STAT448 - Assignment 1

NDU31 - Nicole Dunn

2022-08-10

Question 1 Part a

a)

OLS method by hand using the puzzle inverse method and step by step multiplication of matrices. Coefficients are -11 and 5.

Question 1 part a)

Minimise S

$$S = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \epsilon'\epsilon$$

$$= (y - X\beta)'(y - X\beta)$$

$$\frac{\partial S}{\partial \beta} = -2X'(y - X\beta) \rightarrow \text{set equal to zero.}$$

$$\hat{\beta} = (X'X)^{-1}X'y \quad \text{- overall calculation for OLS}$$

$$y = \begin{bmatrix} 4 \\ 9 \\ 14 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \qquad X' = \begin{bmatrix} 1 & 1 & 1 \\ 3 & 4 & 5 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1\times1+1\times1+1\times1 & 1\times3+1\times4+1\times5 \\ 3\times1+4\times1+5\times1 & 3\times3+4\times4+5\times5 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 12 \\ 12 & 50 \end{bmatrix}$$

Determinant $= (3\times50) - (12\times12)$

$= 150 - 144$

$= 6$

Inversing the 2×2 matrix:

$$\begin{bmatrix} 3 & 12 \\ 12 & 50 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \rightarrow 4R1 + R2,$$ multiplying the first row by 4 and adding row 2 to replace row 2.

$$\begin{bmatrix} 3 & 12 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix} \rightarrow 6R2 - R1,$$ multiplying the second row by 6 and subtracting the first row

$$\begin{bmatrix} -3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} -2.5 & 6 \\ -4 & 1 \end{bmatrix} \rightarrow R1 \div (-3),$$ dividing row 1 by −3

$$\rightarrow R2 \times 0.5,$$ multiplying row 2 by 0.5

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 8.3 & -2 \\ -2 & 0.5 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 12 & 50 \end{bmatrix}^{-1}$$

Question 1, Part a and b

b)

The estimates of the residuals $\hat{\epsilon}$ are 0 as there is no residual after the equation is equaled out.

$$(x'x)^{-1}x' = \begin{bmatrix} 8.3 & -2 \\ -2 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 1 \\ 3 & 4 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} 8.3\times1 + -2\times3 & 8.3\times1 - 2\times4 & 8.3\times1 - 2\times5 \\ -2\times1 + 0.5\times3 & -2\times1 + 0.5\times4 & -2\times1 + 0.5\times5 \end{bmatrix}$$

$$= \begin{bmatrix} 2.3 & 0.3 & -1.67 \\ -0.5 & 0 & 0.5 \end{bmatrix}$$

$$(x'x)^{-1}x'y = \begin{bmatrix} 2.3 & 0.3 & -1.67 \\ -0.5 & 0 & 0.5 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 9 \\ 14 \end{bmatrix}$$ multiplication of matrices.

$$= \begin{bmatrix} 2.3\times4 + 0.3\times9 + -1.67\times14 \\ -0.5\times4 + 0\times9 + 0.5\times14 \end{bmatrix}$$

$$= \begin{bmatrix} -11 \\ 5 \end{bmatrix}$$

$\hat{\beta}_0 = -11$ } coefficients
$\hat{\beta}_1 = 5$

Part B)

$\epsilon + (3\times5) - 11 = 4$
$\epsilon + (4\times5) - 11 = 9$
$\epsilon + (5\times5) - 11 = 14$

$\epsilon = 0$ as all the equations above do not leave any residuals to be the error. In other words $\{4, 9, 14\} = Y$

Question 1 Part c and d

c)

Matrix calculations performed into the beta variable

d)

Coefficients match with -11 and 5 which is further checked by using the lm function

```
qdata <- data.frame(x1 = c(3, 4, 5), y = c(4, 9, 14))

par(mfrow=c(1, 1))

y <- c(4, 9, 14)
x1 <- c(3, 4, 5)

x0 <- rep(1, length(y))

Y <- as.matrix(qdata$y)
Y
```

```
##      [,1]
## [1,]    4
## [2,]    9
## [3,]   14
```

```
X <- as.matrix(cbind(x0, qdata$x1))
X
```

```
##      x0
## [1,]  1 3
## [2,]  1 4
## [3,]  1 5
```

