MAT3375_Assignment-2

Rahul Atre

2023-10-05

Exercise 3.1

Consider the National Football League data in Table B.1

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).

First, we must import and load the data from table.b1 using an R package containing this textbook data sets. We can do this by running the following command on the console

```
install.packages("MPV", repos = "http://cran.us.r-project.org")
library(MPV)
```

Now, let x_2 rep. the team's passing yardage, x_7 rep. the percentage of rushing plays, x_8 rep. the opponents' yards rushing and y rep. the number of games won by a particular team.

```r
data(table.b1)
x_2 = table.b1$x2
x_7 = table.b1$x7
x_8 = table.b1$x8
y = table.b1$y
```

Recall that a regression model that involves more than one regressor variable is called a multiple regression model. It is given by $Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_i k + \epsilon_i$. In this case, we will have 4 parameters since there are 3 regressor variables.

We can calculate the parameters b_0 b_1 b_2$ b_3 using the lm() function to fit the multiple linear model:

```r
full_model = lm(y ~ x_2 + x_7 + x_8)
full_model
```

```
##
## Call:
## lm(formula = y ~ x_2 + x_7 + x_8)
##
## Coefficients:
## (Intercept)          x_2          x_7          x_8
##   -1.808372     0.003598     0.193960    -0.004815
```

From the above function call, we obtain b_0 = -1.8084, b_1 = 0.0036, b_2 = 0.1940, b_3 = -0.0048 (rounded to 4 decimal places).

Therefore, the multiple linear regression model will be $y = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8$.

b. Construct the analysis-of-variance table and test for significance of regression.

Ans: The analysis-of-variance (ANOVA) table can be generated using R:

```
anova(full_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x_2        1  76.193  76.193  26.172 3.100e-05 ***
## x_7        1 139.501 139.501  47.918 3.698e-07 ***
## x_8        1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the ANOVA table, we can see that the p-value for x_2, x_7, and x_8 are all very close to 0. This indicates that all regressor variables are **statistically significant**.

c. Calculate t statistics for testing the hypothesis $H_0 : \beta_2 = 0, H_0 : \beta_7 = 0, and H_0 : \beta_8 = 0$. What conclusions can you draw out about the roles the variables x_2, x_7, and x_8 play in the model?

To calculate the t-statistics for the hypothesis, we can generate the summary table:

```
summary(full_model)
```

```
##
## Call:
## lm(formula = y ~ x_2 + x_7 + x_8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x_2          0.003598   0.000695   5.177 2.66e-05 ***
## x_7          0.193960   0.088233   2.198 0.037815 *
## x_8         -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

As stated, the null hypothesis for each is that the beta parameters are equal to 0. If we look at the table, we can see that the t-value for β_2 is 5.177, β_7 is 2.198, and β_8 is -3.771 which are quite small. The p-value for all three parameters are less than 0.05, which is very small. Therefore, we can reject H_0 and conclude that the parameters are **significant** and not equal to 0. In larger context, this means that there is a correlation between the number of games won (y) and the team's passing yardage (x_2), percentage of rushing plays (x_7), opponent's yard rushing (x_8).

d. Calculate R^2 and R^2_{Adj} for this model.

From the above summary table we have generated, $R^2 = 0.7863 = 78.63\%$, and $R^2_{Adj} = 0.7596 = 75.96\%$.

e. Using the partial F test, determine the contribution of x_7 to the model. How is this partial F statistic related to the t test for β_7 calculated in part c above?

