target"
```

Question 4)

```
#Question4a)

auto_data <- read.table("C:/Users/dokan/Downloads/assignment2_r/Auto.data", header=TRUE, na.strings="?"
```

```r
head(auto_data)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```r
model <- lm(mpg ~ horsepower, data = auto_data)
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = auto_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
predict(model, newdata = data.frame(horsepower = 98), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```r
predict(model, newdata = data.frame(horsepower = 98), interval = "prediction")
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

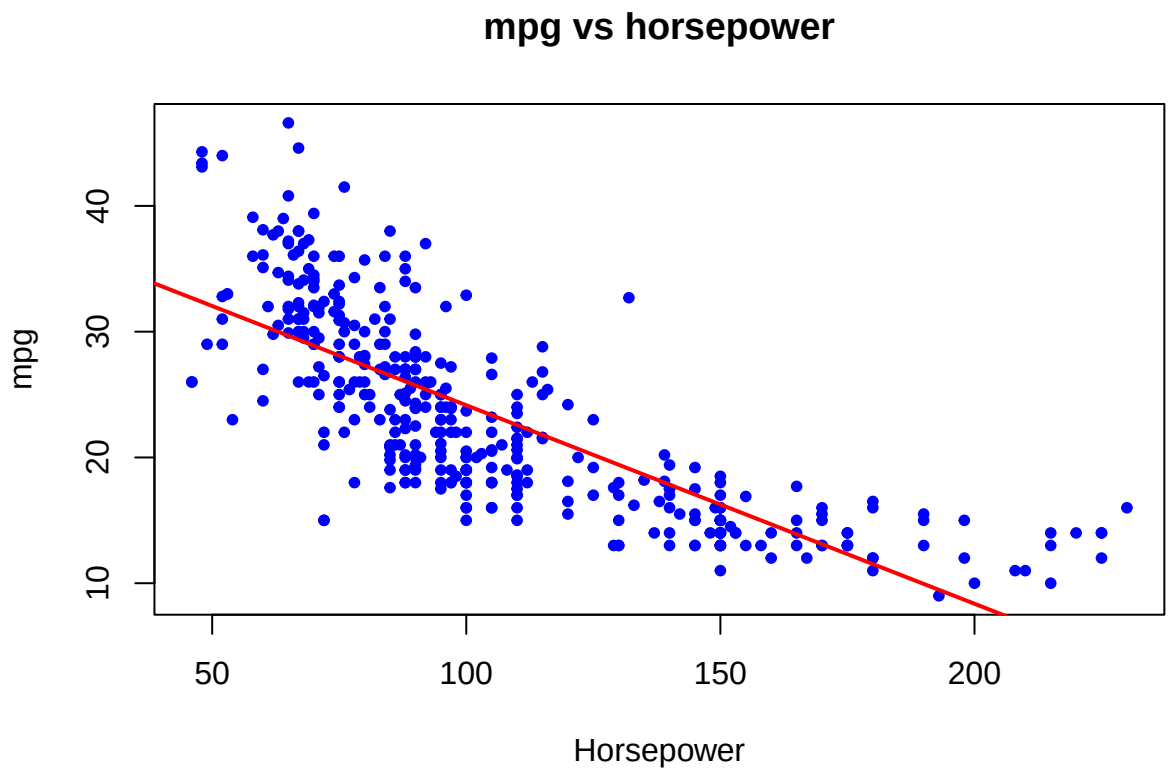
Explanation:

- **i. Relationship:** The `p-value` for the `horsepower` coefficient tells us if there is a statistically significant relationship between the predictor (horsepower) and the response (mpg).
- **ii. Strength:** The R^2 value shows how much of the variation in `mpg` is explained by `horsepower`. A higher value indicates a stronger relationship.
- **iii. Direction of the relationship:** The coefficient sign (positive or negative) for `horsepower` shows whether the relationship between `mpg` and `horsepower` is positive or negative.
- **iv. Predicted `mpg` at `horsepower = 98`:** We use the `predict()` function to obtain the predicted value and the 95% confidence and prediction intervals.

This R Notebook will produce the outputs for each question based on the data provided.

