STAT448 Assignment 1

Thomas Waldin - 17775654

2025-03-01

Part 1

Three observations for a random response variable Y are {3, 9, 15}; the corresponding values observed for the explanatory variable X are {6, 7, 8}. A linear model is assumed: Y $=\beta_0 + \beta_1$X+ ϵ

a). The ordinary least square estimates of the coefficients β_0 and β_1 were calculated by hand as follows.

$$\hat{\beta} = (x'x)^{-1} \times \hat{y}$$

$$= \frac{1}{n\sum(x_i - \overline{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$y = \begin{bmatrix} 3 \\ 9 \\ 15 \end{bmatrix} \times = \begin{bmatrix} 6 \\ 7 \\ 8 \end{bmatrix}$$

$$\hat{\beta} = \frac{1}{3(1^2 + 0^2 + 1^2)} \begin{bmatrix} 149 & -21 \\ -21 & 3 \end{bmatrix} \begin{bmatrix} 27 \\ 201 \end{bmatrix}$$

$$= \frac{1}{3(2)} \begin{bmatrix} 149(27) + -21(201) \\ -21(27) + 3(201) \end{bmatrix}$$

$$= \frac{1}{6} \begin{bmatrix} -198 \\ 36 \end{bmatrix} = \begin{bmatrix} -33 \\ 6 \end{bmatrix}$$

$$\beta_0 = -33, \beta_1 = 6$$

b). The estimates of the residuals were also calculated by hand.

$$\hat{\xi} = y - \hat{X}\beta$$

$$= \begin{bmatrix} 3 \\ 9 \\ 15 \end{bmatrix} - \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} -33 \\ 6 \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ 9 \\ 15 \end{bmatrix} - \begin{bmatrix} -33 & +6(6) \\ -33 & +7(6) \\ -33 & +8(6) \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ 9 \\ 15 \end{bmatrix} - \begin{bmatrix} 3 \\ 9 \\ 15 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\hat{\xi} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

c). The calculations above were repeated in R.

```
# Observations
y = matrix(c(3, 9, 15))
x = matrix(c(6, 7, 8))

# Design matrix X
ones = rep(1,3)
X = matrix(c(ones, x), nrow=3, ncol=2)

# Bhat = (X'X)^-1 X'y
Bhat = solve(t(X) %*% X) %*% t(X) %*% y
Bhat
```

```
##      [,1]
## [1,]  -33
## [2,]    6
```

```
# ehat = y - XBhat
ehat = y - X%*%Bhat
ehat
```

```
##              [,1]
## [1,] 8.526513e-14
## [2,] 6.394885e-14
## [3,] 4.263256e-14
```

The values of β_0 and β_1 were calculated as -33 and 6 respectively. While the residuals are not exactly zero due to numerical precision limits, they are effectively zero.

d). The coefficients can also be calculated using a simple linear regression model.

```
# Simple linear regression model
df = data.frame(x=X, y=y)
model = lm(y ~ x, data=df)

# Show coefficients
coeffs = summary(model)$coefficients
```

```
## Warning in summary.lm(model): essentially perfect fit: summary may be
## unreliable
```

```
coeffs
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -33          0    -Inf        0
## x                  6          0     Inf        0
```

The coefficients calculated are the same as the hand solution (β_0 and β_1 as -33 and 6 respectively). A warning is flagged as the model is a near perfect fit in this case, which makes sense with the estimated residuals being zero.

Part 2

In the context of Part 1, the case was considered where the values observed for the explanatory variable X are {5, 5, 5}.

a). The new coefficient estimates are as follows.

```
# Create the model again with new values in the data frame
x = matrix(c(5, 5, 5))
ones = rep(1,3)
X = matrix(c(ones, x), nrow=3, ncol=2)
df = data.frame(x=X, y=y)
model = lm(y ~ x, data=df)

# Show coefficients
coeffs = summary(model)$coefficients
coeffs
```

```
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)        9   3.464102 2.598076 0.1216899
```

```
# Look at model summary, as only one coefficient is shown
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##          1          2          3
## -6.000e+00 -1.332e-15  6.000e+00
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.000      3.464   2.598    0.122
## x                 NA         NA      NA       NA
##
## Residual standard error: 6 on 2 degrees of freedom
```

The summary showed that one of the coefficients was not defined because of singularities, the other was calculated to be 9 (intercept).

b). Statistically this is due to there being no variation in X observations, which means there is no variance to explain changes in Y. This is a single-variable equivalent of perfect multicollinearity. To estimate coefficients, the matrix X^TX must not be singular. In this case, X^TX is singular, the determinant is zero, and the matrix is therefore non-invertable meaning there is no unique solution for β_1.

c). Geometrically, the X matrix represents a unit vector (intercept column) and a vector (observations column) with equal length in all three dimensions. As these vectors are linearly dependent, they lie on top of each other and the volume of the parallelepiped between them is zero. This volume is what the determinant of the matrix represents.

Part 3

A provided CSV file contains data of student scores (response) and hours of study (explanatory variable). From this data, a simple linear regression model was generated to describe the relationship between them. Student scores are in the range 0 - 100 and hours of study are in the range 0-10.

```
# Load csv
scores = read.csv('Student_Scores_Dataset.csv')

# Linear regression model
model = lm(Scores ~ Hours, data=scores)

summary(model)
```

```
##
## Call:
## lm(formula = Scores ~ Hours, data = scores)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.8550  -3.2910  -0.0868   3.2460  15.4159
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.95989    0.35780   16.66   <2e-16 ***
## Hours        9.91401    0.05947  166.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.942 on 998 degrees of freedom
## Multiple R-squared:  0.9653, Adjusted R-squared:  0.9653
## F-statistic: 2.779e+04 on 1 and 998 DF,  p-value: < 2.2e-16
```

a). The model summary above provides the coefficients $\beta_0 = 5.960$ and $\beta_1 = 9.914$. From these, the regression equation for student score can be formulated as follows.

