PSTAT 120C Midterm

TJ Sipin

2022-08-19

Goal

Should miles per gallon be predicted based on weight alone, or on the linear combination of weight and displacement?

```
data <- read.csv('data.csv')
```

Problems

1. Answer the following based on a *simple* linear regression, predicting *mpg* (y) with *weight* (x_1).
   a. Fit the specified model. Write the model equation.

```
(fit1 <- lm(mpg ~ weight, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = data)
##
## Coefficients:
## (Intercept)       weight
##   40.267655    -0.004678
```

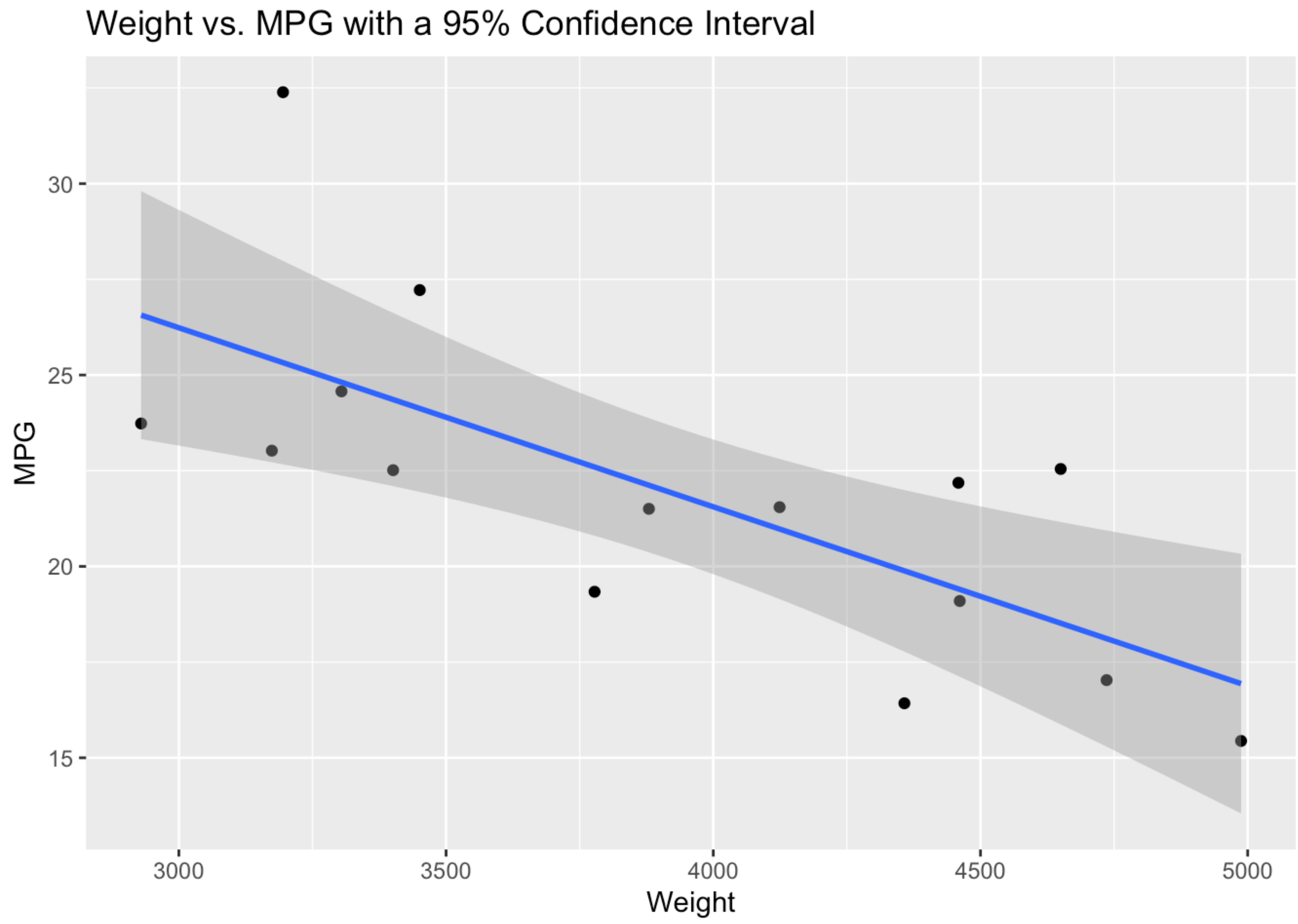
Model equation:

$$mpg = 40.267655 - 0.004678x_{weight,i} + \epsilon_i$$

b. Create a scatterplot of *mpg* and *weight*. Add a line representing the model, with 95% confidence bands. Does the model appear to fit the data?

```
ggplot(data,
       aes(x = weight,
                 y = mpg)) +
  geom_point() +
  stat_smooth(method = lm, level = 0.95) +
  xlab('Weight') +
  ylab('MPG') +
  ggtitle('Weight vs. MPG with a 95% Confidence Interval')
```

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



The model doesn't appear to fit the data too well, as the variance isn't constant.

c. Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Do not interpret the conclusion of this test.

Hypotheses: $H_0 : \beta_1 = 0, \quad H_\alpha : \beta_1 \neq 0$

Note: We want to test $H_0 : \beta_1 = a^T\beta = 0$, where $a^T = (0, 1)$.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4600 -2.1210 -0.6158  1.6716  7.0659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.267655   5.038457   7.992 2.26e-06 ***
## weight      -0.004678   0.001267  -3.692  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.131 on 13 degrees of freedom
## Multiple R-squared:  0.5119, Adjusted R-squared:  0.4744
## F-statistic: 13.63 on 1 and 13 DF,  p-value: 0.002709
```

Test statistic:

$$\begin{aligned}|T| &= \frac{a^T\hat{\beta} - 0}{S\sqrt{a^T(X^TX)^{-1}}} \\ &= 3.692\end{aligned}$$

Rejection region(s):

Note: we have $n - 2 = 15 - 2 = 13$ degrees of freedom.

$$\text{reject if } |T| \geq t_{\alpha/2,13} = t_{0.025,13} = 2.160.$$

P-value: 0.00271, according to the summary of the fit.

2. Answer the following based on a *multiple* linear regression, predicting *mpg* with *weight* (x_1) and *engine displacement* (x_2).

a. Fit the specified model. Write the model equation, including your estimates.