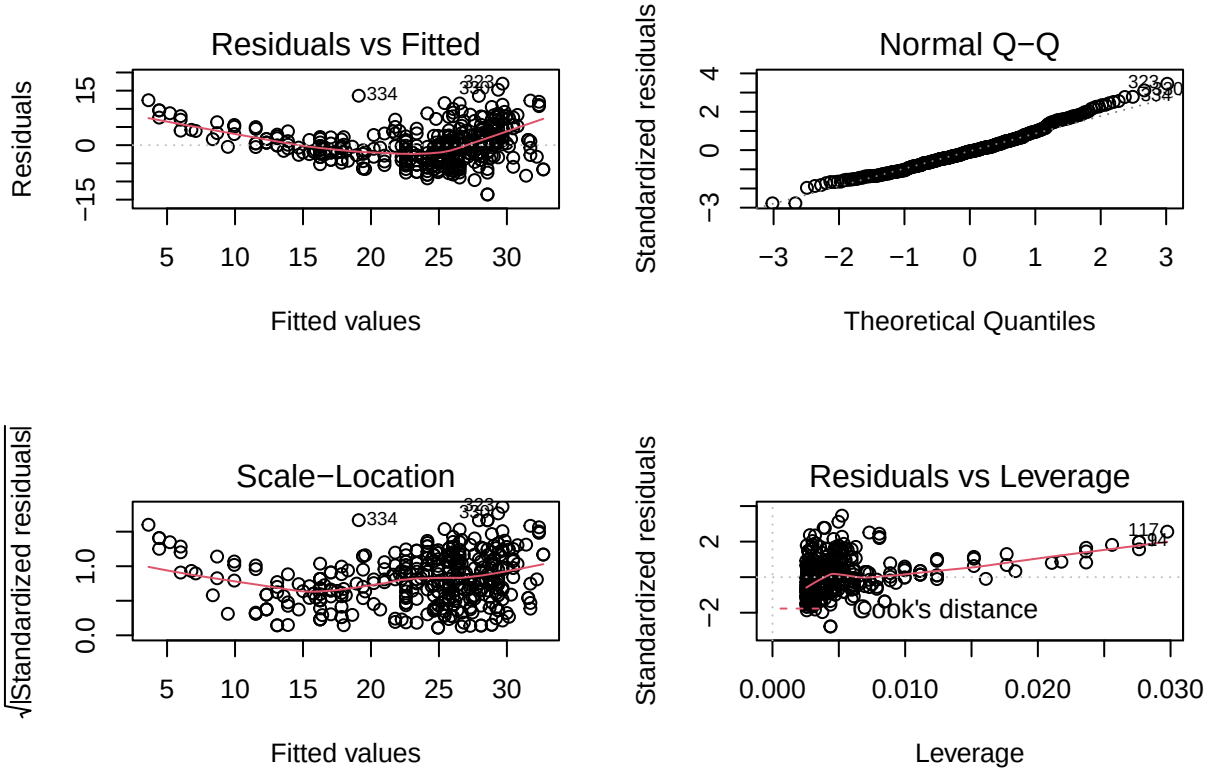
```
#Question8)b

plot(auto_data$horsepower, auto_data$mpg, main="mpg vs horsepower",
     xlab="Horsepower", ylab="mpg", pch=20, col="blue")
abline(model, col="red", lwd=2)
```




```
#Question 8)c

par(mfrow = c(2, 2))
plot(model)
```



Comments on Diagnostic Plots

Residuals vs Fitted: Random scatter indicates a good fit. A pattern suggests non-linearity. Normal Q-Q: Points on the line indicate normal residuals. Deviations suggest non-normality. Scale-Location: A horizontal line indicates constant variance. A funnel shape suggests heteroscedasticity. Residuals vs Leverage: High-leverage points with large residuals may influence the model excessively.

Question 11)

Question 11 a) Part (a): Simple Linear Regression of y onto x (No Intercept)

```
set.seed(1)
x <- rnorm(100)
y <- 2 * x + rnorm(100)


model_a <- lm(y ~ x + 0)


summary(model_a)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
```

```
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

#Interpretation:

The coefficient for x is statistically significant (small p-value), meaning x significantly explains the variation in y. The positive coefficient suggests that y increases with x, confirming the expected positive relationship.

```
#Question 11 b)

model_b <- lm(x ~ y + 0)


summary(model_b)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

#Interpretation:

The regression of x onto y shows similar results to the first model. The coefficient is also statistically significant, confirming the reciprocal relationship between x and y. This symmetry is reflected in similar t-statistics and p-values.