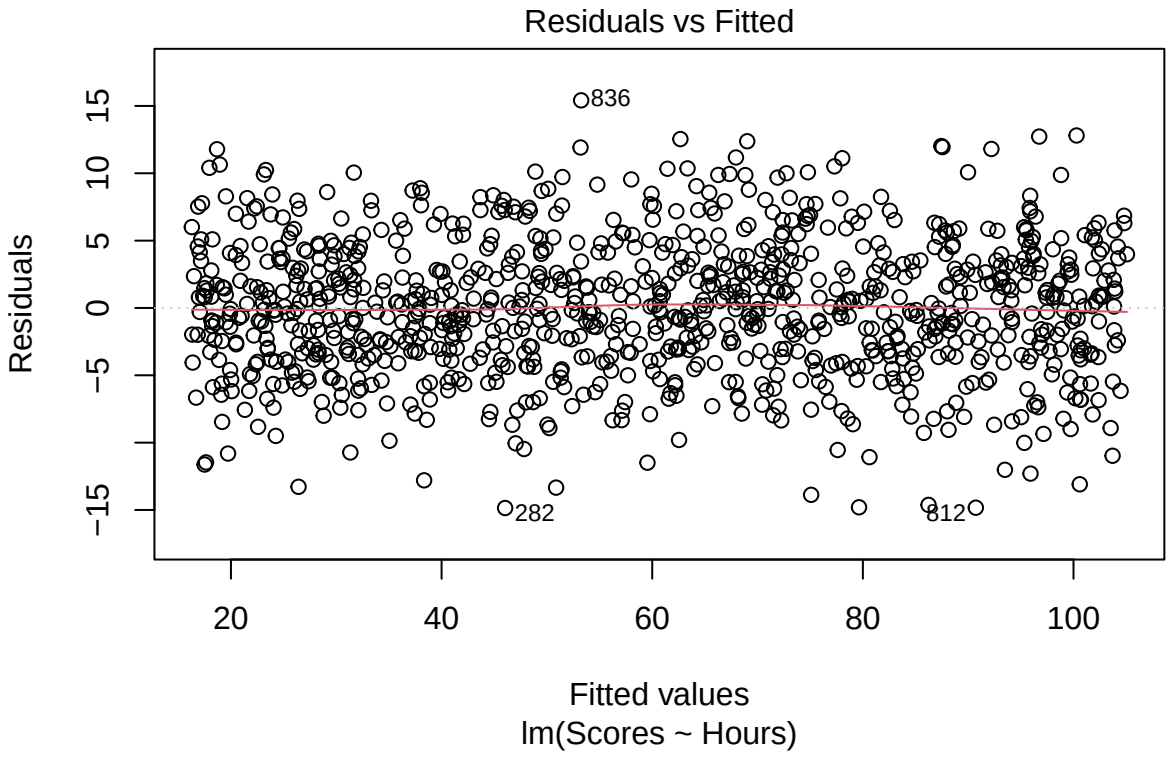
$Y = 5.960 + 9.914X + \epsilon$

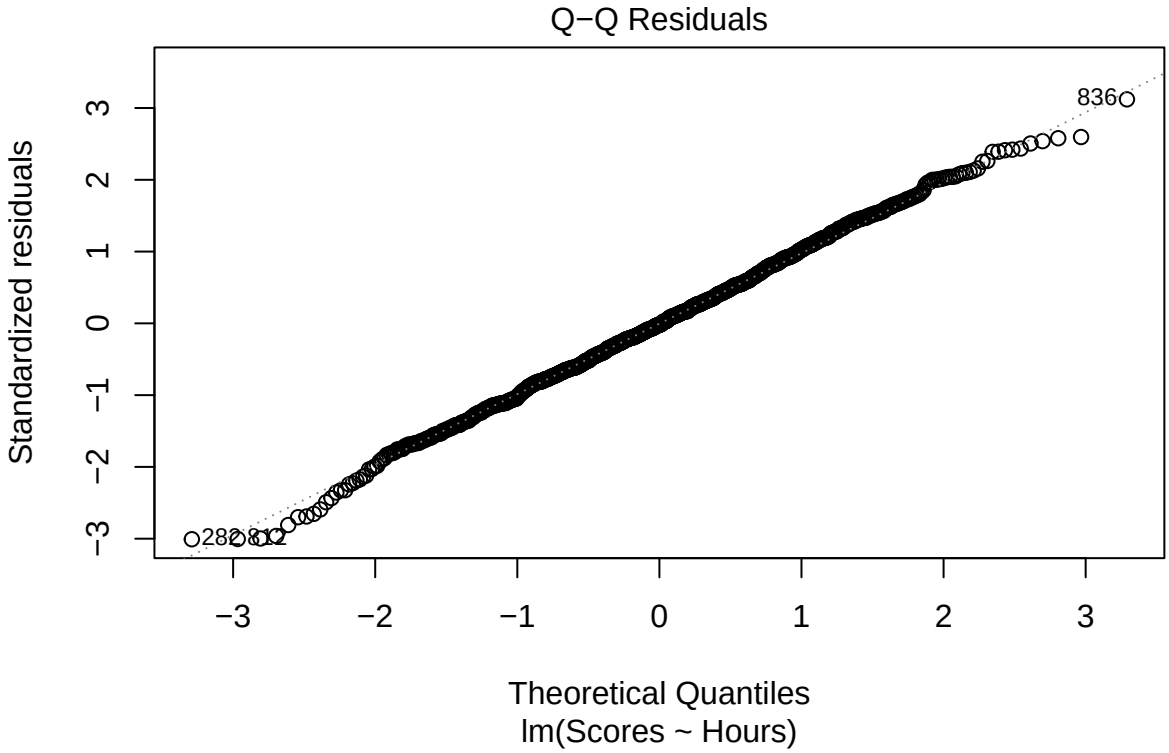
b). A slope of 9.914 means the regression model predicts that for every hour of study the student score would increase by 9.914.

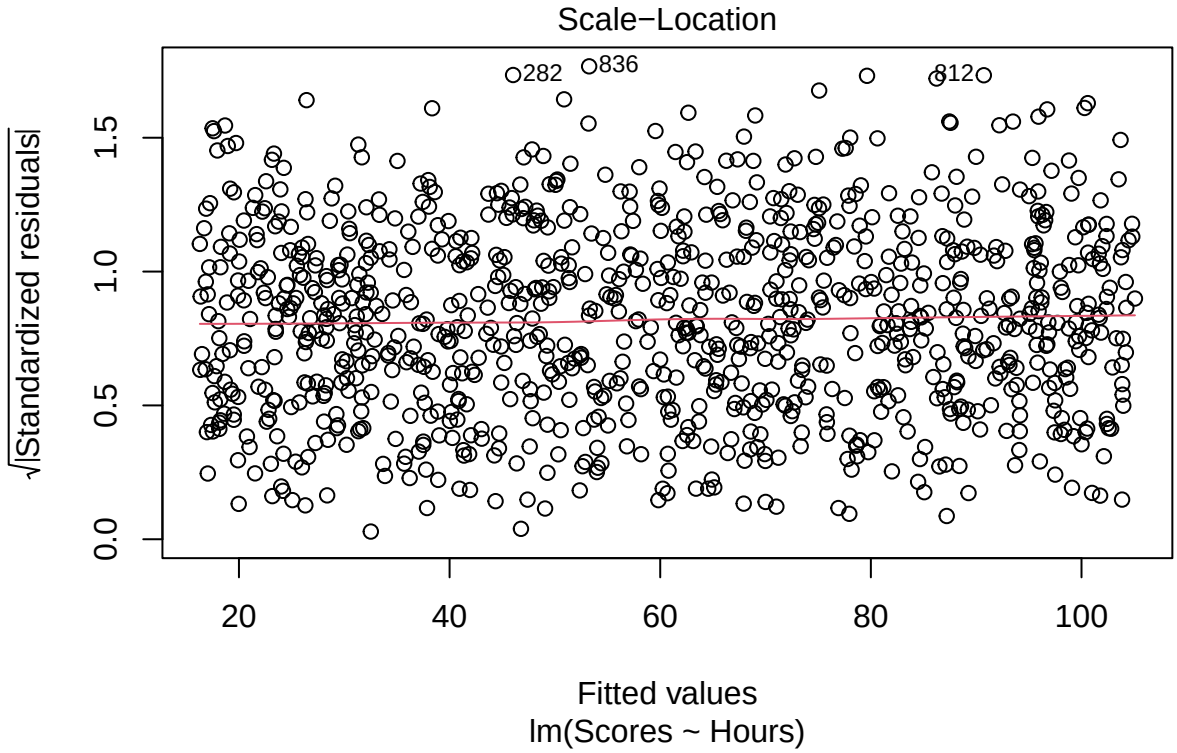
c). From the model summary, hours of study likely have an effect on the student score. The p-value is much less than 0.05 signifying that X has a significant effect on Y.