We can perform a partial F test by comparing the full model and reduced model using anova(). The full model is shown above, and the reduced model contains all predictor variables excluding x_7.

```
reduced_model = lm(y ~ x_2 + x_8) #Reduced model that includes everything but x_7
anova(full_model, reduced_model)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x_2 + x_7 + x_8
## Model 2: y ~ x_2 + x_8
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     24 69.870
## 2     25 83.938 -1   -14.068 4.8324 0.03782 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above function call, we obtain an F-statistic of 4.8324. This value is the square of the t-statistic of x_7 in part c.

Exercise 3.3

Refer to Problem 3.1.

a. Find a 95% CI on β_7.

Ans: Since we want the CI on the slope β_7, we can use the formula $b_7 \pm t_{1-\alpha/2,n-2} * s(b_7)$. We know that $n = 28$, $b_7 = 0.193960$, $\alpha = 0.95$. We can generate the CI using R:

```
confint(full_model, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) -18.114944410 14.498200293
## x_2           0.002163664  0.005032477
## x_7           0.011855322  0.376065098
## x_8          -0.007451027 -0.002179961
```

Therefore, from the above function call, we are 95% confident that β_7 is in the interval [0.0119, 0.3761].

b. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300, x_7 = 56.0, x_8 = 2100$.

Ans: In order to find this, we need to set the predictor variables to the given values above to get a 95% confidence interval, and use the point estimate $\hat{Y} = b_0 + b_1 x_2 + \beta_2 x_7 + \beta_3 x_8$. So, the CI is $\hat{Y} \pm t_{1-\alpha/2, n-2} * s(\hat{Y})$. Generating the CI using R:

```
new.dat = data.frame(x_2 = 2300, x_7 = 56.0, x_8 = 2100)
predict(full_model, newdata = new.dat, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 7.216424 6.436203 7.996645
```

Therefore, a 95% CI on the mean number of games won by team given the above constraints is [6.4362, 7.9966].

Exercise 3.4

Reconsider the National Football League data from Problem 3.1. Fit a model to these data using only x_7 and x_8 as the regressors.

- We can create a reduced model that will only contain x_7 and x_8 as the regressors by applying the same lm() function:

```
reduced_model = lm(y ~ x_7 + x_8) #Excluding x_2
reduced_model
```

```
##
## Call:
## lm(formula = y ~ x_7 + x_8)
##
## Coefficients:
## (Intercept)          x_7          x_8
##   17.944319     0.048371    -0.006537
```

The reduced model is $\hat{Y} = y = 17.9443 + 0.0484 x_7 - 0.0065 x_8$

a. Test for significance of regression.

The analysis-of-variance (ANOVA) will generate the following output for significance of regression:

```
summary(reduced_model)
```

```
##
## Call:
## lm(formula = y ~ x_7 + x_8)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.7985 -1.5166 -0.5792  1.9927  4.5248
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.944319   9.862484   1.819  0.08084 .
## x_7          0.048371   0.119219   0.406  0.68839
## x_8         -0.006537   0.001758  -3.719  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.432 on 25 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.5115
## F-statistic: 15.13 on 2 and 25 DF,  p-value: 4.935e-05
```

From the above function call, we can see that the F-statistic for the reduced model is 15.13 which is smaller than 30, and a p-value which is very close to 0. This means that a model with only x_7 and x_8 is **statistically significant**.

b. Calculate R^2 and R^2_{Adj}. How to these quantities compare to the values computed for the model in Problem 3.1, which included an additional regressor (x_2)?

From the above summary table we have generated, $R^2 = 0.5477 = 54.77\%$, and $R^2_{Adj} = 0.5115 = 51.15\%$. These quantities are **close to 20% less** than the model in Problem 3.1, which included x_2.

c. Calculate a 95% CI on β_7. Also find a 95% CI on the mean number of games won by a team when $x_7 = 56.0$ and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from Problem 3.3.

```r
confint(reduced_model, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) -2.36784828 38.256485319
## x_7         -0.19716429  0.293906022
## x_8         -0.01015637 -0.002916818
```

From the above function call, we are 95% confident that β_7 is in the interval [-0.1972, 0.2939].