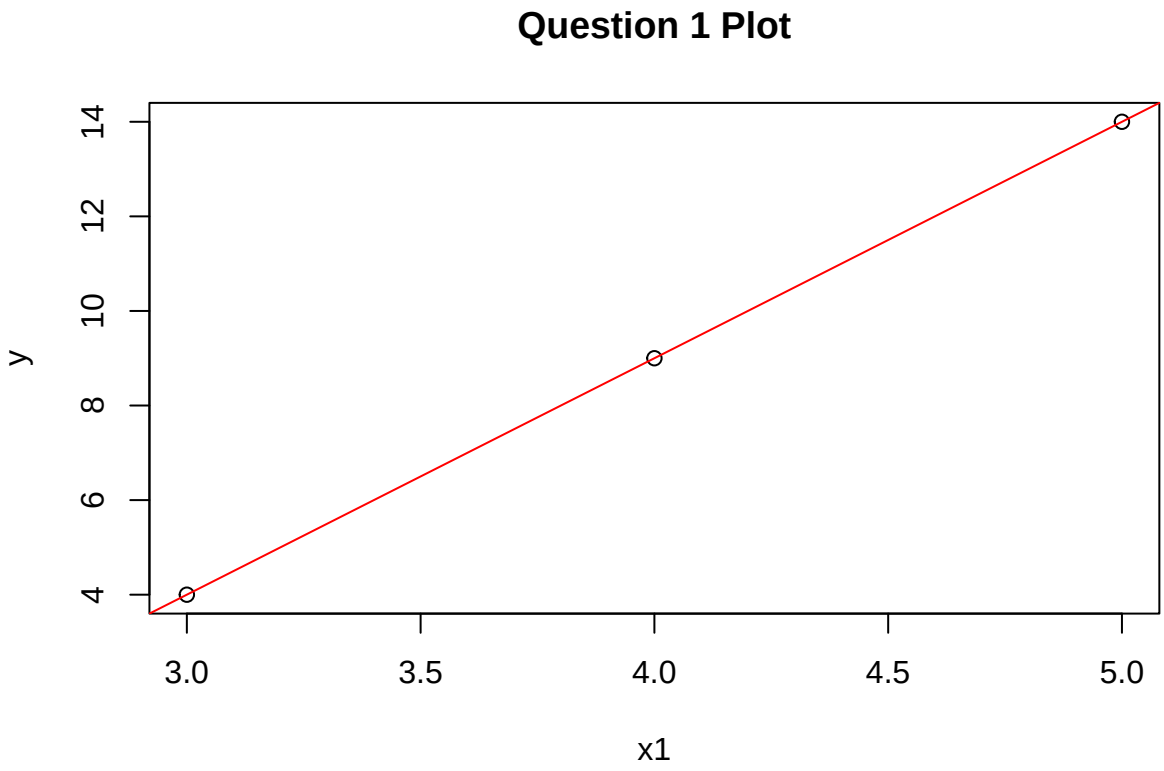
```
beta <- solve(t(X) %*% X) %*% (t(X) %*% Y)
beta
```

```
##    [,1]
## x0  -11
##       5
```

```
fit1 <- lm(y ~ x1)
fit1
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Coefficients:
## (Intercept)           x1
##         -11            5
```

```r
plot(qdata, xlab="x1", ylab="y", main="Question 1 Plot")
abline(beta, col="red")
```



Question 2 Part a and b

a)

The coefficients show as 9 and NA, solving the problem provides errors as they are invalid values for this equation. The plot shows where the data points are sitting and why this would throw errors for the R equations.

b)

The lm functions shows that there is a NA intercept as the line is just a straight line as shown in the below plot. This implies that it could go on for infinity and that there is no correlation between the values.

```r
q2data <- data.frame(x2 = c(2, 2, 2), y = c(4, 9, 14))

x2 <- c(2, 2, 2)

Xmat <- as.matrix(cbind(x0, x2))
Xmat
```