```
(fit2 <- lm(mpg ~ weight + displacement, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ weight + displacement, data = data)
##
## Coefficients:
##  (Intercept)        weight  displacement
##   36.5095516    -0.0003083    -0.0717513
```

Model equation:

$$mpg_i = 36.5095516 - 0.0003083x_{weight,i} - 0.0717513x_{displacement,i} + \epsilon_i$$

b. Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypothesis, test statistic, rejection region(s), and p-value. Interpret the conclusion of this test at $\alpha = 0.05$.

Hypotheses: $H_0 : \beta_1 = 0, \quad H_\alpha : \beta_1 \neq 0$

Note: We want to test $H_0 : \beta_1 = a^T\beta = 0$, where $a^T = (0, 1, 0)$.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + displacement, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1342 -0.9828 -0.6934  1.4039  5.0779
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.5095516  3.8852963   9.397 6.98e-07 ***
## weight       -0.0003083  0.0015820  -0.195    0.849
## displacement -0.0717513  0.0209294  -3.428    0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 12 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7123
## F-statistic: 18.33 on 2 and 12 DF,  p-value: 0.0002248
```

Test statistic:

$$\begin{aligned}|T| &= \frac{a^T\hat{\beta} - 0}{S\sqrt{a^T(X^TX)^{-1}}} \\ &= 0.195\end{aligned}$$

Rejection region(s):

Note: we have $n - 2 = 15 - 2 = 13$ degrees of freedom.

$$\text{reject if } |T| \geq t_{\alpha/2,13} = t_{0.025,13} = 2.160.$$

P-value: 0.849, according to the summary of the fit.

We fail to reject the null hypothesis. That is, there is insufficient evidence to support the claim that the slope of x_1, β_1, is equal to zero and that *weight* does not contribute information for the prediction of *mpg*.

c. Consider $x_1^* = 3000$ and $x_2^* = 150$. Calculate a 95% confidence interval for $\mathbb{E}[Y|x_1 = x_1^*, x_2 = x_2^*]$. Calculate a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$. Interpret both of these intervals in context.

We can use the equation to find a 95% confidence interval:

$$\mathbf{a}'\hat{\beta} \pm t_{\alpha/2}S\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$$

For the model,

$$\mathbb{E}(Y) = \beta_0 + \beta_1x_1 + \beta_2x_2 = \mathbf{a}'\beta, \qquad \mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

We're interested in the 95% confidence interval for $E[Y|x_1 = 3000, x_2 = 150]$, so we have

$$E[Y|x_1 = 3000, x_2 = 150] = \begin{bmatrix} 1 \\ 3000 \\ 150 \end{bmatrix}\beta$$

```
x <- matrix(data = c(rep(1, 15),
                     data$weight,
                     data$displacement), nrow = 15, byrow = F)
xt <- t(x)
xtx_inv <- solve(xt %*% x)
```

```
data
```

```
##    manufacturer   model_year      mpg   weight displacement class horsepower
## 1           Kia         2015 21.54716 4124.129     178.5575 Truck   237.6715
## 2          Ford         2007 17.02911 4736.041     236.0139 Truck   236.3425
## 3         Mazda         2003 19.33781 3777.898     179.4107 Truck   186.9290
## 4            GM         1986 23.02399 3174.024     190.2972   Car   115.2296
## 5           BMW         2017 22.54566 4650.112     164.4554 Truck   284.4615
## 6           Kia Prelim. 2021 32.38923 3194.868     114.4701   Car   161.8138
## 7           All         2000 22.51440 3400.909     168.2990   Car   168.2936
## 8        Nissan         2014 22.18444 4458.683     208.4433 Truck   245.0790
## 9      Mercedes         1999 21.50476 3879.585     197.3525   Car   222.9132
## 10       Subaru         2012 27.21958 3450.740     137.7964   Car   170.3608
## 11           VW         1990 23.73371 2929.358     122.0215   Car   118.9145
## 12       Toyota         1998 24.57349 3304.248     142.4937   Car   151.3054
## 13       Toyota         2007 19.09633 4461.215     218.8619 Truck   229.8358
## 14           GM         2002 15.44052 4987.675     302.1571 Truck   257.2493
## 15         Ford         1988 16.42429 4357.654     239.6896 Truck   149.6339
```

Through the fit equation, we have that

$$\hat{\beta} \approx \begin{bmatrix} 36.5 \\ -0.000308 \\ -0.0718 \end{bmatrix}.$$

Additionally, $t_{\alpha/2,13} = 2.160$ and $S = $ 2.2254756.

Substituting each element into the equation

$$\mathbf{a}'\hat{\beta} \pm t_{\alpha/2} S \sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}},$$

we get

```
a <- c(1, 3000, 150) %>% as.matrix()
at <- t(a)
s <- ((sse <- sum((fitted(fit2) - data$mpg)^2))/(15-2)) %>% sqrt()
bhat <- c(36.5, -0.000308, -0.0718) %>% as.matrix()
at %*% bhat + 2.16* s * sqrt(at %*% xtx_inv %*% a)
```

```
##          [,1]
## [1,] 27.15418
```

```
at %*% bhat - 2.16* s * sqrt(at %*% xtx_inv %*% a)
```

```
##          [,1]
## [1,] 22.45782
```

The 95% confidence interval for $\mathbb{E}[Y|x_1 = 3000, x_2 = 150]$ is $(22.458, 27.154)$.

Similarly, the prediction interval can be given by

$$\mathbf{a}'\hat{\beta} \pm t_{\alpha/2} S \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$$

```
at %*% bhat + 2.16* s * sqrt(1 + at %*% xtx_inv %*% a)
```

```
##         [,1]
## [1,] 30.1559
```

```
at %*% bhat - 2.16* s * sqrt(1 + at %*% xtx_inv %*% a)
```

```
##         [,1]
## [1,] 19.4561
```

The above code yields the prediction interval to be $(19.456, 30.156)$.