#Question 11 c)

The t-statistics from both regressions are symmetric, confirming that the linear relationship between x and y is reciprocal. Whether we regress y on x or x on y, the strength of the relationship remains the same.

```
#Question 11 d)
n <- length(x)


beta_hat <- sum(x * y) / sum(x^2)
```

```r
residuals <- y - beta_hat * x


SE_beta <- sqrt(sum(residuals^2) / (n - 1)) / sqrt(sum(x^2))


t_stat_manual <- beta_hat / SE_beta
t_stat_manual
```

```
## [1] 18.72593
```

```r
t_stat_lm <- summary(model_a)$coefficients[1, "t value"]
t_stat_lm
```

```
## [1] 18.72593
```

#Interpretation:

The manually calculated t-statistic matches the t-statistic from the model output in part (a). This consistency validates the model's accuracy and confirms the statistical significance of the relationship between y and x.

Question 11 e)

As expected, the t-statistics for both regressions are identical. This reaffirms the symmetry in the relationship between y and x, showing that the strength of the relationship remains consistent in both directions.

```r
#Question 11 f)

model_with_intercept_a <- lm(y ~ x)
summary(model_with_intercept_a)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
model_with_intercept_b <- lm(x ~ y)
summary(model_with_intercept_b)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

#Interpretation:

Including the intercept in the regression does not change the t-statistic for the slope, confirming that the strength of the relationship between y and x is unaffected by the intercept. The intercept only shifts the regression line vertically, but the relationship remains equally significant.

Question 14)

Question 14 a)

```
set.seed(1)


x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100) / 10
y <- 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

#Interpretation:

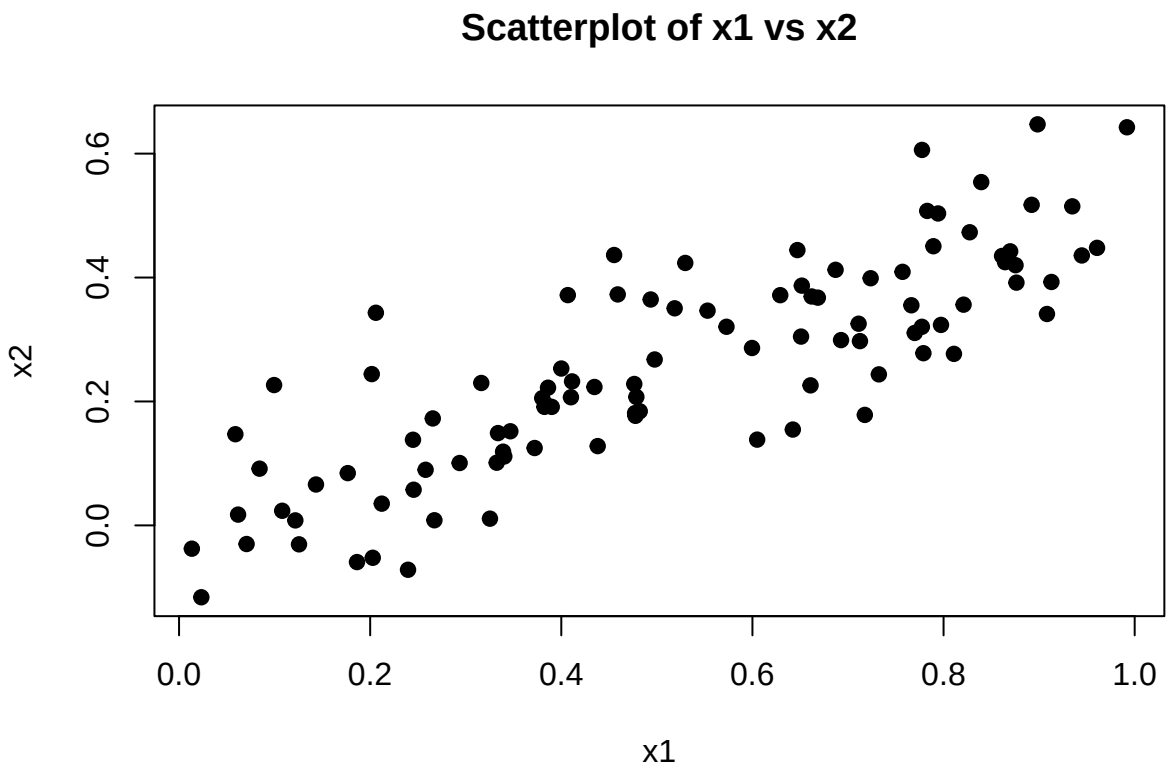
The generated data represents a model where both x1 and x2 influence y. The correlation between x1 and x2 is expected to introduce multicollinearity, which may affect the regression results.

```
#Question 14 b)

correlation <- cor(x1, x2)
correlation
```

```
## [1] 0.8351212
```

```
plot(x1, x2, main = "Scatterplot of x1 vs x2", xlab = "x1", ylab = "x2", pch = 19)
```