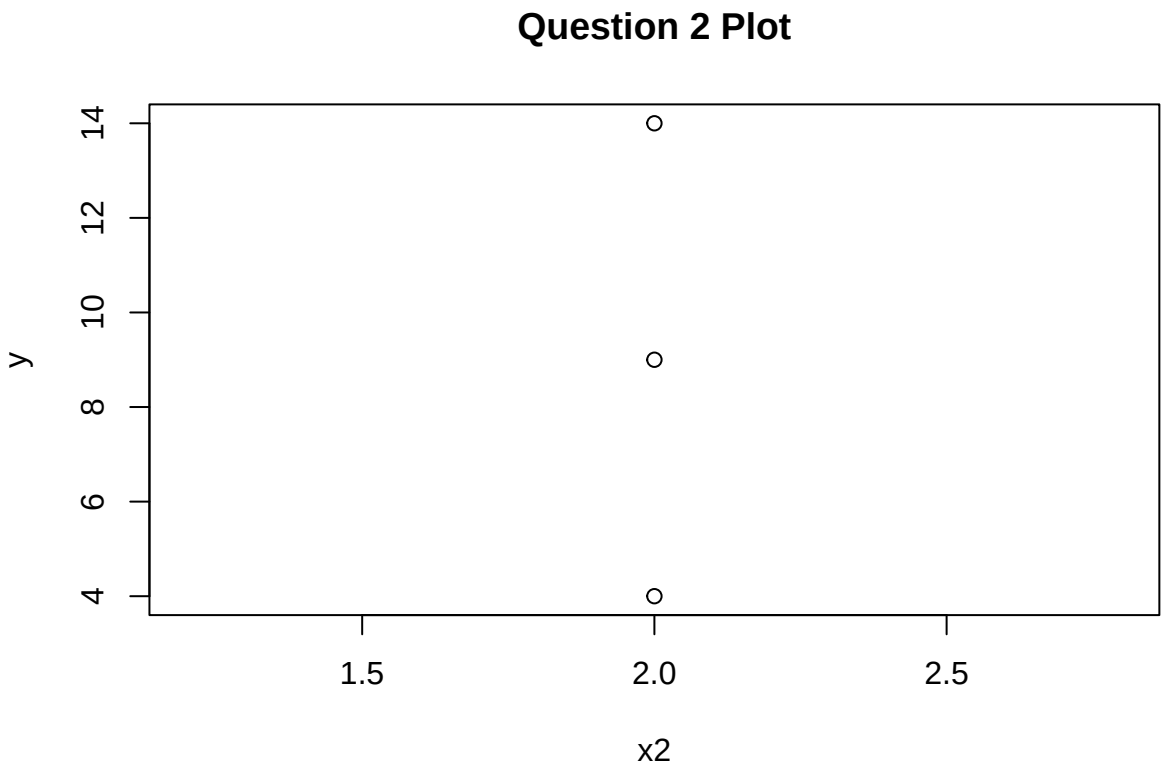
```
##      x0 x2
## [1,]  1  2
## [2,]  1  2
## [3,]  1  2
```

```
# beta2 <- solve(t(Xmat) %*% Xmat) %*% (t(Xmat) %*% Y)
# This doesn't work as proved by the lm function below

fit2 <- lm(y ~ x2)
fit2
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Coefficients:
## (Intercept)           x2
##           9           NA
```

```
plot(q2data, xlab="x2", ylab="y", main = "Question 2 Plot")
```



Question 3 Part a, b and c

a)

$y_{happiness} = 0.204 + 0.714(x_{income}) + \epsilon$

b)

On average, an increase of one unit of income which is $10k increases happiness by 0.714 on a scale of 1-10.

c)

The significance level of income shows that there is very high significance. There are 3 *** which shows that the significance is so close to 0 that it almost guarantees that whatever significance level is chosen it is guaranteed to be significant.

```
data = read.csv('happy.csv')

q3 <- lm(happiness ~ income, data = data)
summary(q3)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