This means that when *weight* is 3000 and *engine displacement* is 150, then the mean value of *mpg* with 95% confidence is between 22.458 and 27.154 miles per gallon. On the other hand, with those same values of *weight* and *engine displacement*, the predicted value of *mpg* with 95% confidence is between 19.456 and 30.156.

d. Which model constitutes the "complete" model and which the "reduced" model? Can x_2 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

Our hypotheses:

$$H_0 : \beta_2 = 0, \qquad H_\alpha : \beta_2 \neq 0$$

The complete model:

$$mpg_i = \beta_0 + \beta_1 x_{weight,i} + \beta_2 x_{displacement,i} + \epsilon_i$$

The reduced model:

$$mpg_i = \beta_0 + \beta_1 x_{weight,i} + \epsilon_i$$

We can perform an F test:

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SEE_C/(n-k-1)}$$

First, we need to find the SSE of both. We use an ANOVA table:

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value   Pr(>F)
## weight     1 133.66 133.665  13.634 0.002709 **
## Residuals 13 127.44   9.803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: mpg
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## weight        1 133.665 133.665  24.912 0.0003139 ***
## displacement  1  63.060  63.060  11.753 0.0050018 **
## Residuals    12  64.386   5.365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that $SSE_C = 64.386$ and $SSE_R = 127.44$.

$$\begin{aligned} F &= \frac{(SSE_R - SSE_C)/(k-g)}{SEE_C/(n-k-1)} \\ &= \frac{(127.44 - 64.386)/(2-1)}{64.386/(15-2-1)} \\ &= 11.75175 \end{aligned}$$

The test statistic follows an F distribution with $df_1 = 1, df_2 = 12$ under the null hypothesis. The p-value is 0.005, which is less than 0.05, so we reject the null. That is, there is insufficient evidence to say that we can drop x_2 from the model without losing predictive information.

3. Consider your answers to the previous questions, then answer the following. Suppose that the true population relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Further suppose that there is a relationship between x_1 and x_2 given by:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

where γ_1 and β_2 are non-zero.

a. Find the expected values of β_0 and β_1 if the independent variable x_2 is omitted from the regression.

We have $\hat{\beta}_0$ and $\hat{\beta}_1$ defined as follows:

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\end{aligned}
$$

So, we have

$$

$$
\begin{aligned}
\mathbb{E}\hat{\beta}_0 &= \mathbb{E}\left[\bar{y} - \hat{\beta}_1\bar{x}\right] \\
\mathbb{E}\hat{\beta}_1 &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i - (\beta_0 + \beta_1\bar{x} + \bar{\epsilon}_i))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 x_i + \epsilon_i - (\beta_1\bar{x} + \bar{\epsilon}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1 x_i - \beta_1\bar{x} + \epsilon_i - \bar{\epsilon}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1(x_i - \bar{x}) + (\epsilon_i - \bar{\epsilon}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}\beta_1(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^{n}\beta_1(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] + \mathbb{E}\left[\frac{\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \beta_1 \\
\implies \mathbb{E}\hat{\beta}_0 &= \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}_1\bar{x}] \\
&= \mathbb{E}\left[\beta_0 + \beta_1\bar{x} + \bar{\epsilon} - \beta_1\bar{x}\right] \\
&= \mathbb{E}\left[\beta_0 + \bar{\epsilon}\right] \\
&= \mathbb{E}\beta_0 + 0 \\
&= \beta_0
\end{aligned}
$$

$$

b. Calculate the bias (if any) of β_0 and β_1 when x_2 is omitted.

Since $\mathbb{E}\hat{\beta}_0 = \beta_0$ and $\mathbb{E}\hat{\beta}_1 = \beta_1$, we say they are both unbiased (bias $= 0$.)

c. What values of γ_1 and β_2 would result in β_0 and β_1 remaining unbiased?

Since we have

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2(\gamma_0 + \gamma_1 x_1 + \delta) + \epsilon \\
&= \beta_0 + \beta_1 x_1 + \beta_2\gamma_0 + \beta_2\gamma_1 x_1 + \beta_2\delta + \epsilon,
\end{aligned}
$$

we want to find some choices β_2 and γ_1 that equate the above equation to $y = \beta_0 + \beta_1 x_1 + \epsilon$.

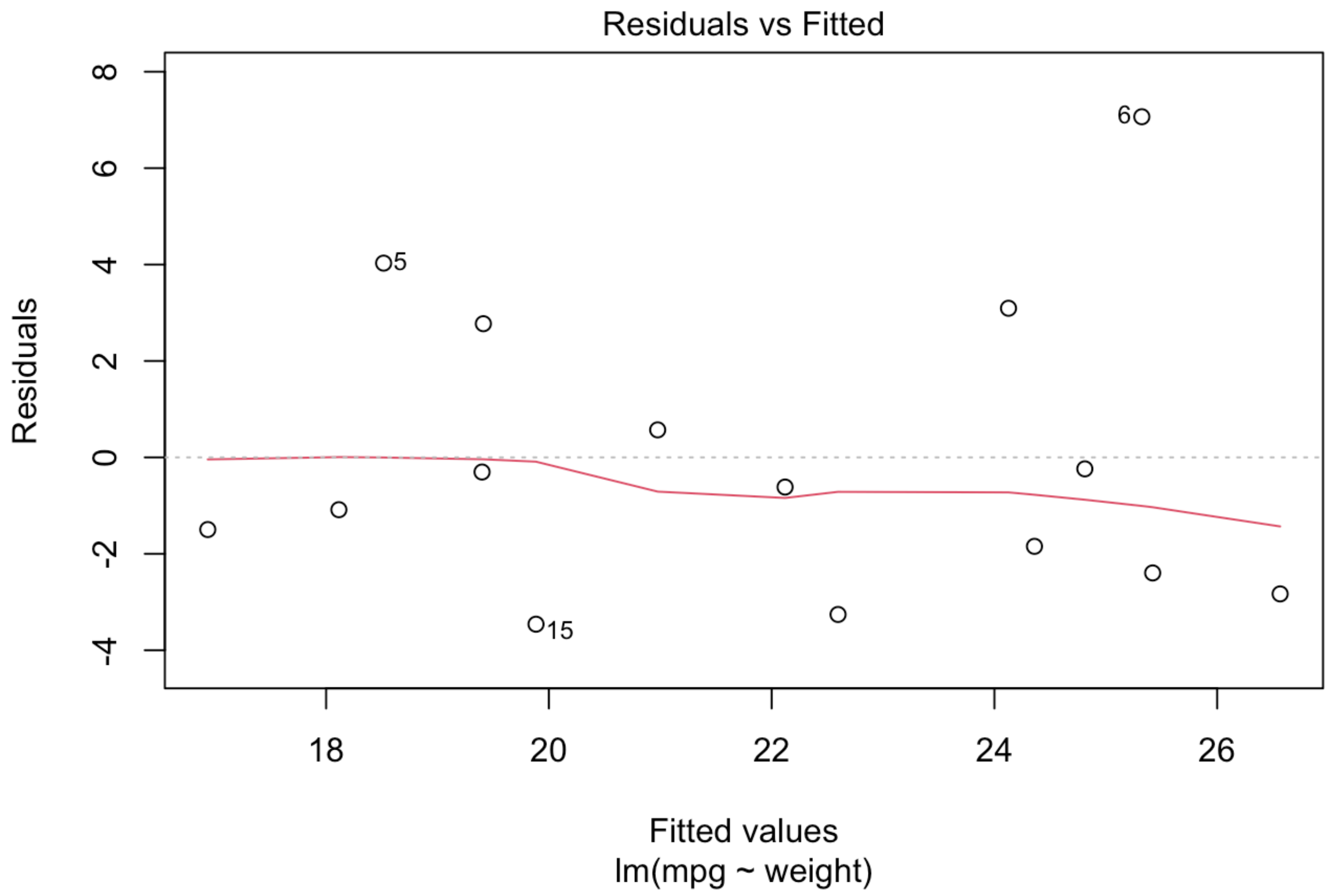
A clear choice is $\beta_2 = 0$. We can also set $\beta_2(\gamma_0 + \gamma_1 x_1 + \delta) = 0$. We find that the choice of γ_1 that satisfies this is $\gamma_1 = -\frac{\gamma_0+\delta}{x_1}$