#Interpretation:

The correlation between x1 and x2 is moderately high, indicating multicollinearity. This suggests that it may be difficult to separate the individual effects of x1 and x2 on y, and it may inflate the standard errors of the coefficients in the regression model.

```
#Question 14 c)

model_c <- lm(y ~ x1 + x2)

summary(model_c)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

#Interpretation:

The coefficients for x1 and x2 are both significant, closely aligning with the true values of 2 and 0.3. Multicollinearity may inflate the standard errors slightly, but both predictors significantly contribute to explaining the variation in y.

```
#Question 14 d)

model_d <- lm(y ~ x1)


summary(model_d)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

#Interpretation:

When regressing y on x1 alone, the coefficient for x1 is overestimated compared to the full model, as the effect of x2 is not accounted for. Despite this bias, x1 remains significant, though the estimate is less reliable due to the omitted variable.

```
#Question 14 e)

model_e <- lm(y ~ x2)


summary(model_e)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

#Interpretation:

The regression of y on x2 alone underestimates the effect of x2 compared to the full model. This is due to the omission of x1, which shares some explanatory power with x2. Nevertheless, x2 remains a significant predictor, although the model is missing important information from x1.

Question 14 f)

In part (c), both x1 and x2 were included in the model. The multicollinearity between them affected the standard errors of their coefficients, potentially leading to less precise estimates. In parts (d) and (e), we regressed y onto x1 and x2 individually, which avoids the multicollinearity problem. However, those models might have biased estimates since they ignore the effect of the other variable.

In summary, multicollinearity in the full model causes inflated standard errors, making it harder to detect the significance of individual predictors. In the simpler models, the exclusion of one variable leads to biased coefficients, overestimating or underestimating the true effects of x1 and x2 on y.

```r
#Question 14 g)

x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)


model_c_new <- lm(y ~ x1 + x2)
summary(model_c_new)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
```