```
q3$coefficients[1]
```

```
## (Intercept)
##   0.2042704
```

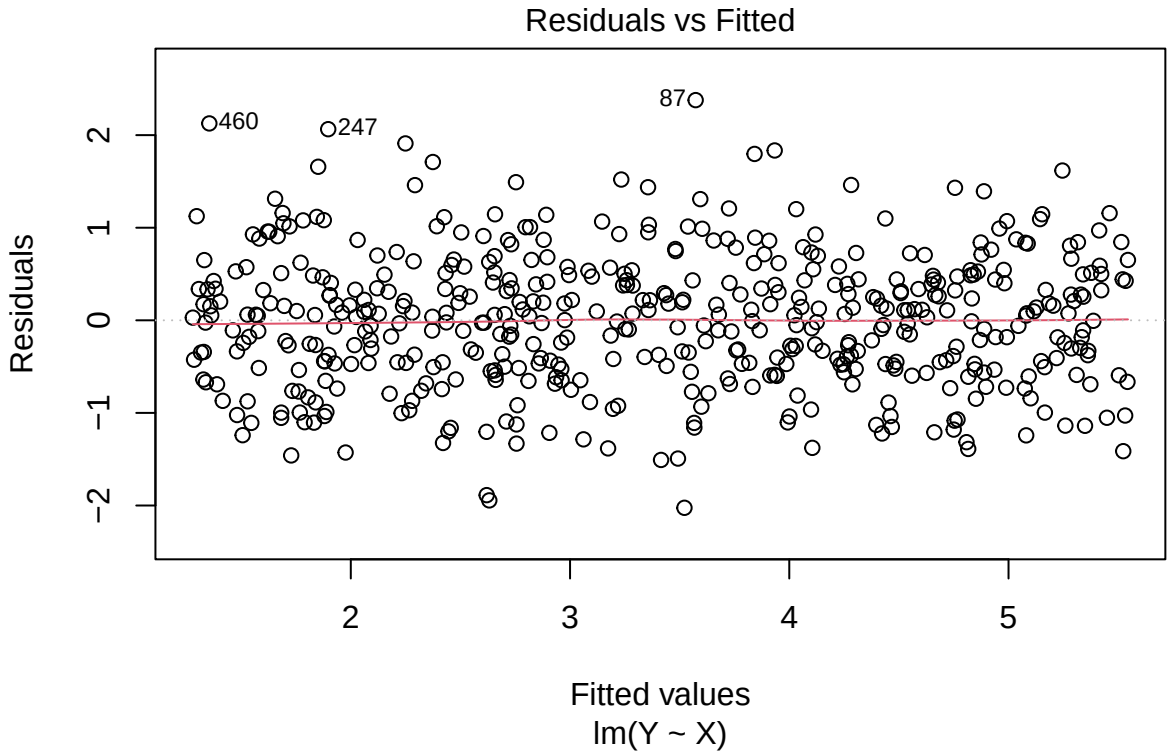
```
q3$coefficients[2]
```

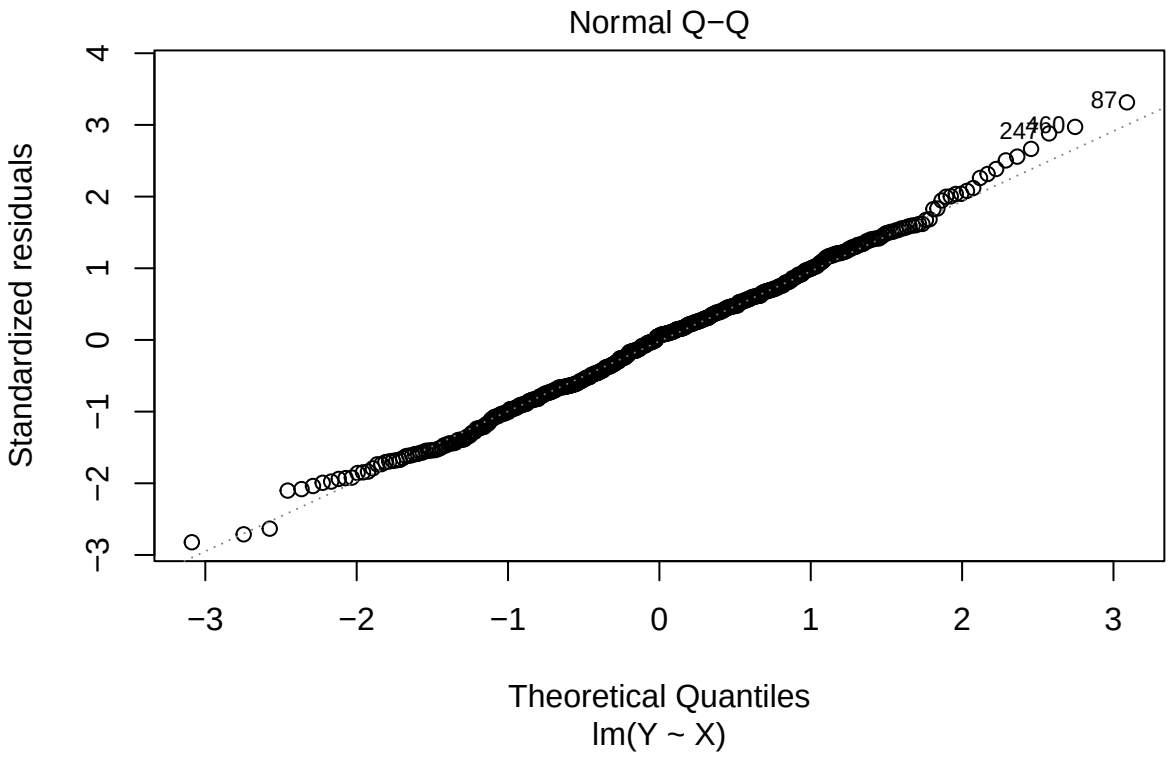
```
##    income
## 0.7138255
```

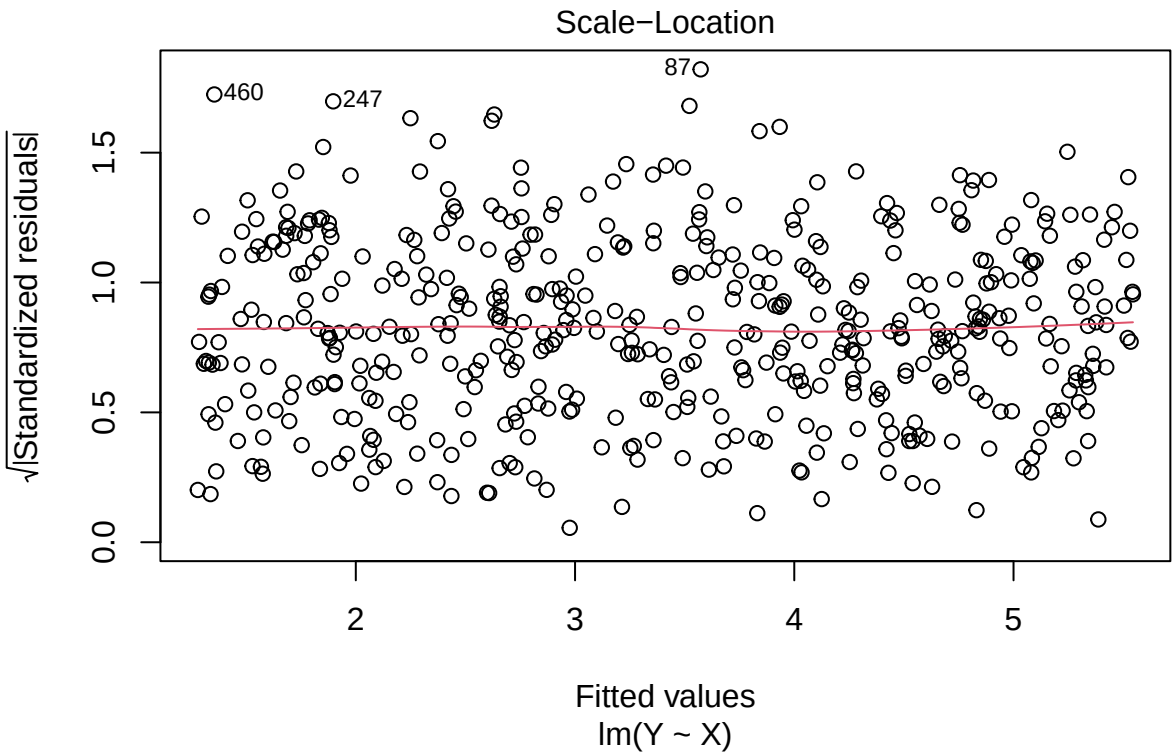