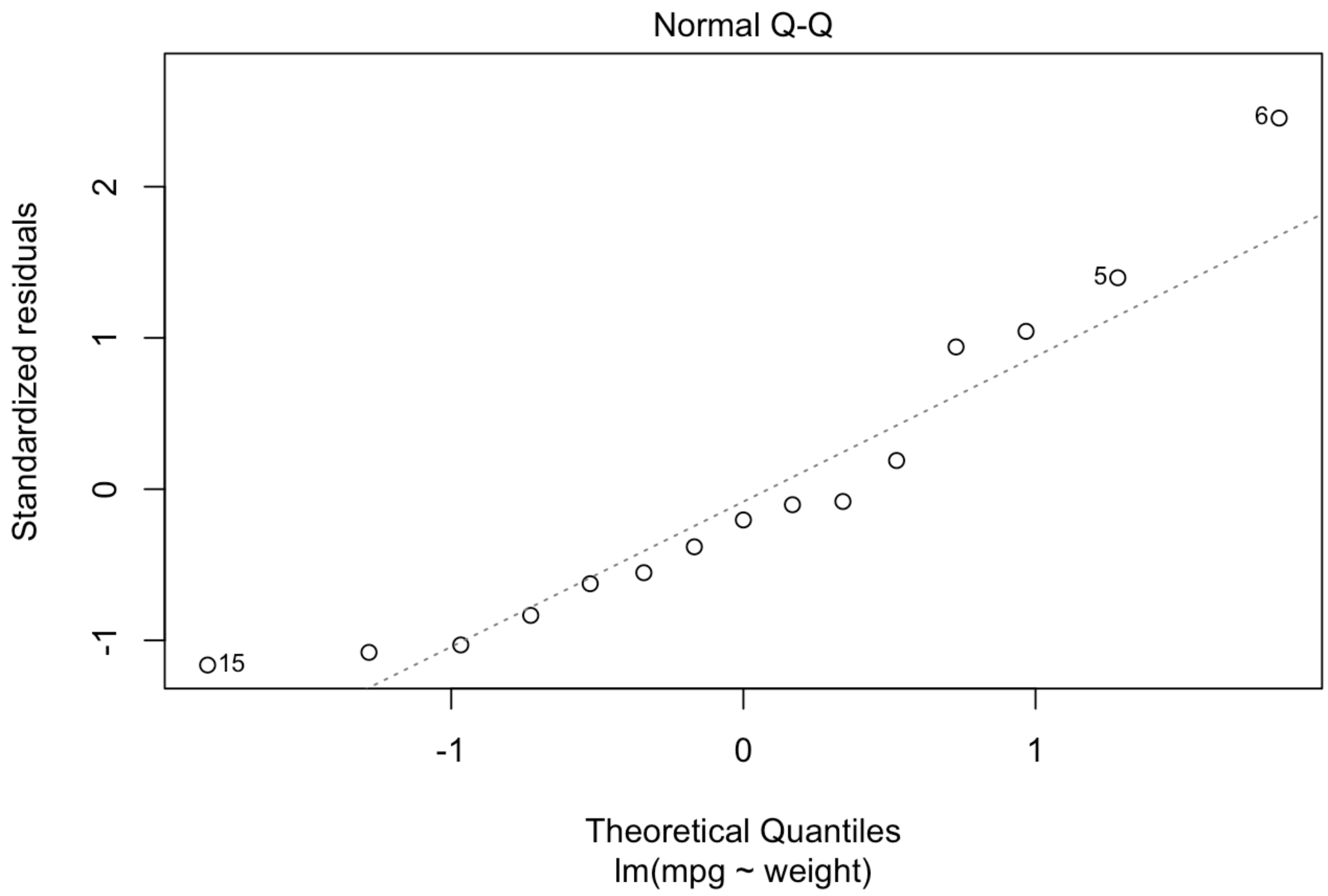
d. In light of the above:

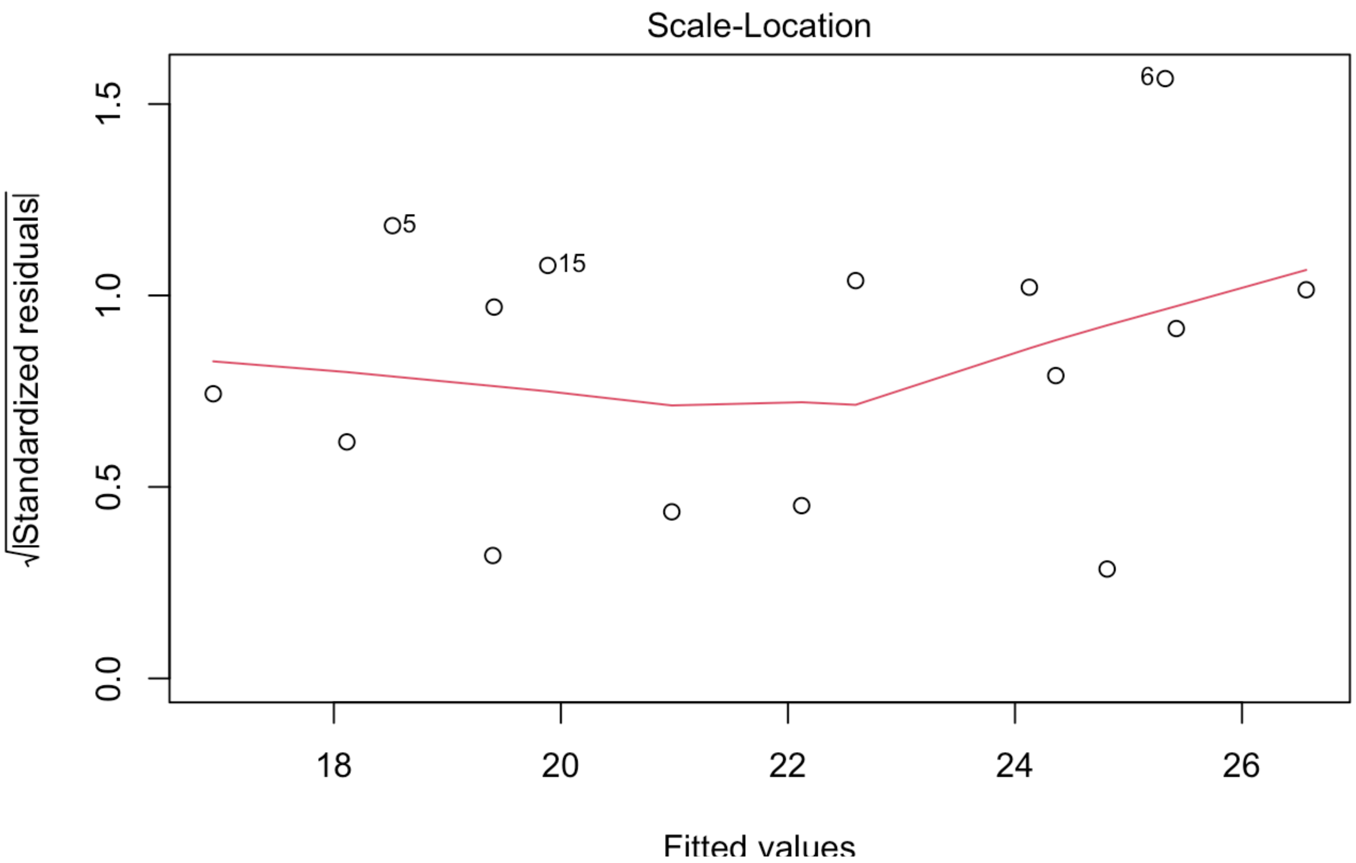
i. What assumption of linear regression is being violated in Question 1? Is this assumption met in Question 2?

The assumption of homoscedasticity is violated in Question 1. After removing outliers, the assumption may be considered to not be violated in Question 2, however the answer may differ depending on the analyst.