```r
new.dat = data.frame(x_7 = 56.0, x_8 = 2100)
predict(reduced_model, newdata = new.dat, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 6.926243 5.828643 8.023842
```

A 95% CI on the mean number of games won by team with the reduced model constraints is [5.8286, 8.0238].

Both the confidence intervals have **higher values** in comparison to the full model (i.e. values that are shifted to a greater value), although from a high-level overview, they seem to have a similar range between the confidence intervals.

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from the model?

By removing an important regressor from the model, there is a significant drop in the total variability (R^2) explained by the model. In addition, the confidence interval shifted slightly to the right, and the standard errors of coefficients also changed.

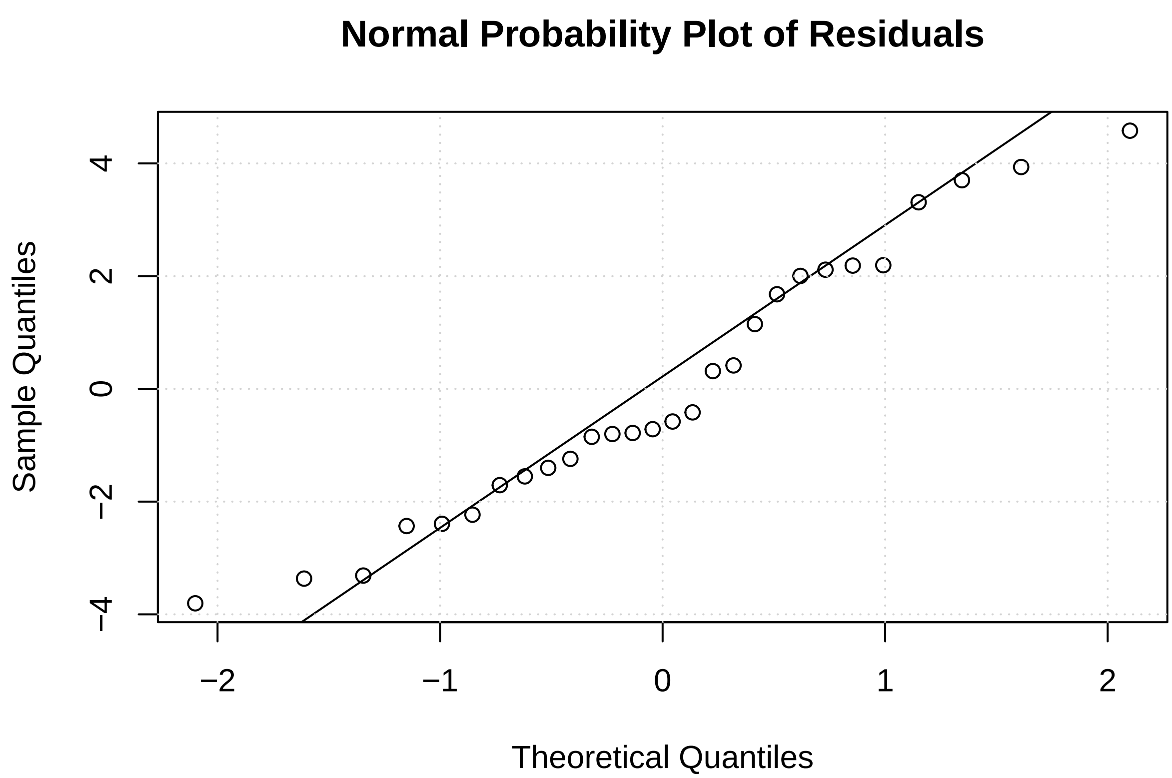
Exercise 4.1

Consider the simple regression model fit to the National Football League team performance data in Problem 2.1.

a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

We first need to obtain the residuals of the dataset (same as previous questions):

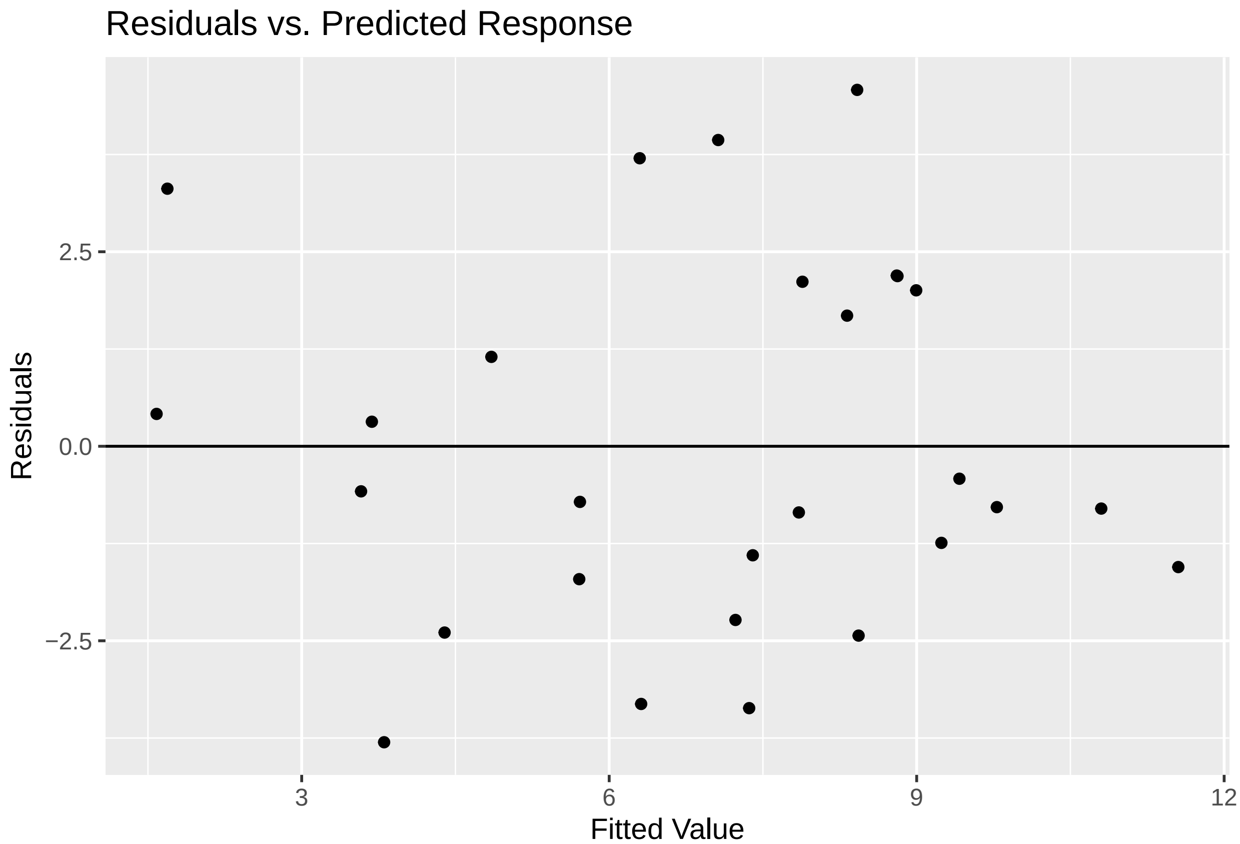
```
full_model2 = lm(y ~ x_8)
residuals <- resid(full_model2) #Obtain residuals from the full_model
qqnorm(residuals, main = "Normal Probability Plot of Residuals")
qqline(residuals) #Add line of best fit
grid()
```



Above is the graph for the normal probability plot of residuals. Since most of the residual points follow the line of best fit, there appears to be no problem with the normality assumption.

b. Construct and interpret a plot of the residuals versus the predicted response.