```
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

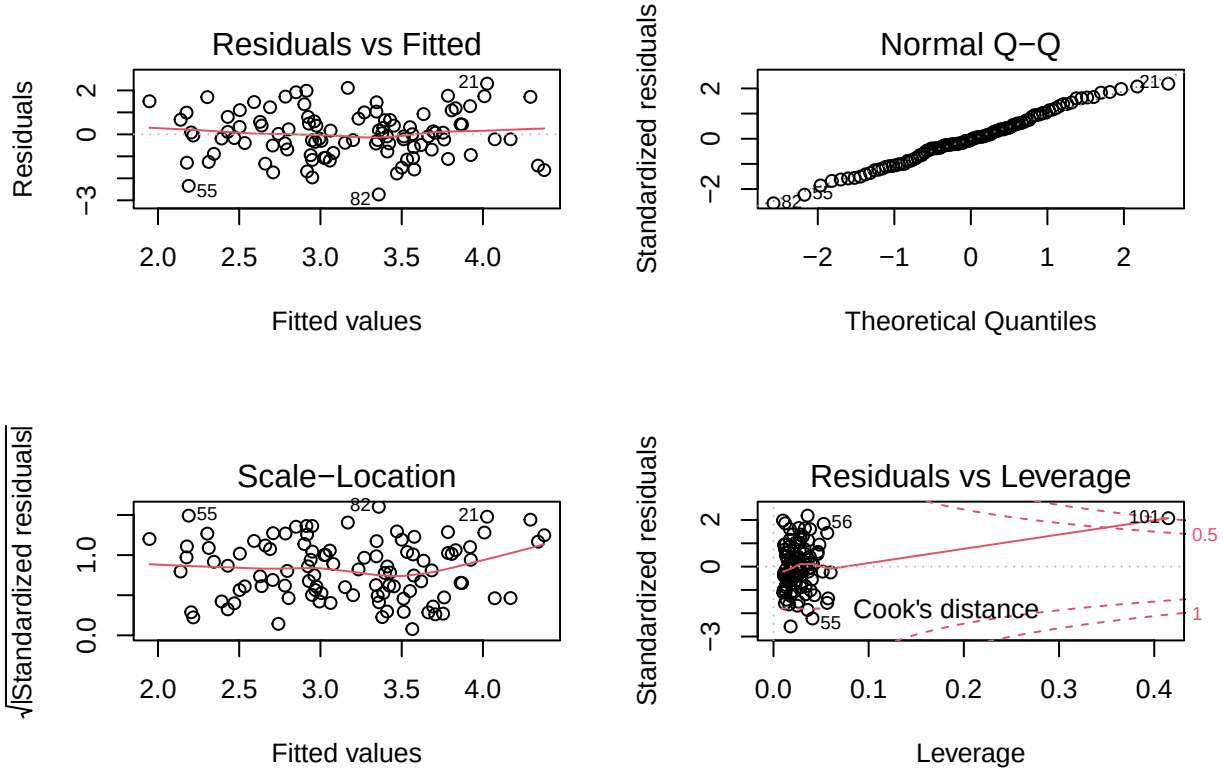
```
model_d_new <- lm(y ~ x1)
summary(model_d_new)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

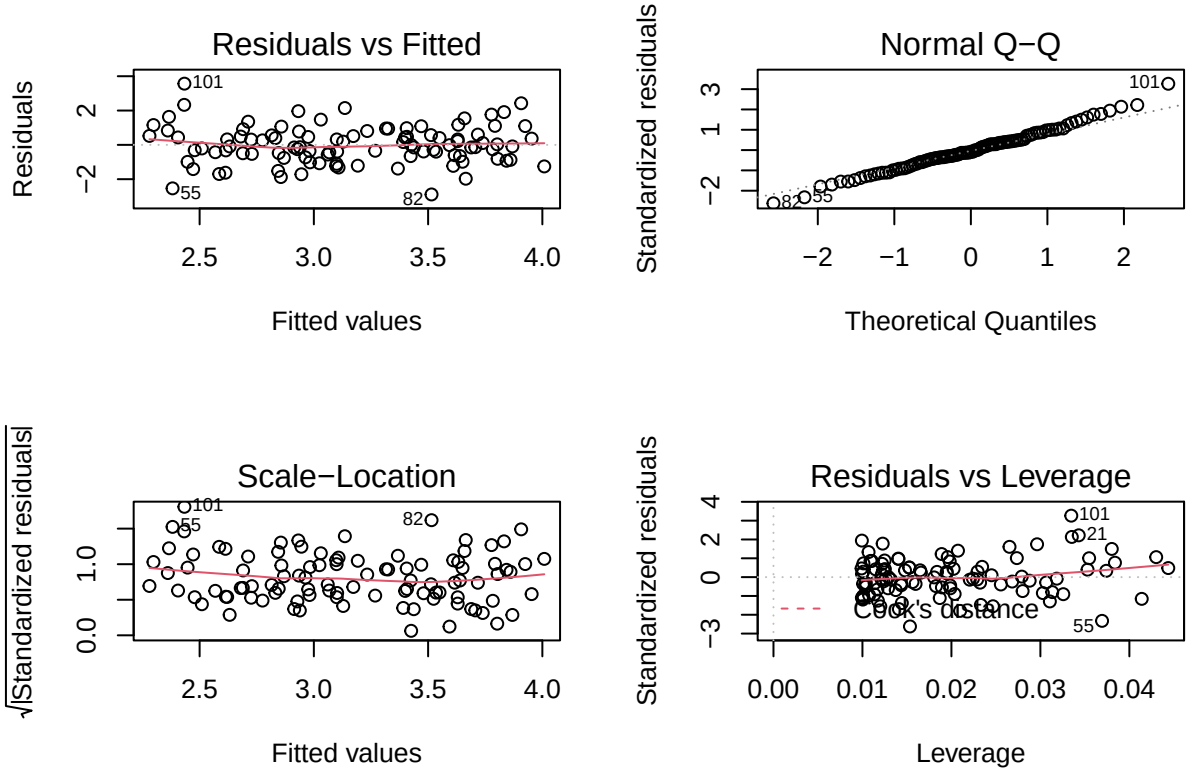
```
model_e_new <- lm(y ~ x2)
summary(model_e_new)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