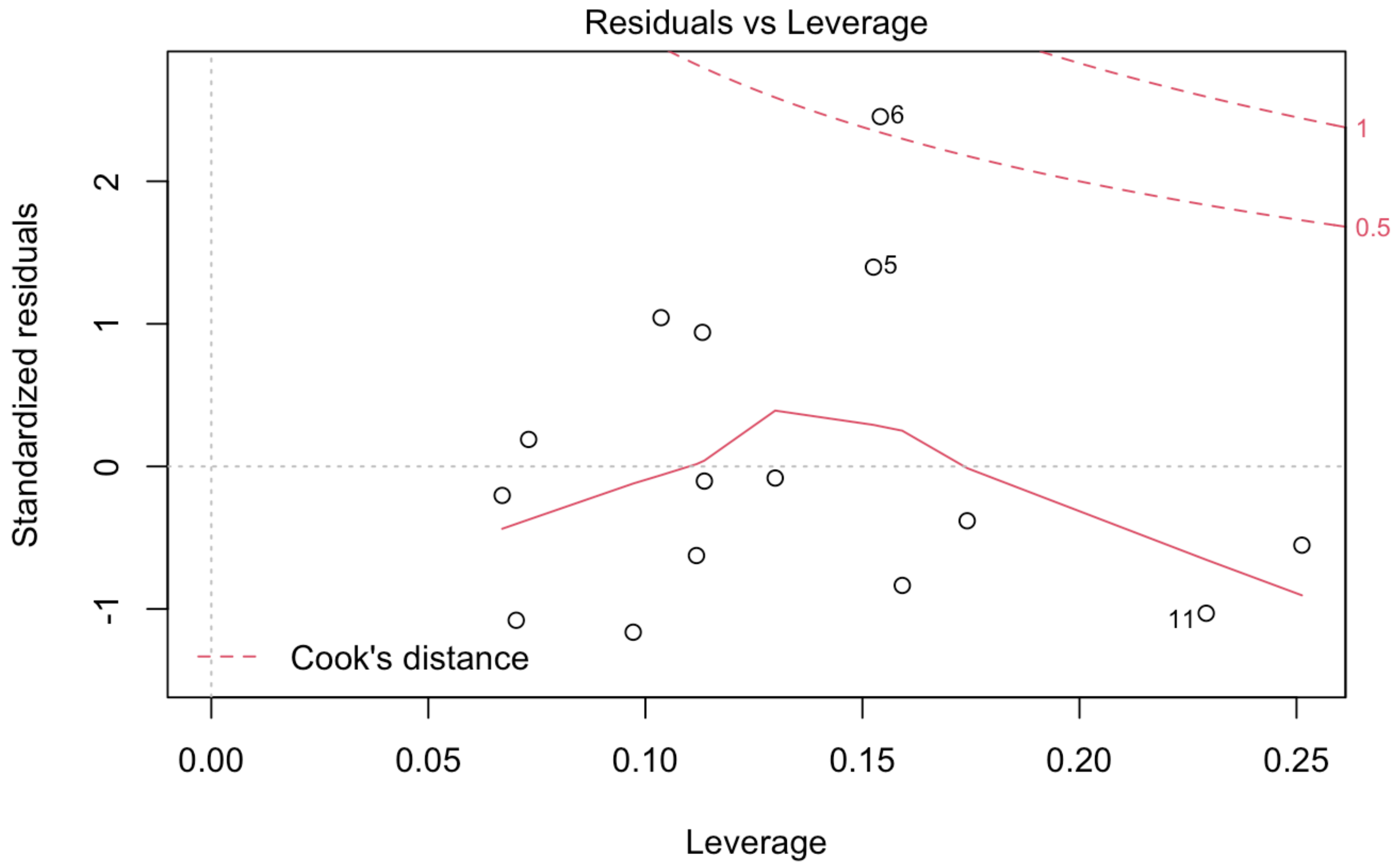
```
fit1 %>% plot()
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ weight)
6
5
15
8
6
4
2
0
-2
-4
18
20
22
24
26

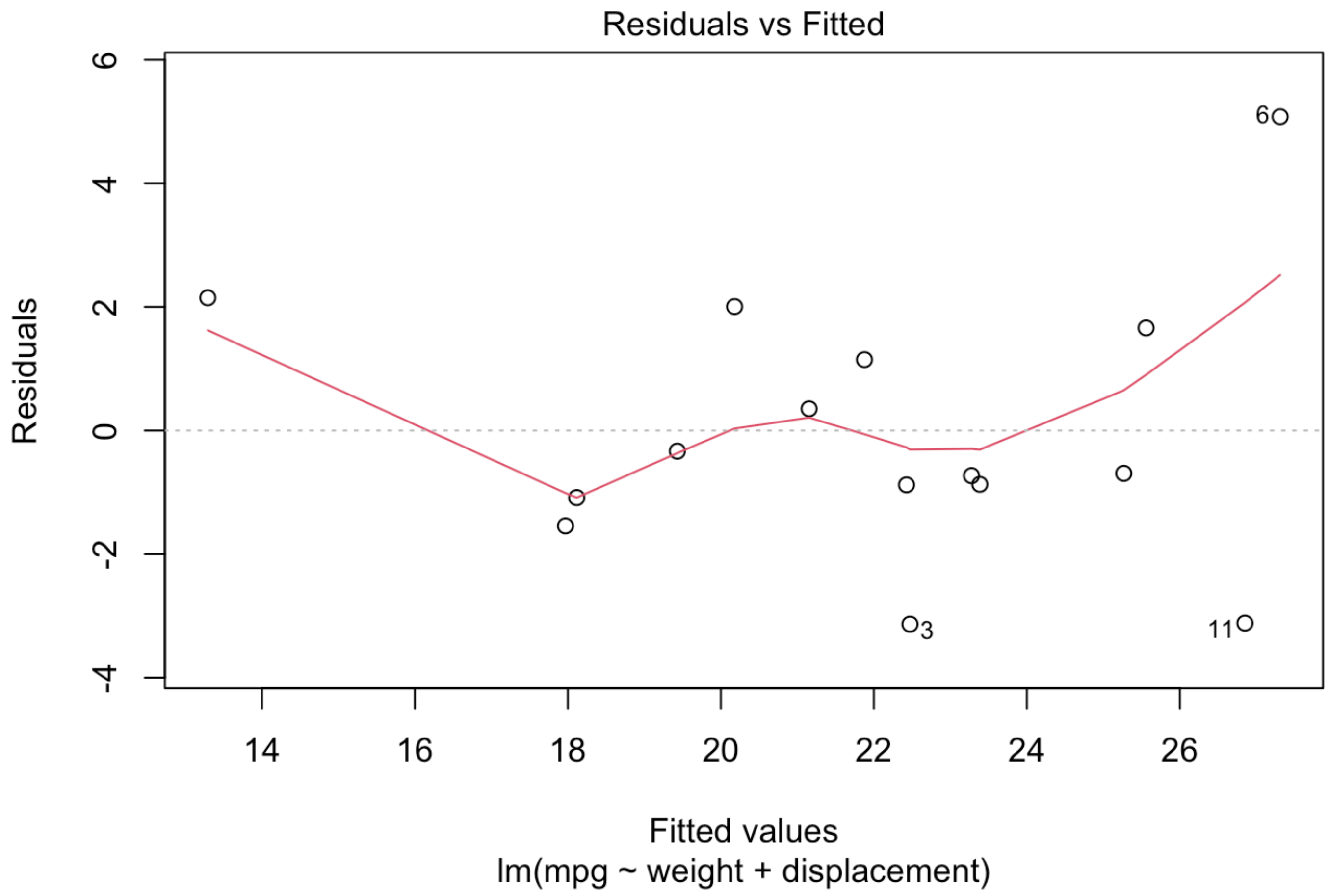

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ weight)
6
5