```
ggplot(full_model2, aes(.fitted, .resid)) + geom_point() + geom_hline(yintercept = 0) + labs(title = "R
```

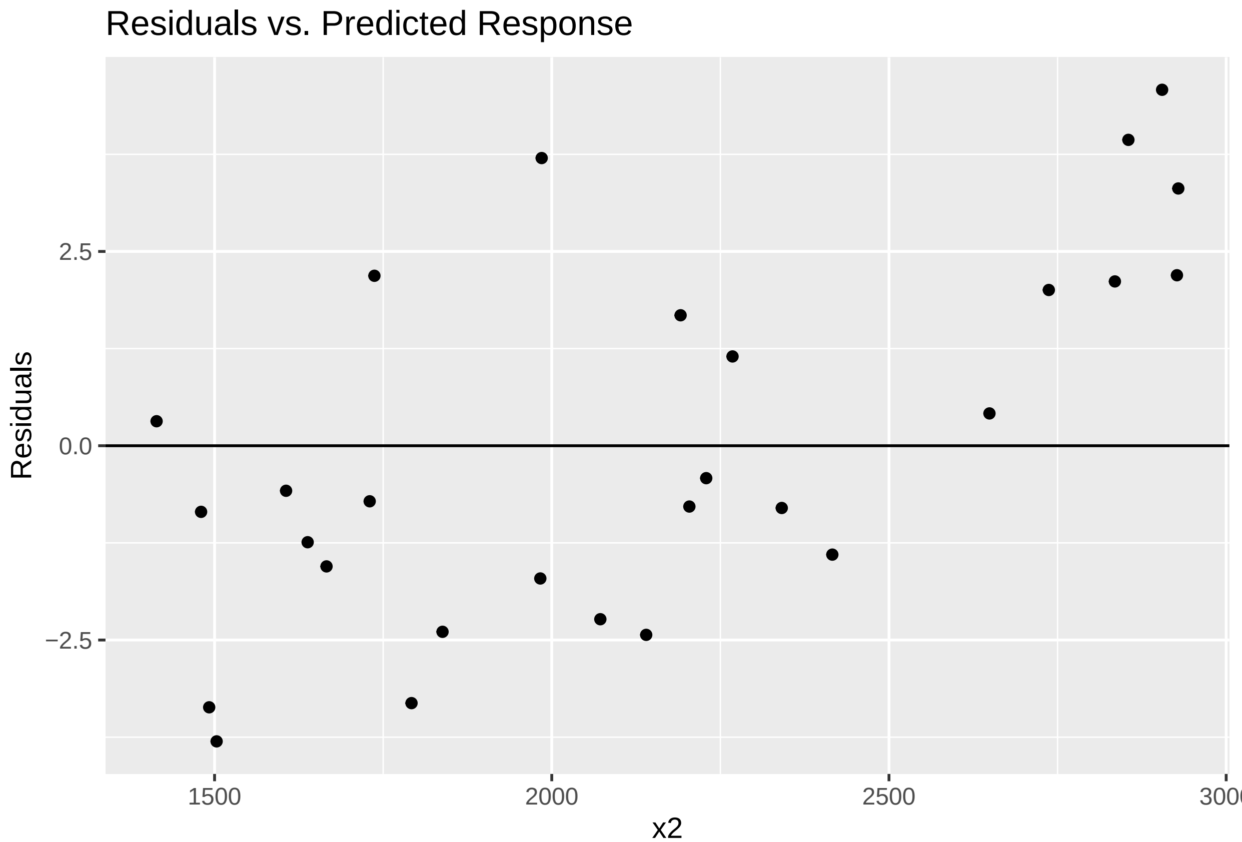


Above is the plot of residuals versus the predicted response. Most of the points appear to be close to the horizontal line where residual = 0. Since there is no clear pattern among the points, it implies that the residuals are random and not correlated to the predicted response.