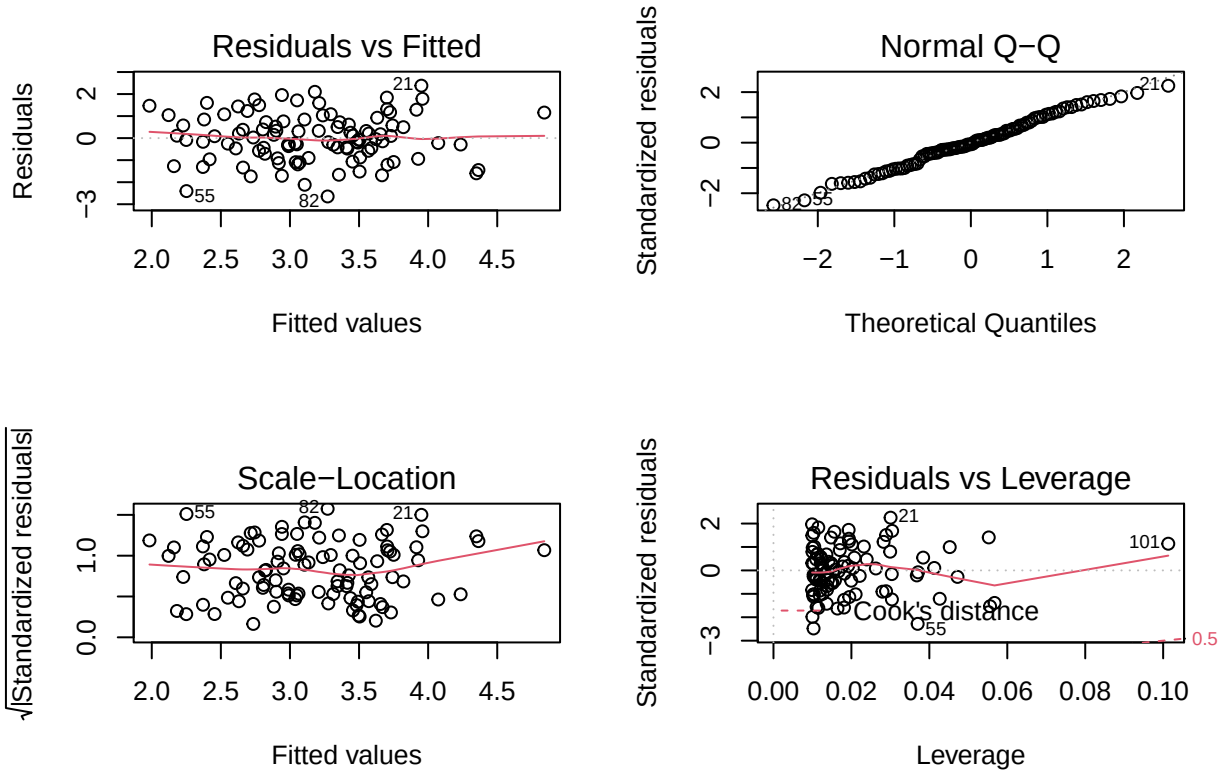
```
par(mfrow = c(2, 2))
plot(model_c_new)
```




```
par(mfrow = c(2, 2))
plot(model_d_new)
```




```
par(mfrow = c(2, 2))
plot(model_e_new)
```



#Interpretation:

The addition of the new data point has a noticeable impact on the models, especially in terms of the coefficients and their significance. The diagnostic plots suggest that this new observation is likely an outlier or a high-leverage point, meaning it has a disproportionate influence on the regression results. This underscores the sensitivity of the models to such points, which may distort the true relationship between the predictors and y.

Question 14 Bonus)

If we were to simulate this data many times, the estimated coefficients would, on average, converge to the true values of $\beta_0 = 2$, $\beta_1 = 2$, and $\beta_2 = 0.3$. This is due to the fact that linear regression is an unbiased estimator when its assumptions hold. Over multiple simulations, the fluctuations in the estimates should average out to the true underlying coefficients.

Additionally, the presence of noise in each simulation will cause some variability in the coefficient estimates, but as the number of simulations increases, the estimates' distribution should tighten around the true values. This convergence reflects the law of large numbers in statistics, which assures that the sample estimates approach the population parameters as the sample size increases.