d)

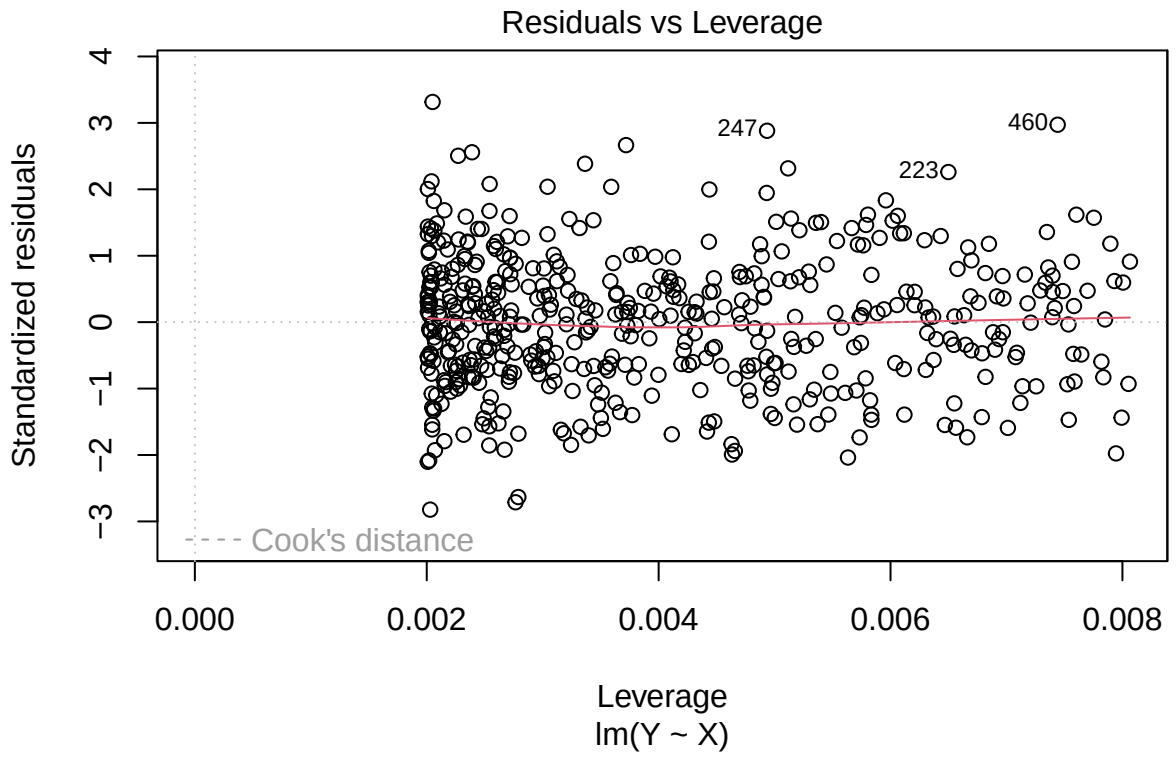
Using the adjusted R-Squared value we can see that it is 0.7488. We use the Adjusted variable rather than the multiple as it the Adjusted R-Squared normalizes Multiple R-Squared by taking into account how many samples the data includes and how many variables are being used.Further to justify this point, plotting fit shows how well the model fit as the residuals appear random centered around zero, with constant variance.