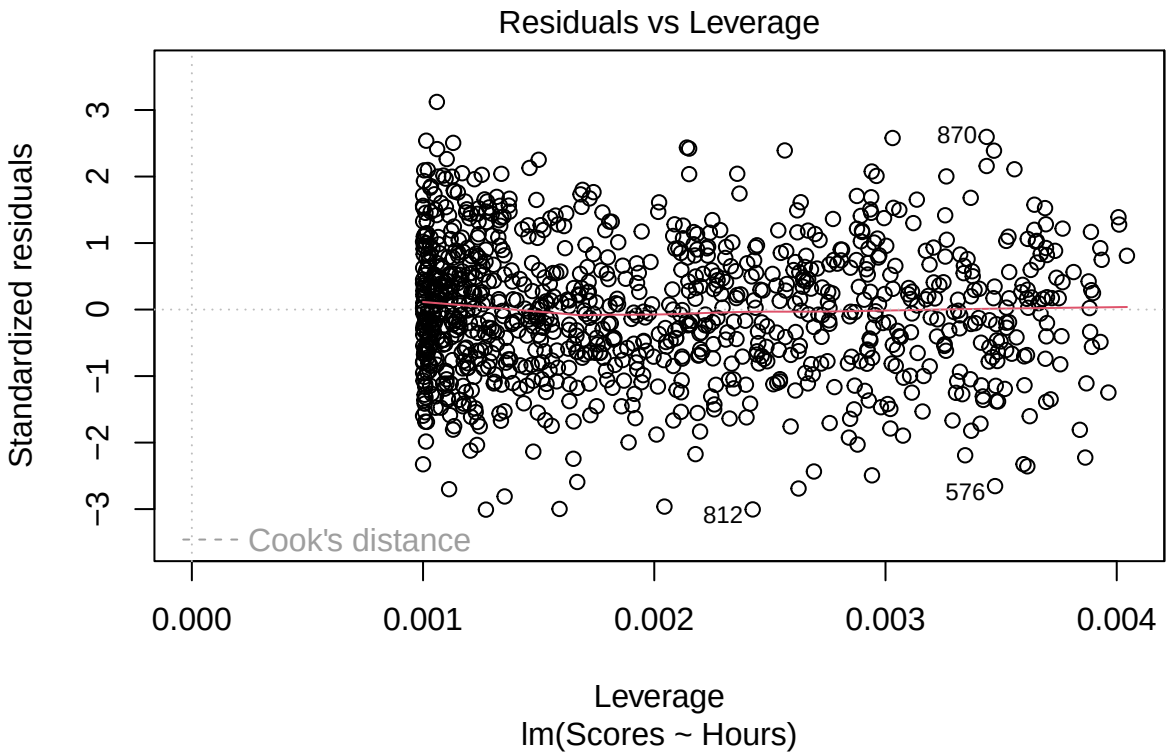
d). The R^2 value is 0.965 which is close to 0, suggesting a large amount of the variance of Y is explained by X. The RSE is smaller than the intercept ($4.942 < 5.960$) which suggests a good fit as a rule of thumb. The F-statistic also has a p-value much less than 0.05, signifying that the model is valid as a whole.

e). The model residual plots can help validate the model. The plots are shown below.




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Scores ~ Hours)
836
282
812
−3
−2
−1
0
1
2
3

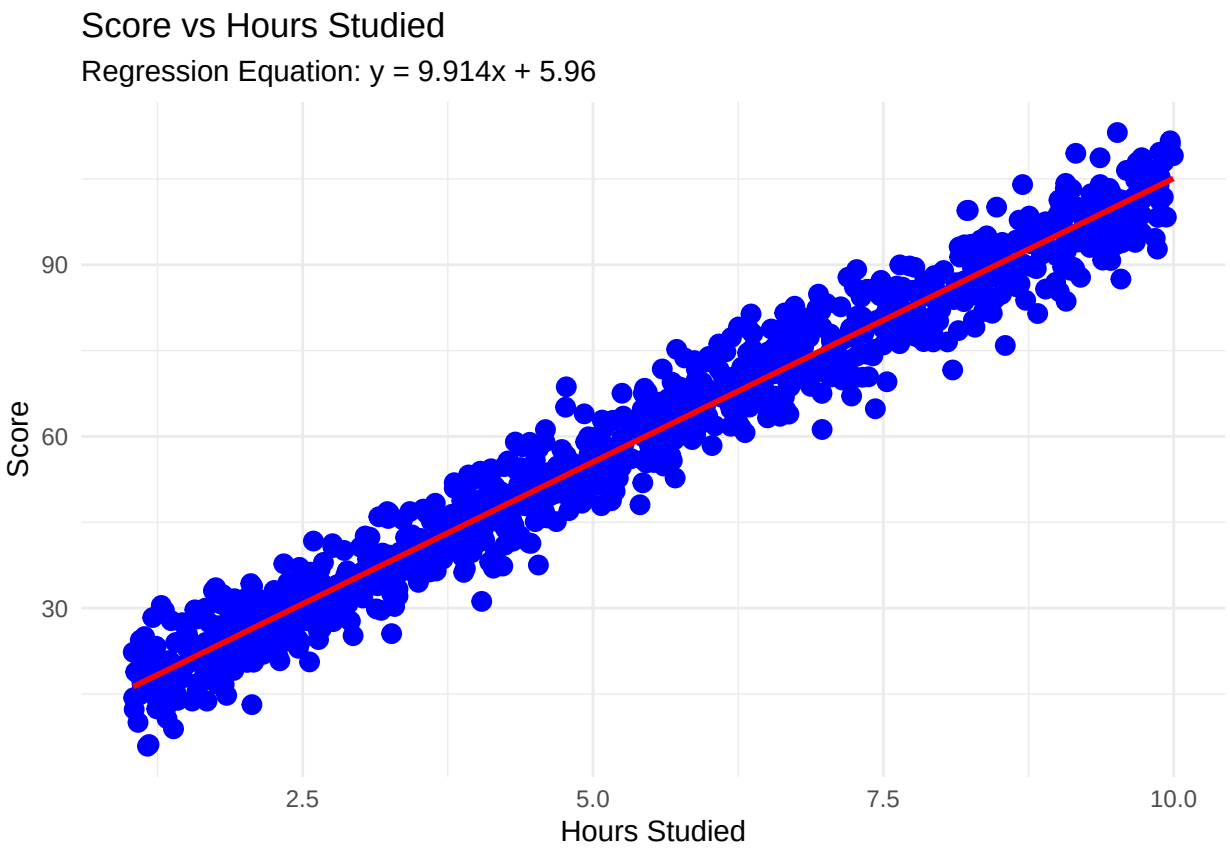

Scale−Location
282
836
812
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
40
60
80
100
Fitted values
lm(Scores ~ Hours)



In the 'Residuals vs Fitted' plot, the residual values appear to be scattered randomly around zero. This suggests that a linear model is appropriate. The 'Q-Q Residuals' plot has points that mostly follow the diagonal line, with slight deviations at the tails. This suggests that residuals are mostly normally distributed. There is no funnel shape on the 'Scale-Location' plot, suggesting no issues with heteroscedasticity and there seem to be no extreme leverage points (beyond Cook's distance) on the 'Residuals vs Leverage' plot.

f). A plot of the observations and the regression line is provided below.

```
## `geom_smooth()` using formula = 'y ~ x'
```



g). Using the equation of the regression line, score predictions can be made, given hours of study.

```
# Score predictions
X = data.frame(Hours = c(4.36, 6.86, 8.84))
predictions = predict(model, X)
predictions
```

```
##        1        2        3
## 49.18497 73.96999 93.59973
```