Thus, we can conclude that the model's predictions are mostly accurate, with some variance that is reasonable.

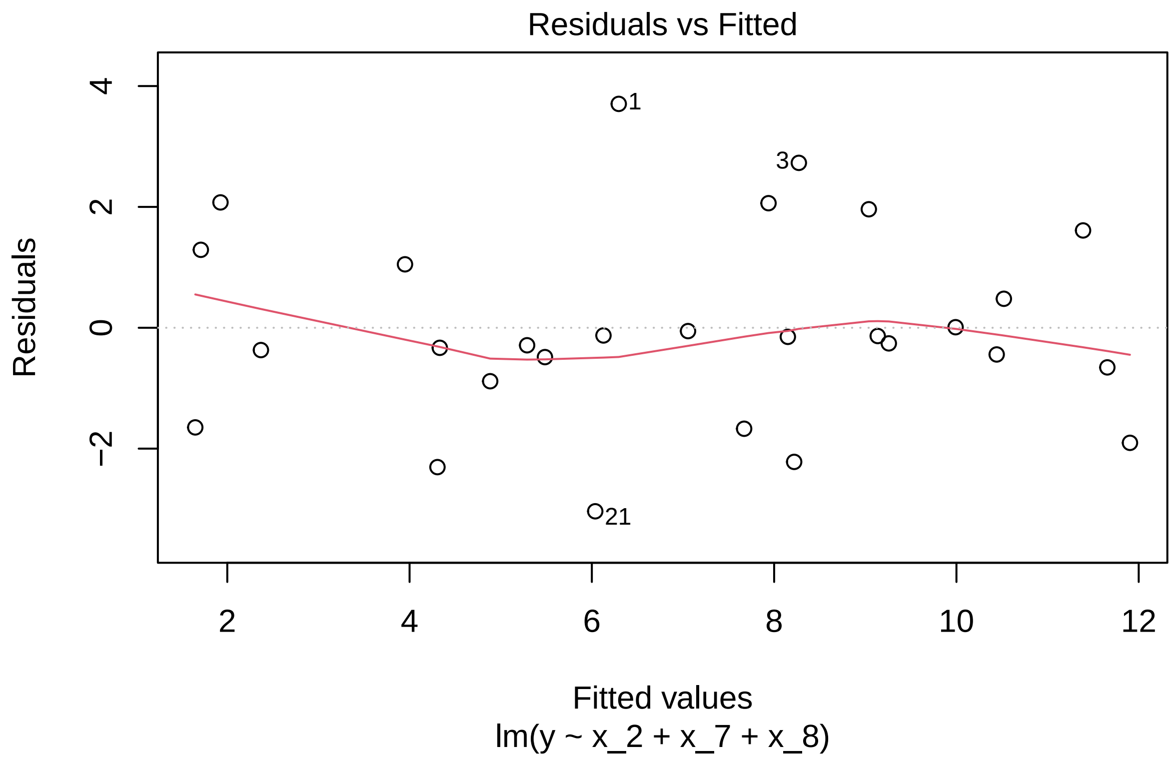
c. Plot the residuals versus the team passing yardage x_2. Does this plot indicate that the model will be improved by adding x_2 to the model?

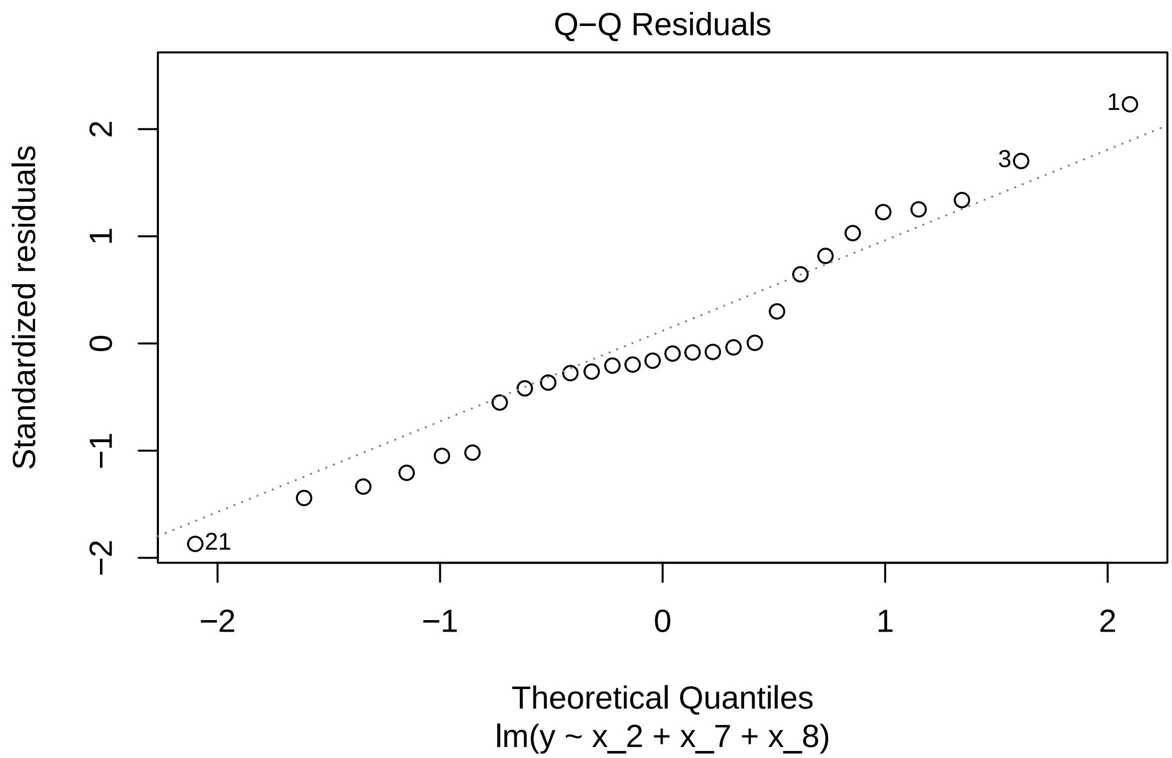
```
ggplot(full_model2, aes(x_2, .resid)) + geom_point() + geom_hline(yintercept = 0) + labs(title = "Resid