15
2
1
0
-1
-1
0
1

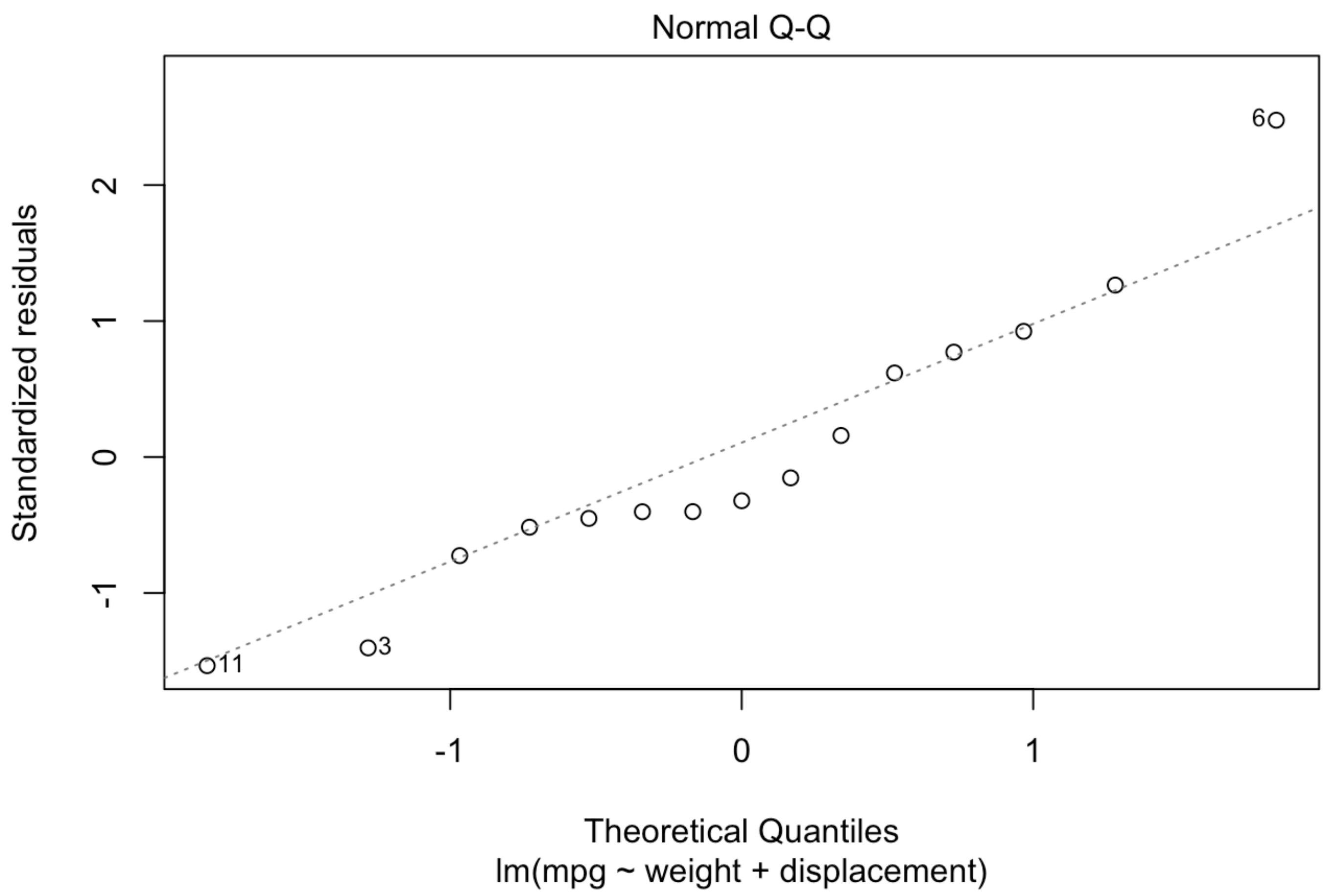

Scale-Location
√|Standardized residuals|
Fitted values
6
5
15
1.5
1.0
0.5
0.0
18
20
22
24
26

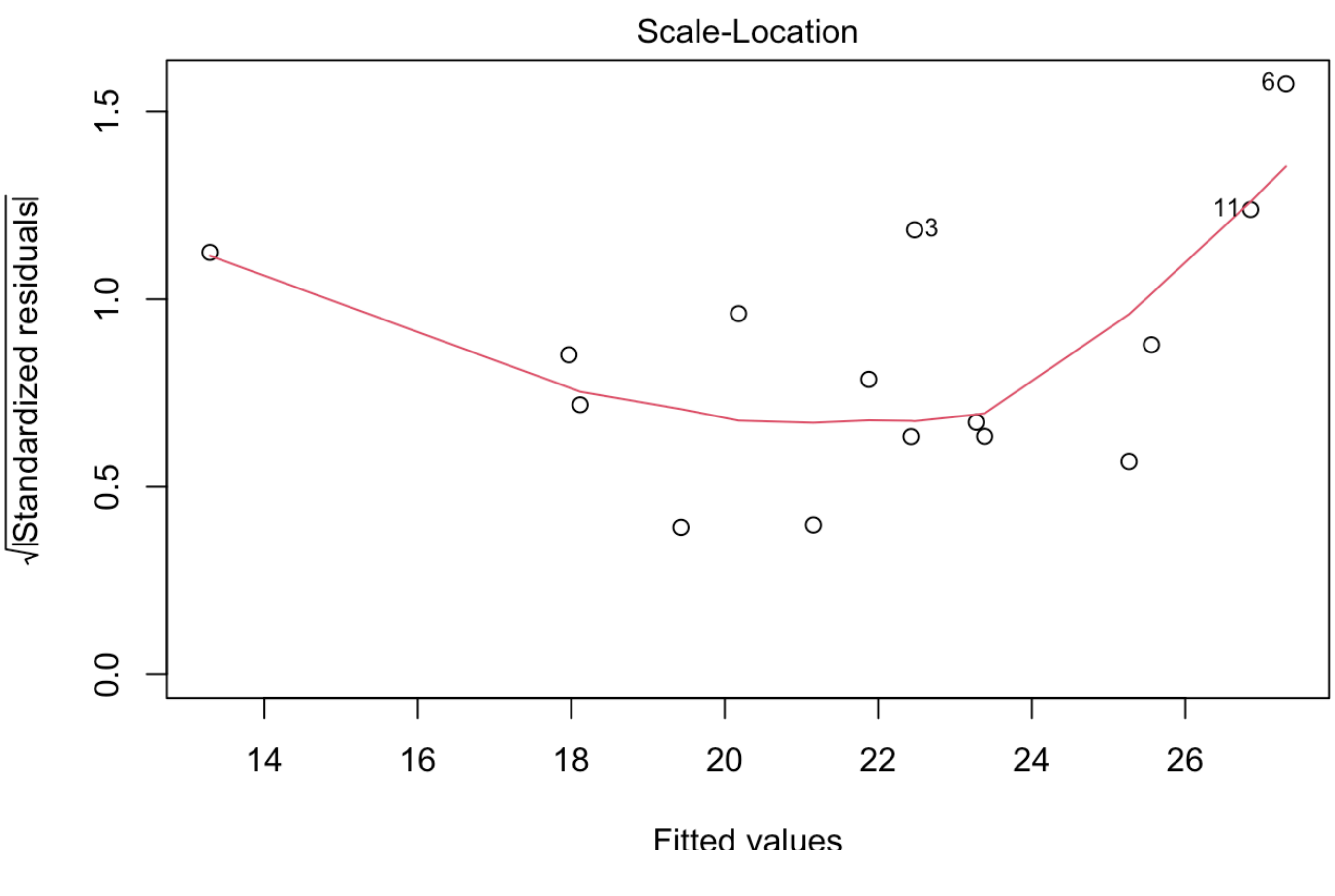
Fitted values
lm(mpg ~ weight)



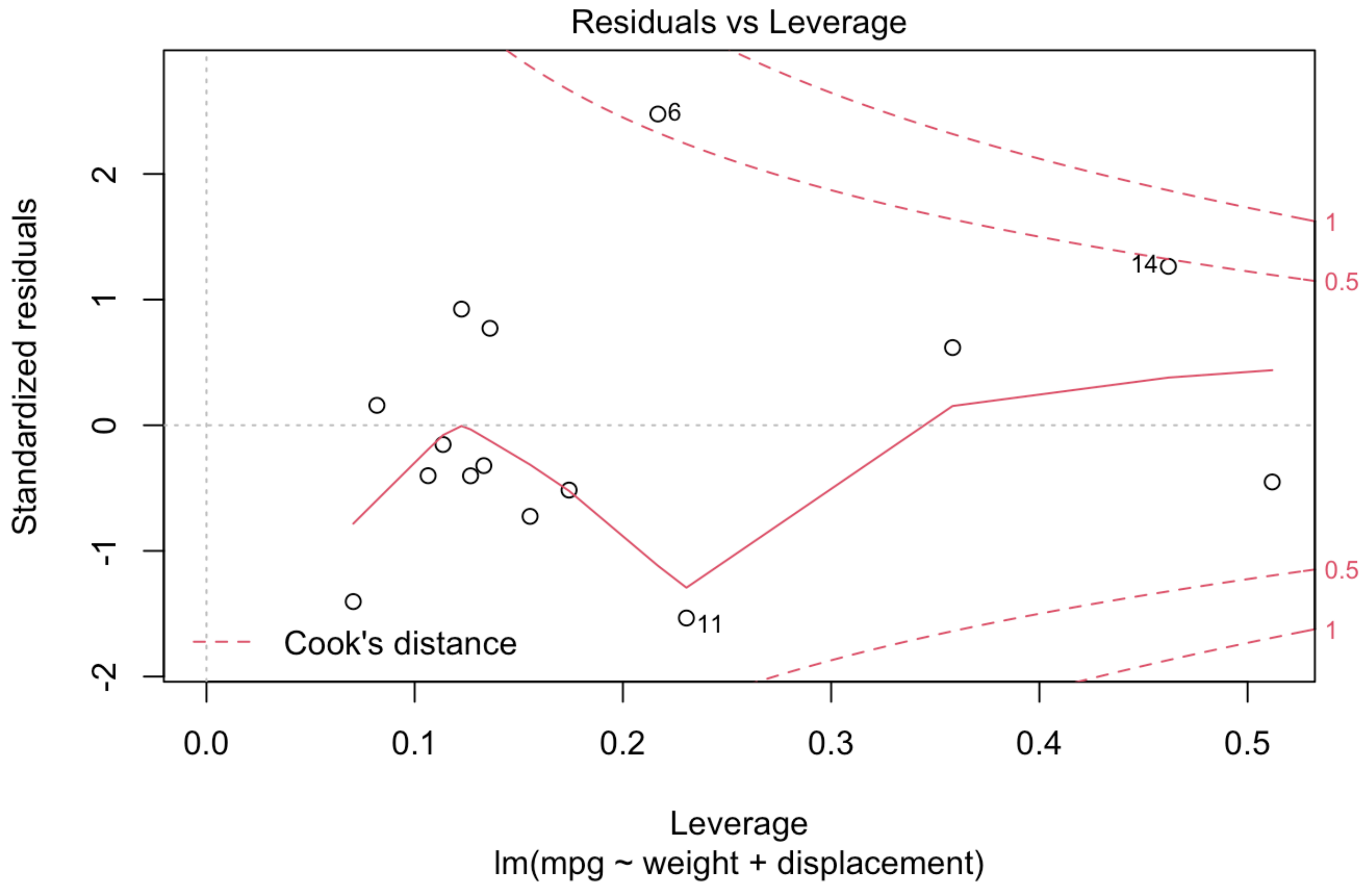