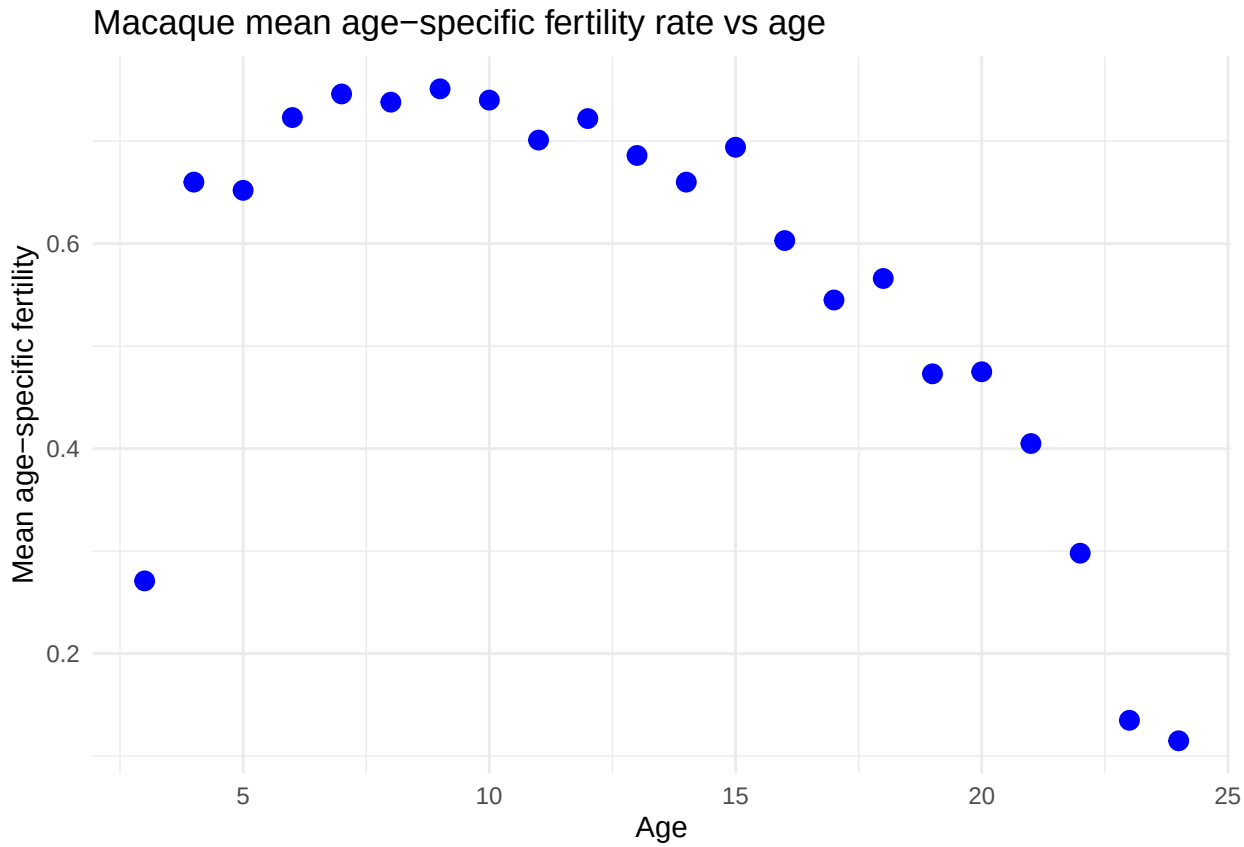
The model predicts scores of 49.18, 73.97, and 93.60 for study hours of 4.36, 6.86, and 8.84 respectively. The data in this model has a range of 0-10 hours and scores of 0-100. Predictions outside of this range (i.e. predicting a score for hours > 10) require extrapolation. This means predictions are based on assumptions rather than real data, which means the predictions are unlikely to be valid. Using this model assumes that the relationship between hours and scores remains linear which is unlikely to be true. Studying beyond a certain amount of hours will likely have diminishing returns so the assumption of a linear relationship is poor.

Part 4

A simple linear regression model was employed to investigate the connections between fertility rate and age in female rhesus macaques from Cayo Santiago. Reproductive data from Cayo Santiago rhesus macaque females was used, as documented in Luevano et al. (2022). The goal was to determine whether female fertility is influenced by age through the application of simple linear regression analysis. The mean age-specific fertility

rate is defined as the number of offspring produced at age 'x' divided by the total number of females of age 'x'.

a). The mean age specific fertility is plotted against age below.



As the age of the Macaques increases, the mean age-specific fertility appears to increase until approximately age 8, and then decrease. The association appears to be non-linear.

b). A simple linear regression is modeled below.

```
model = lm(mean_fertility ~ age, data=data)
coeffs = summary(model)$coefficients
coeffs
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  0.82176398 0.080312528 10.232077 2.148106e-09
## age         -0.01925861 0.005384164 -3.576899 1.886890e-03
```

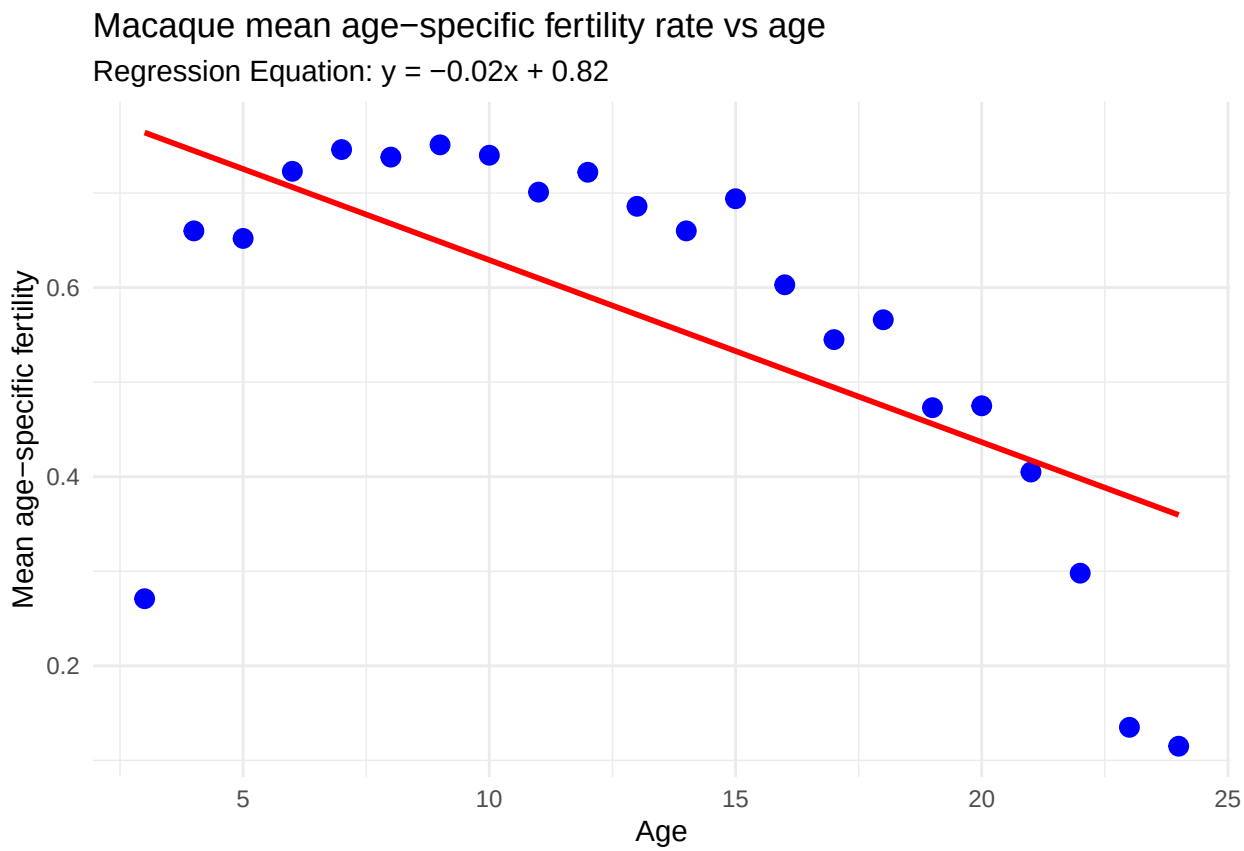
From this model a linear regression equation can be formed.

y = 0.822 -0.019x

This model predicts that for every year that age increases, the mean age-specific fertility rate decreases by 0.019. It also predicts that for Macaques of age 0 the fertility rate is 0.822 (intercept).