```

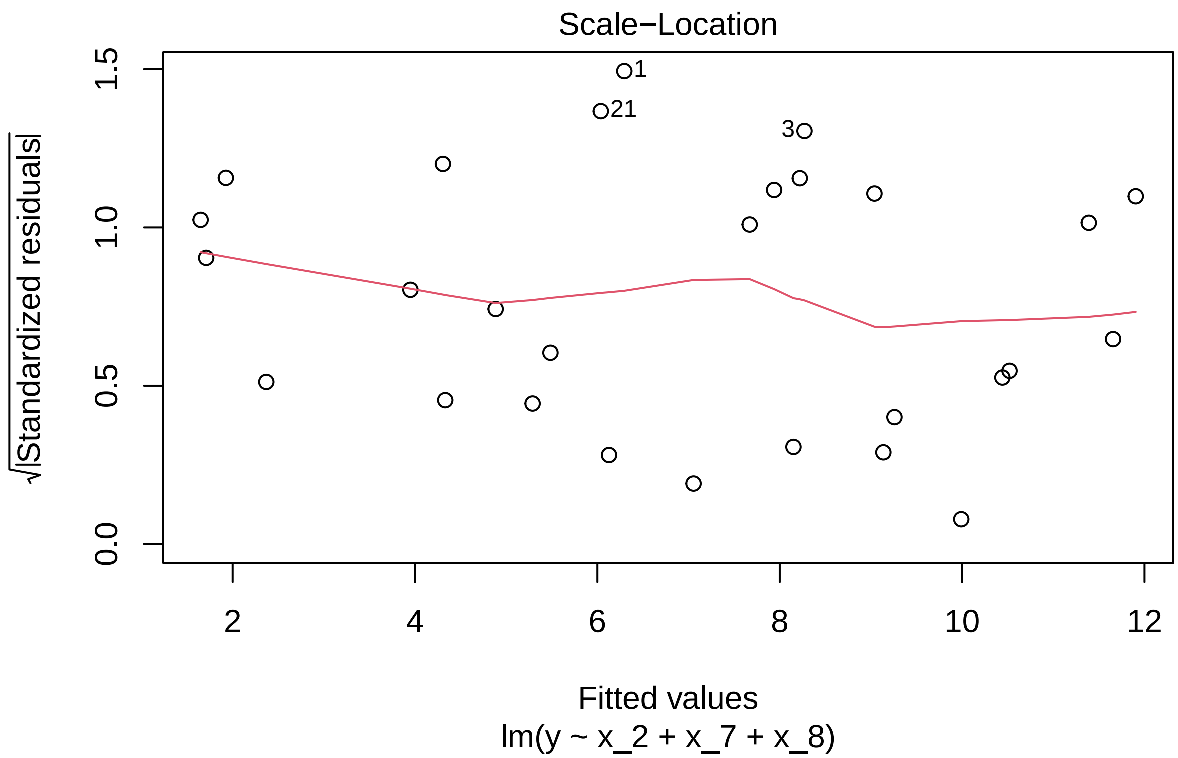


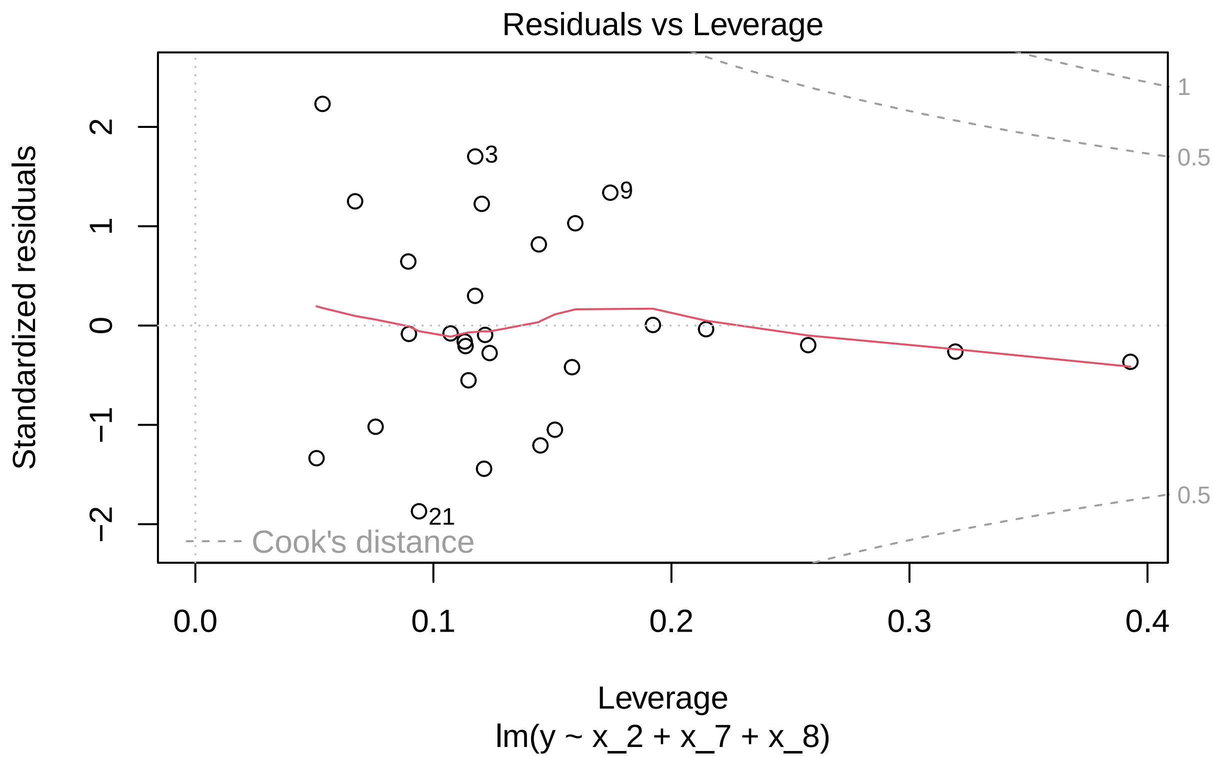
For the plot with only x_2, there is no clear pattern among the points and shows a higher concentration of points around the center of the graph. Thus, we can conclude that when the model only contains x_2 as its predictor variable, the model is more accurate.

```
myplot = mfrow = c(4, 2)
plot(full_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x_2 + x_7 + x_8)
4
2
0
−2
2
4
6
8
10
12
1
3
21


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(y ~ x_2 + x_7 + x_8)
−2
−1
0
1
2
1
3
21


Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ x_2 + x_7 + x_8)
1
21
3
0.0
0.5
1.0
1.5
2
4
6
8
10
12


Residuals vs Leverage
Standardized residuals
Leverage
lm(y ~ x_2 + x_7 + x_8)
Cook's distance
−2
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
1
0.5
0.5
3
9
21