```
fit2 %>% plot()
```


Residuals vs Fitted
Residuals
6
4
2
0
-2
-4
14
16
18
20
22
24
26
Fitted values
lm(mpg ~ weight + displacement)
6
3
11


Normal Q-Q
Standardized residuals
2
1
0
-1
-1
0
1
Theoretical Quantiles
lm(mpg ~ weight + displacement)
6
3
11


Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
14
16
18
20
22
24
26
Fitted values
6
3
11

lm(mpg ~ weight + displacement)



ii. In Question 1, are the estimates of β_0 and β_1 BLUE? Why or why not?

The estimates of β_0 and β_1 are not BLUE since the assumption of homoscedasticity is not met.

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
data <- read.csv('data.csv')
(fit1 <- lm(mpg ~ weight, data = data))
ggplot(data,
       aes(x = weight,
                 y = mpg)) +
  geom_point() +
  stat_smooth(method = lm, level = 0.95) +
  xlab('Weight') +
  ylab('MPG') +
  ggtitle('Weight vs. MPG with a 95% Confidence Interval')
summary(fit1)
(fit2 <- lm(mpg ~ weight + displacement, data = data))
summary(fit2)
x <- matrix(data = c(rep(1, 15),
                     data$weight,
                     data$displacement), nrow = 15, byrow = F)
xt <- t(x)
xtx_inv <- solve(xt %*% x)
data
a <- c(1, 3000, 150) %>% as.matrix()
at <- t(a)
s <- ((sse <- sum((fitted(fit2) - data$mpg)^2))/(15-2)) %>% sqrt()
bhat <- c(36.5, -0.000308, -0.0718) %>% as.matrix()
at %*% bhat + 2.16* s * sqrt(at %*% xtx_inv %*% a)
at %*% bhat - 2.16* s * sqrt(at %*% xtx_inv %*% a)
at %*% bhat + 2.16* s * sqrt(1 + at %*% xtx_inv %*% a)
at %*% bhat - 2.16* s * sqrt(1 + at %*% xtx_inv %*% a)
anova(fit1)
anova(fit2)
fit1 %>% plot()
fit2 %>% plot()
```