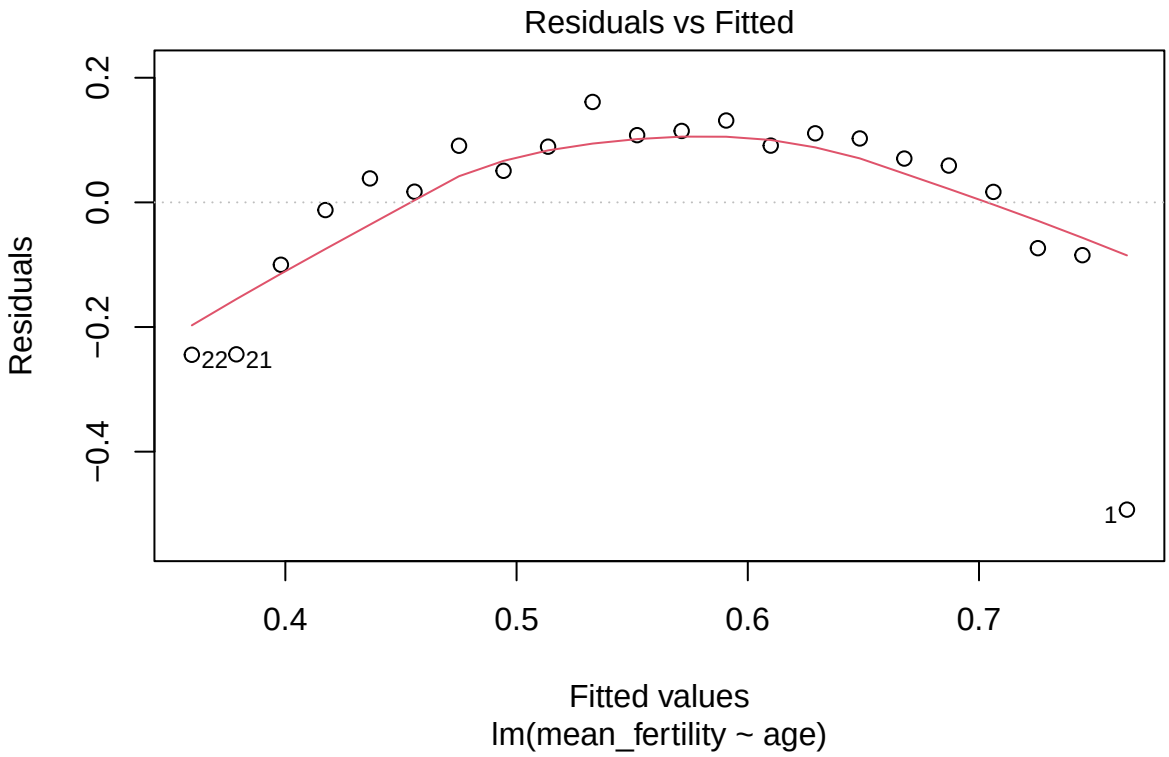
c). The simple linear regression is plotted onto the graph of the observations below.

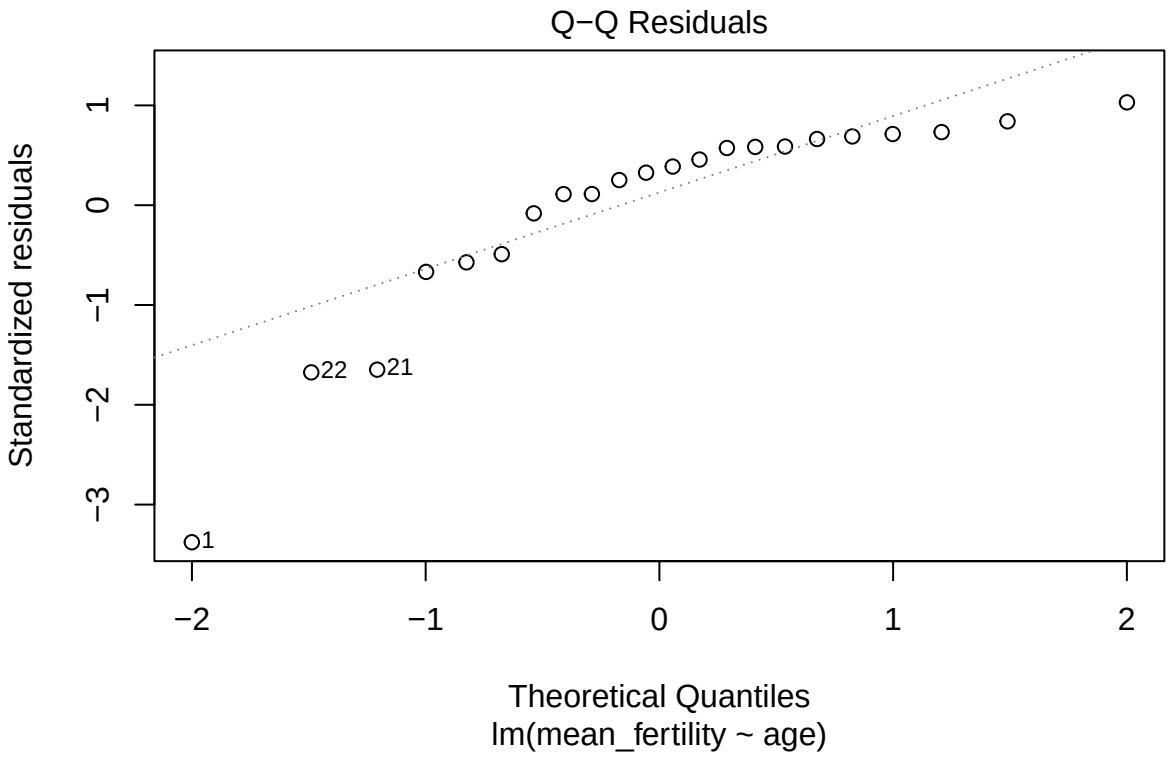
```
## `geom_smooth()` using formula = 'y ~ x'
```