```
Y <- data$happiness
X <- data$income
fit3 <- lm(Y ~ X)
plot(fit3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X)
87
460
247
−3
−2
−1
0
1
2
3
4


Scale−Location
460
247
87
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
3
4
5
Fitted values
lm(Y ~ X)

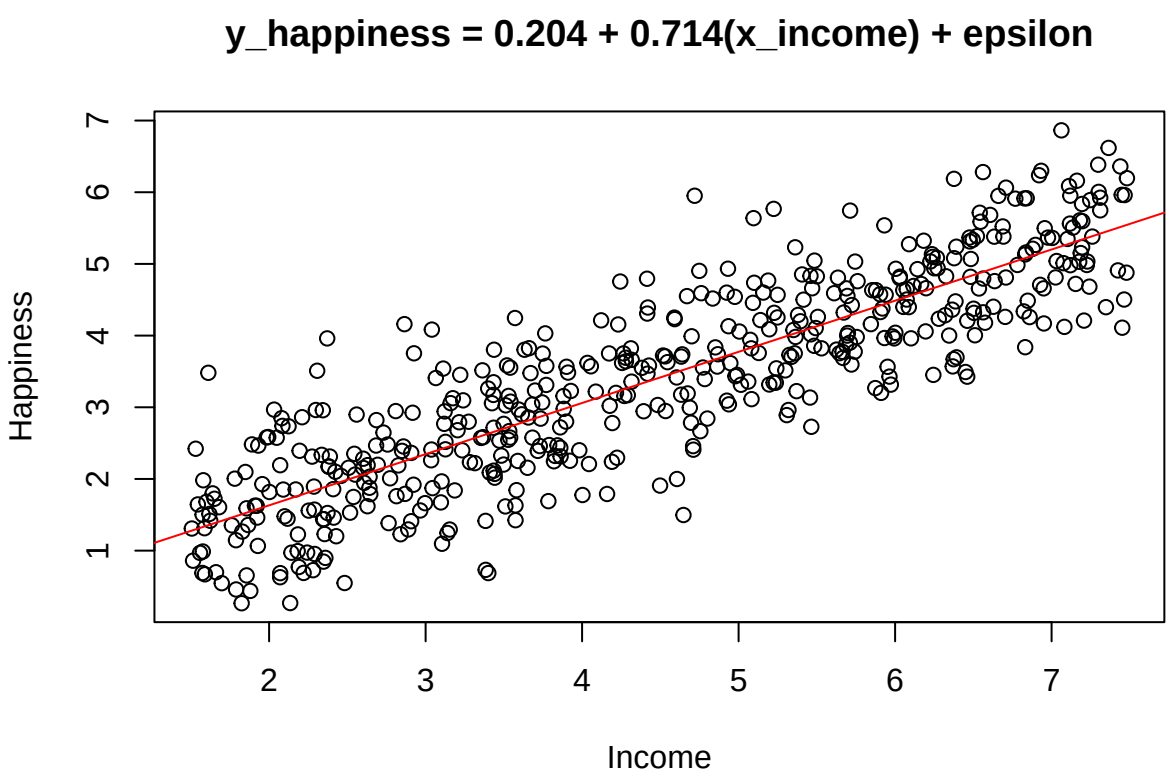


e)

Linear Regression Plot

```
x0 <- rep(1, length(Y))
ymat <- as.matrix(data$happiness)
xmat <- as.matrix(cbind(x0, data$income))
betaq3 <- solve(t(xmat) %*% xmat) %*% (t(xmat) %*% ymat)

plot(data$income, data$happiness, xlab="Income", ylab="Happiness",
     main = "y_happiness = 0.204 + 0.714(x_income) + epsilon")
abline(betaq3, col="red")
```



f)

Happiness predictions for the following income values, 2.75, 5.75, and 8.75:

an income of 2.75 has a happiness prediction of 2.17

an income of 5.75 has a happiness prediction of 4.31

an income of 8.75 has a happiness prediction of 6.45

Our data is only representative of what have, we don't know if data we don't have will follow the observed pattern which is why it wouldn't be strictly valid to make a prediction outside the income range. The max income we have is 7.48 so the previous prediction for 8.75 wouldn't be valid in this context.

```
0.2042704 + 0.7138255*2.75
```

```
## [1] 2.167291
```

```
0.2042704 + 0.7138255*5.75
```

```
## [1] 4.308767
```

```
0.2042704 + 0.7138255*8.75
```

```
## [1] 6.450244
```

```
max(data$income)
```

```
## [1] 7.481521
```