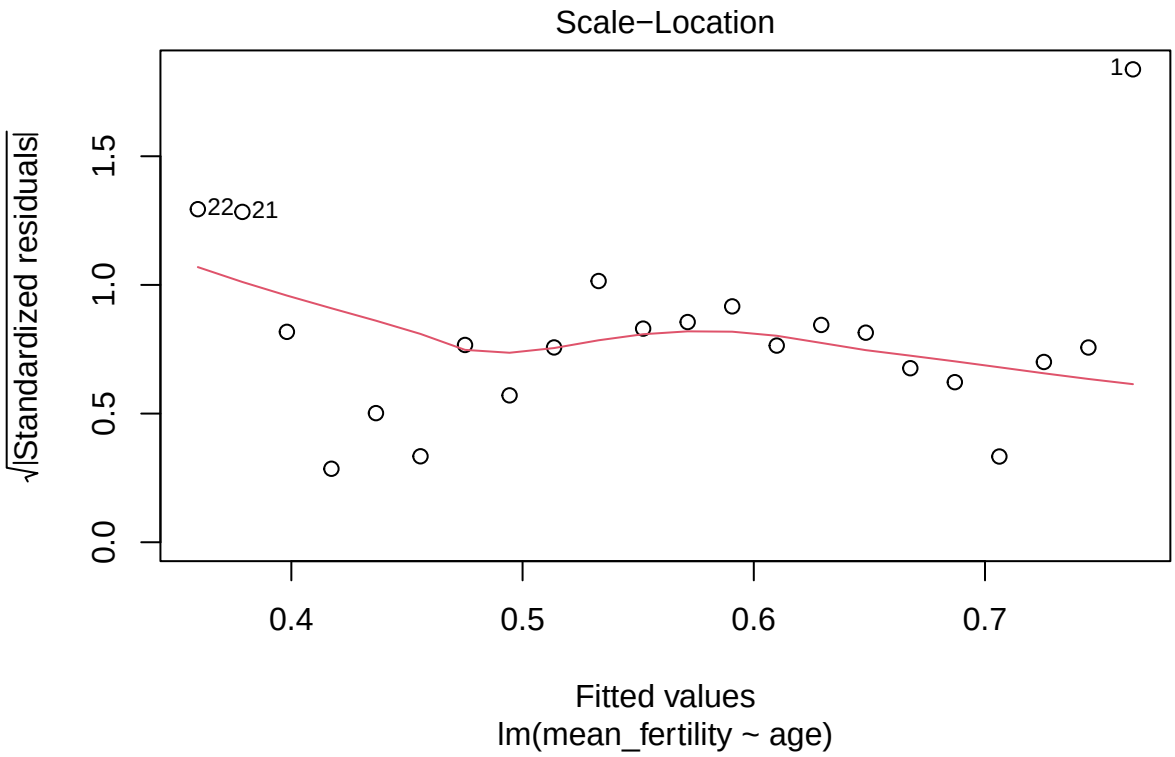


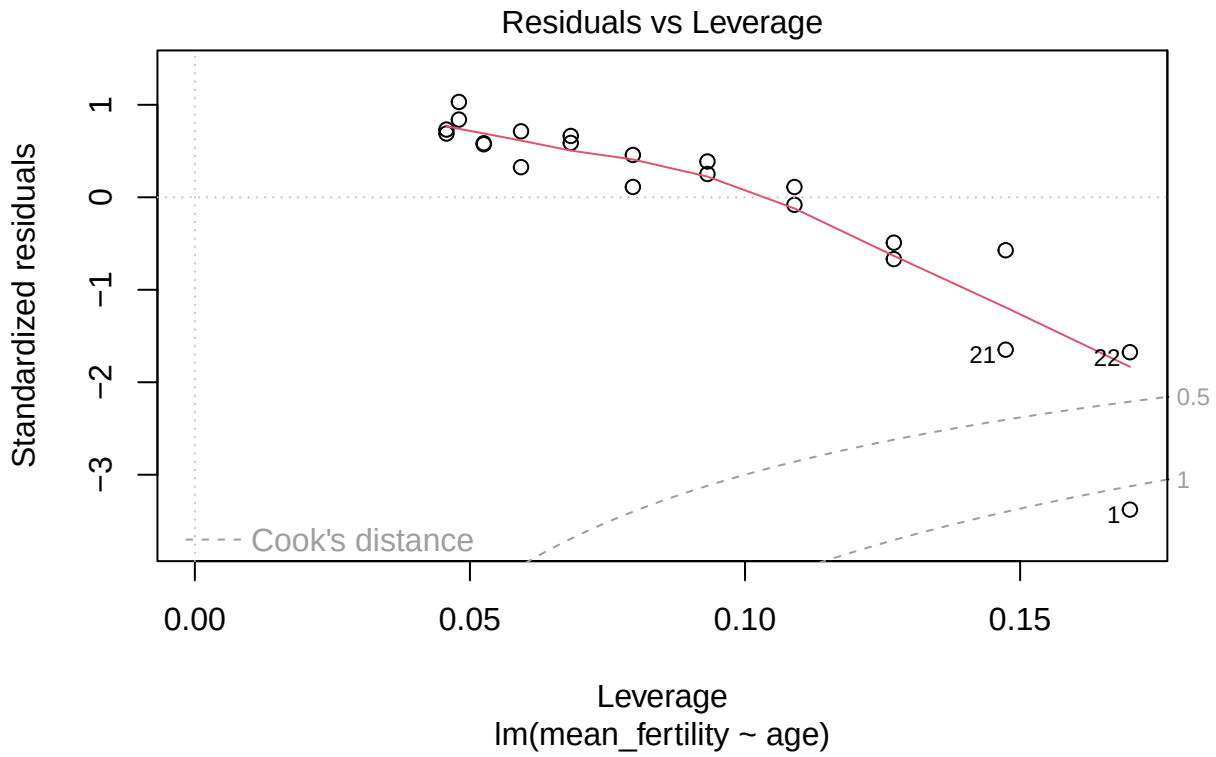
A linear model appears not to model the data well. The observations create a clear quadratic arc which a straight line cannot represent well. It is obvious that the fit is not adequate, but this can be investigated further with residual plots.

```
# Check residuals
plot(model)
```


Residuals vs Fitted
Residuals
0.2
0.0
−0.2
−0.4
22
21
1
0.4
0.5
0.6
0.7
Fitted values
lm(mean_fertility ~ age)


Q−Q Residuals
Standardized residuals
1
0
−1
−2
−3
22
21
1
−2
−1
0
1
2
Theoretical Quantiles
lm(mean_fertility ~ age)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.4
0.5
0.6
0.7
Fitted values
lm(mean_fertility ~ age)
1
22
21



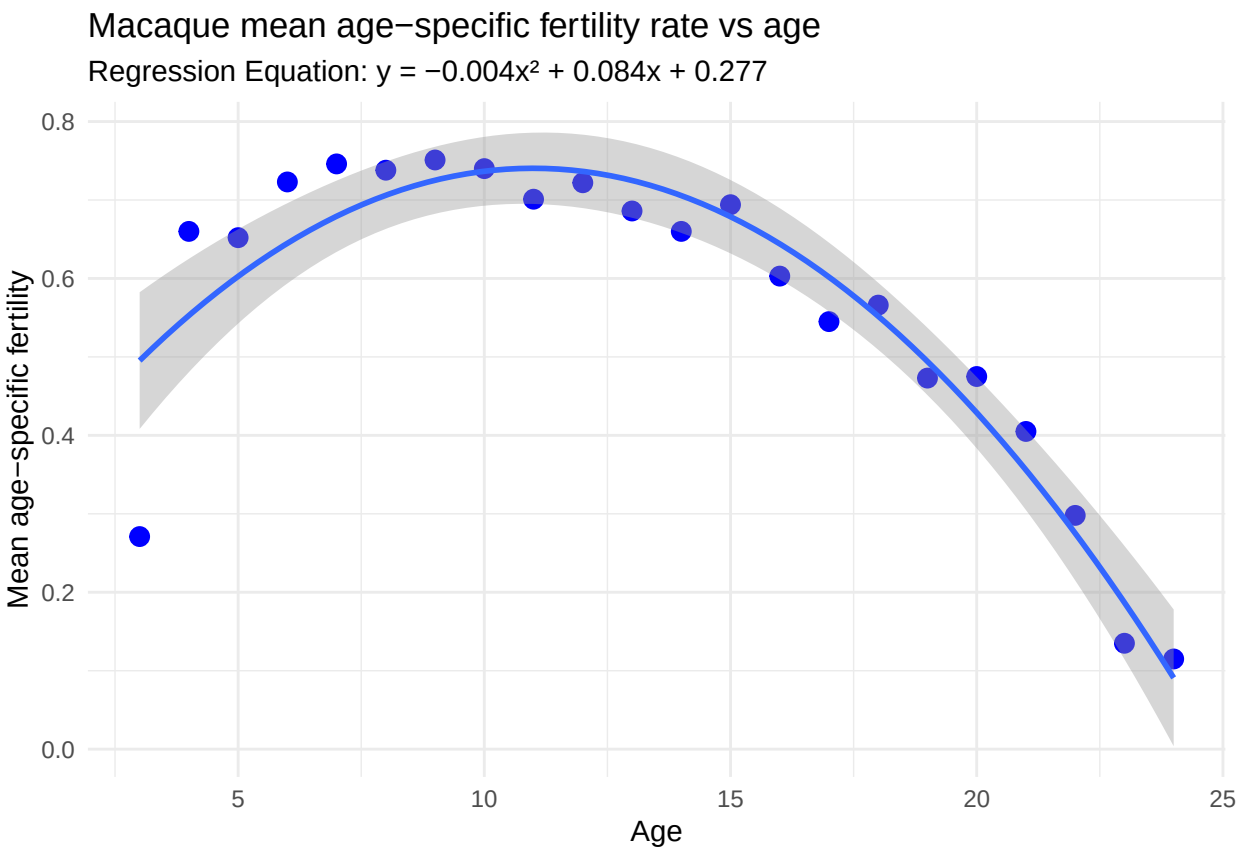
In the 'Residuals vs Fitted' plot, the residual values appear to be in an obvious arc and are not scattered randomly around zero. This suggests that a linear model is not appropriate as was stated earlier. The 'Q-Q Residuals' plot has points that deviate from the diagonal line in an s-curve, with heavy deviations at the tails. This suggests that residuals are not normally distributed which is an assumption of linear regression. There is no funnel shape on the 'Scale-Location' plot, suggesting no issues with heteroscedasticity but there seems to be one extreme leverage point (beyond Cook's distance) on the 'Residuals vs Leverage' plot, that may be distorting the model.

The residual plots suggest a better fit is possible. Adding a quadratic term to the model is explored below.

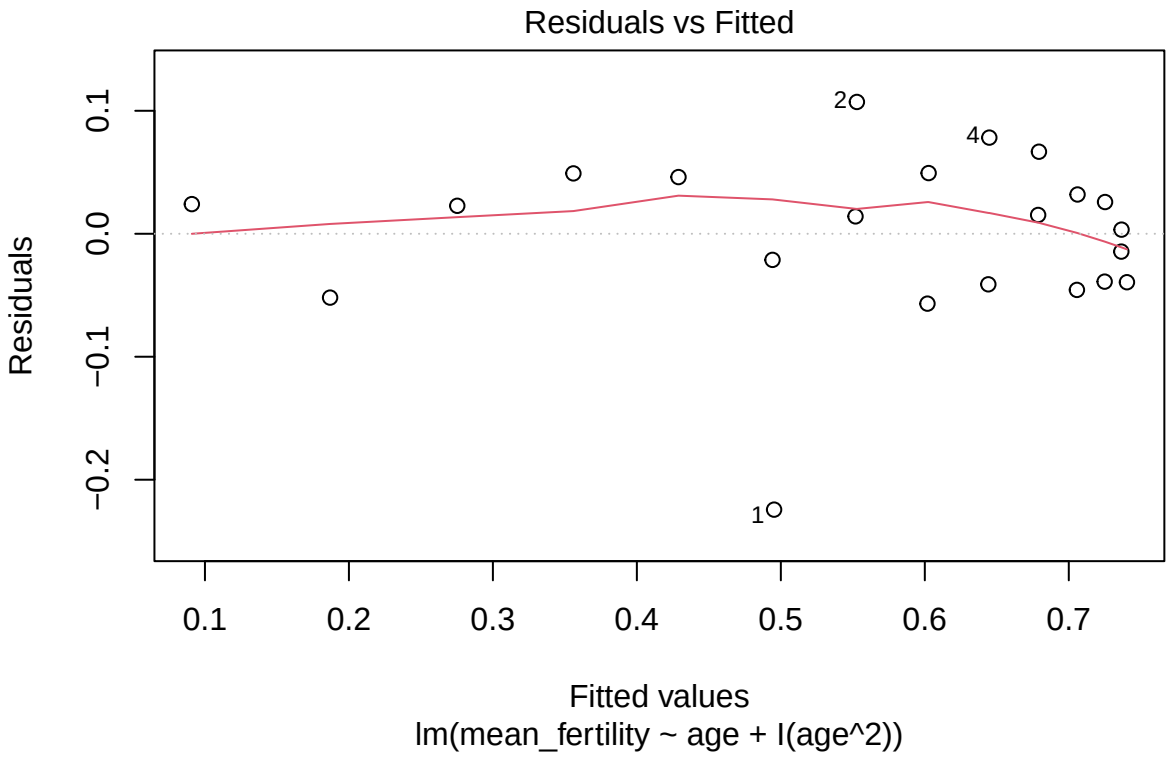
```
# Try adding quadratic term
model_quad = lm(mean_fertility ~ age + I(age^2), data=data)
coeffs = summary(model_quad)$coefficients
coeffs
```

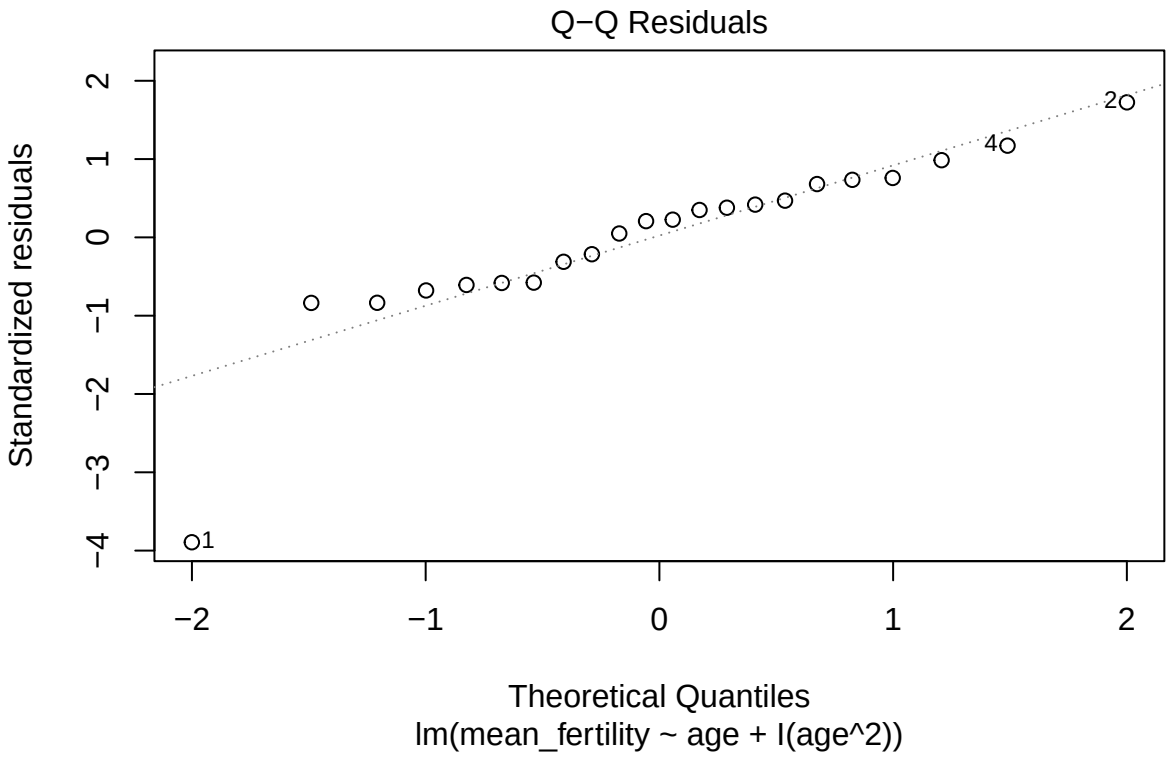
```
##                 Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  0.276765810 0.0697497622  3.967982 8.243305e-04
## age          0.084367801 0.0116487960  7.242620 7.105125e-07
## I(age^2)    -0.003838015 0.0004222685 -9.089039 2.395496e-08
```

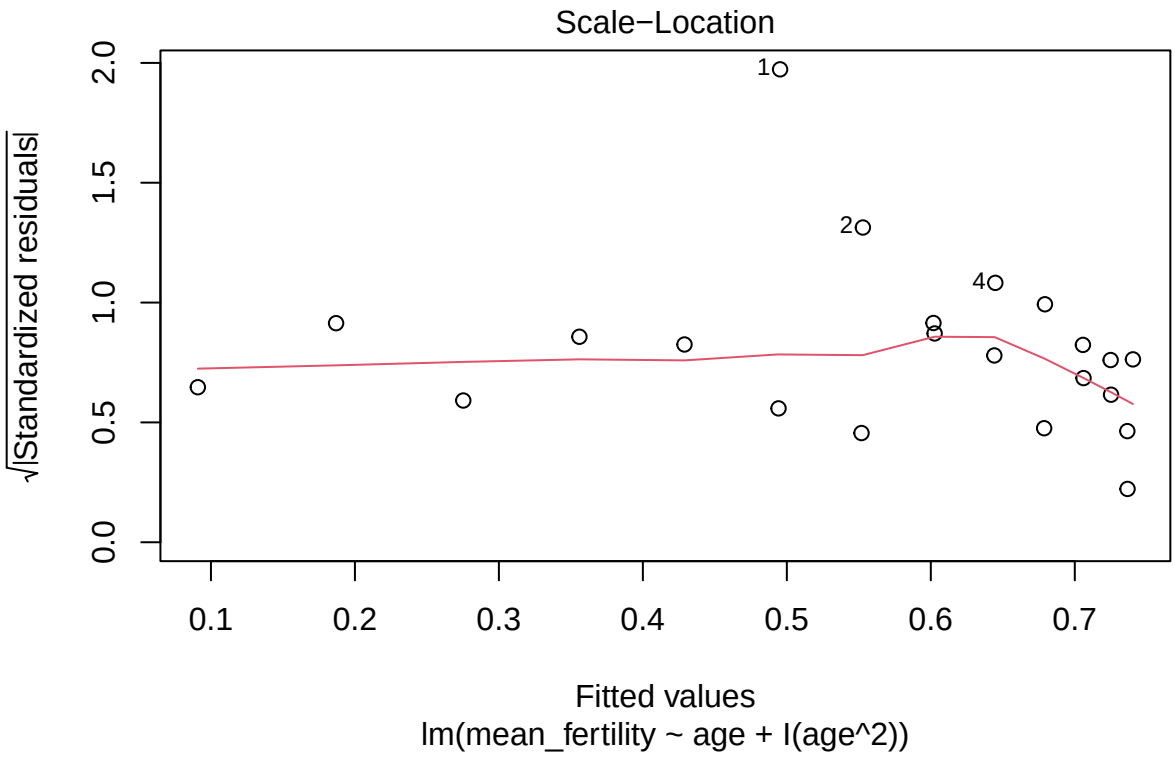
This model was plotted with the observations below.

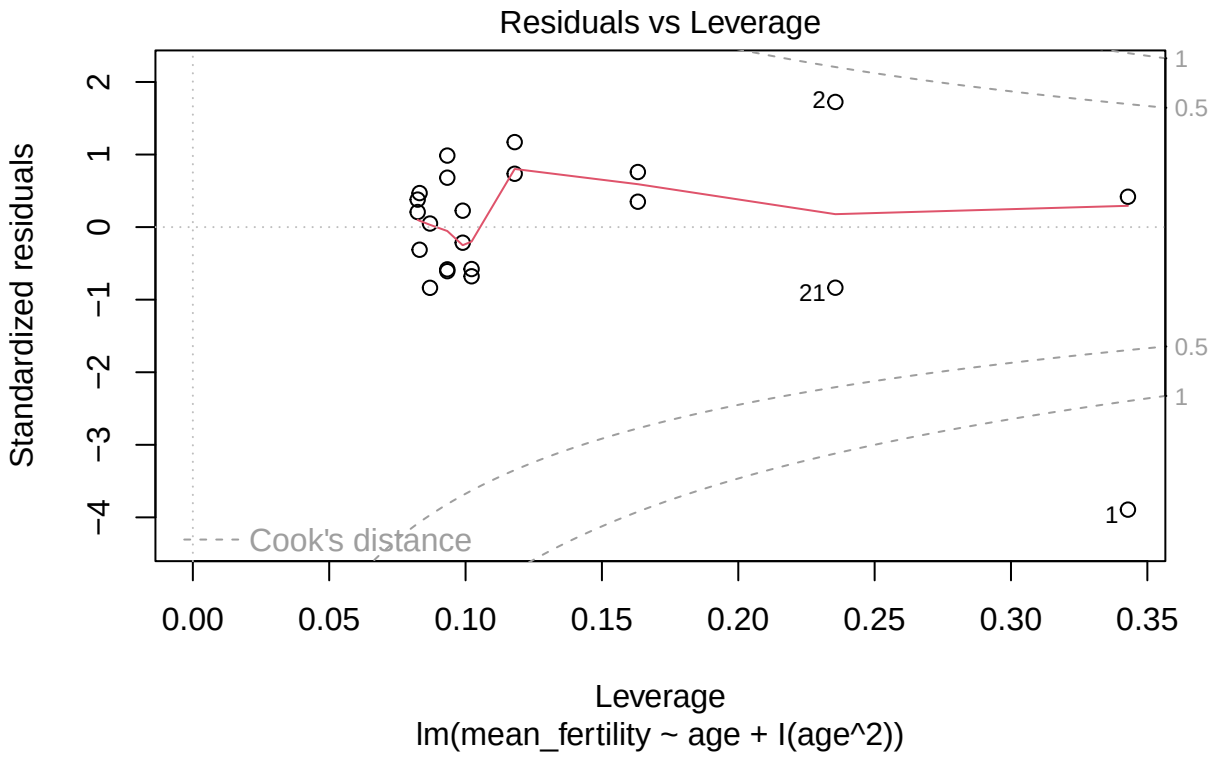


Adding the quadratic term appears to greatly improve the fit of the model. The residual plots of the model are investigated below.


Residuals vs Fitted
Residuals
0.1
0.0
−0.1
−0.2
2
4
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Fitted values
lm(mean_fertility ~ age + I(age^2))


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−4
1
2
4
−2
−1
0
1
2
Theoretical Quantiles
lm(mean_fertility ~ age + I(age^2))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1
2
4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Fitted values
lm(mean_fertility ~ age + I(age^2))



In the 'Residuals vs Fitted' plot, the residual values seem to be scattered fairly randomly around zero. This suggests that the model with the quadratic term is appropriate. The 'Q-Q Residuals' plot has points that follow the diagonal line, but with slight s-curve deviations. This suggests that residuals are mostly normally distributed. There is no funnel shape on the 'Scale-Location' plot, suggesting no issues with heteroscedasticity but there seems extreme leverage point as before (beyond Cook's distance) on the 'Residuals vs Leverage' plot. Excluding this point could potentially improve the fit, but contextually this point is important. The data point that has extreme leverage is the first one, or the point representing the fertility rate of macaques aged one, which is understandably significantly lower than macaques of greater age.

d). Using the equation of the regression line, age-specific fertility rate predictions can be made, given the age of a macaque.

```
# Fertility predictions
X = data.frame(age = c(6.50, 14.25, 18.75))
predictions = predict(model_quad, X)
predictions
```

```
##         1         2         3
## 0.6630004 0.6996500 0.5093598
```

The model predicts fertility rates of 0.66, 0.70, and 0.51 for ages of 6.50, 14.25, and 18.75 respectively.