Quiz 2

Your name

2024-10-21

Quiz 2

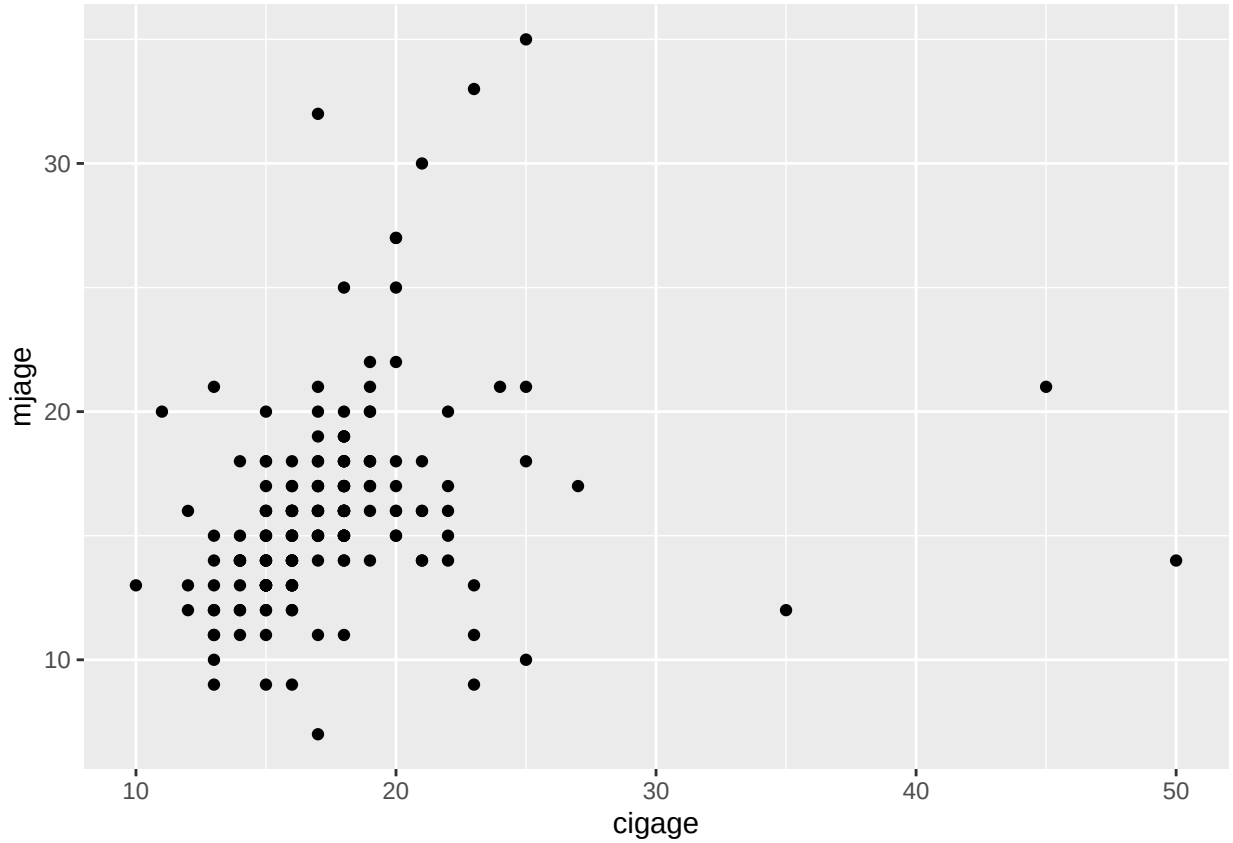
```
library(tidyverse)
```

1. Load the data called dat.nsduh.small.csv.

```
dat <- read_csv("data/dat.nsduh.small.csv")
```

2. Do visual EDA for x=cigage and y=mjage, together. Describe the scatterplot.

```
dat %>% ggplot(aes(x=cigage, y=mjage)) + geom_point()
```



Answer:

3. Fit a linear model of mjage vs. cigage. What are the model's estimated parameters for intercept and slope?

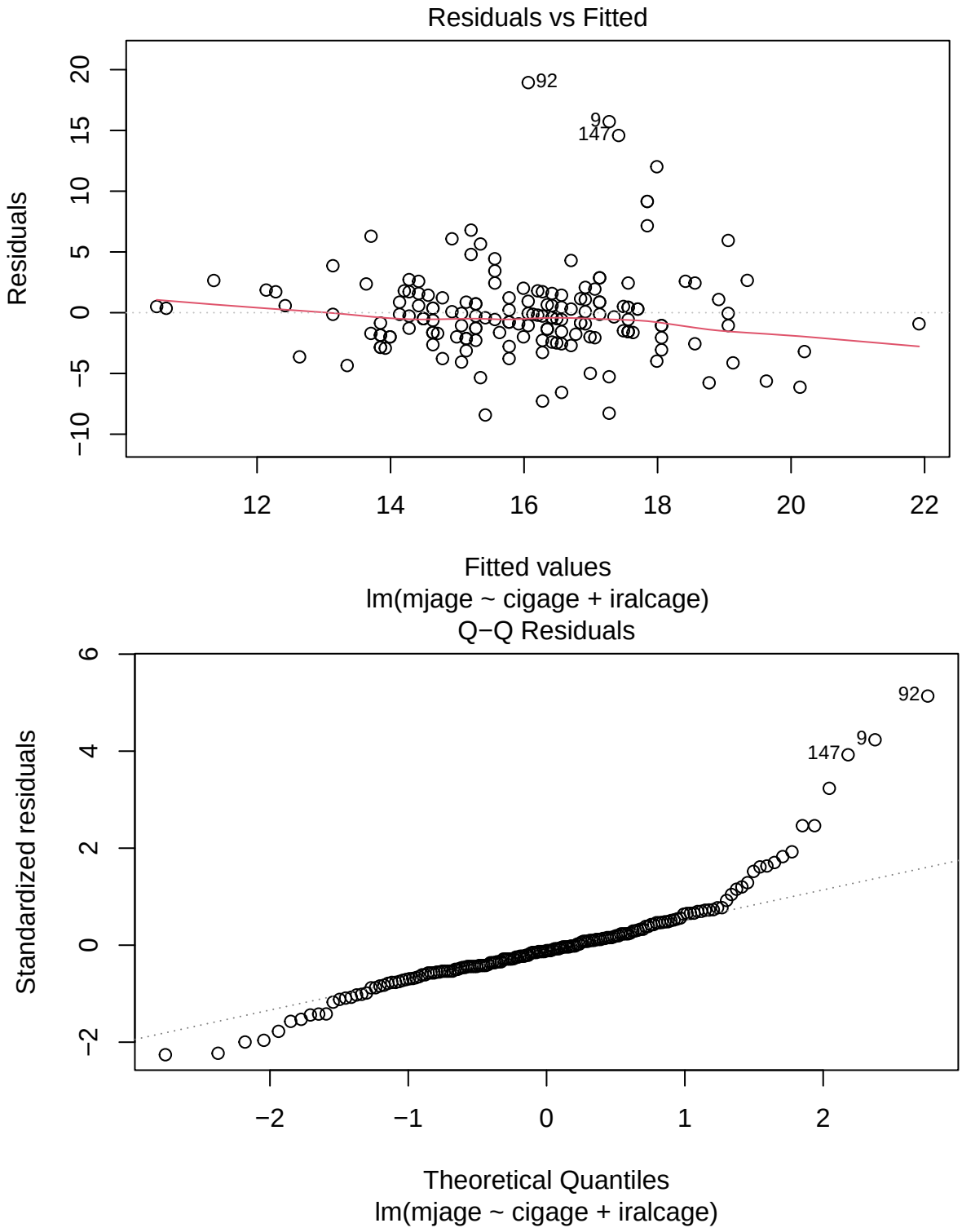
```
out <- lm(mjage ~ cigage + iralcage, data = dat)
summary(out)
```

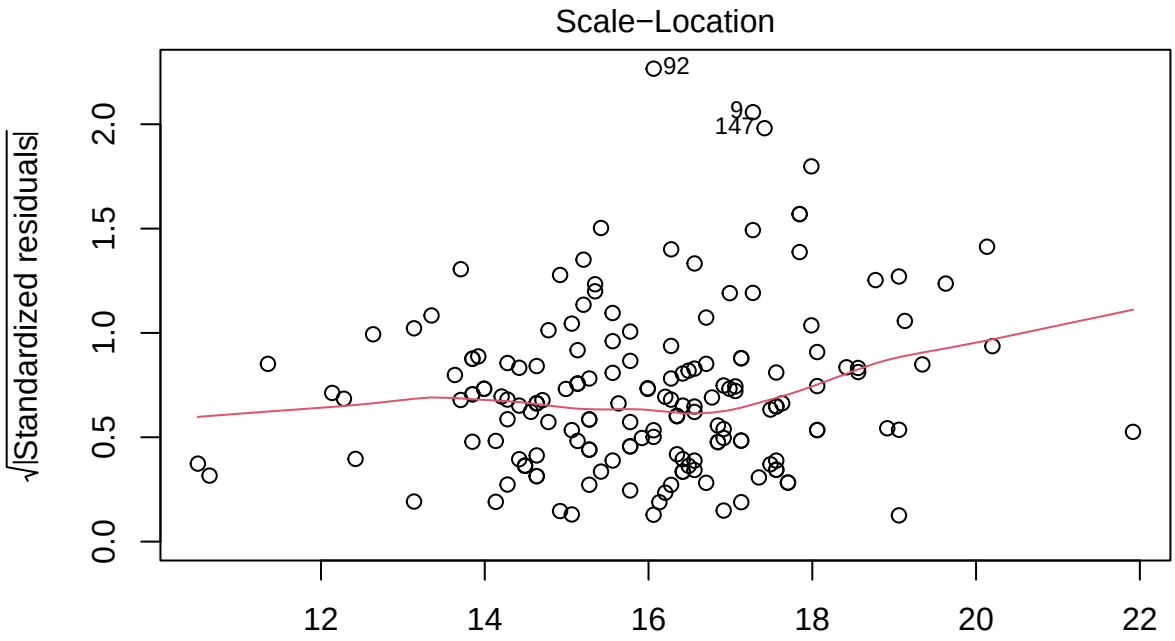
```
##
## Call:
## lm(formula = mjage ~ cigage + iralcage, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4195 -1.9189 -0.4195  1.1879 18.9370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.99942    1.63902   3.660 0.000337 ***
## cigage       0.14286    0.06451   2.215 0.028136 *
## iralcage     0.49940    0.09838   5.076 1.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.739 on 168 degrees of freedom
## Multiple R-squared:  0.1882, Adjusted R-squared:  0.1786
## F-statistic: 19.48 on 2 and 168 DF,  p-value: 2.465e-08
```

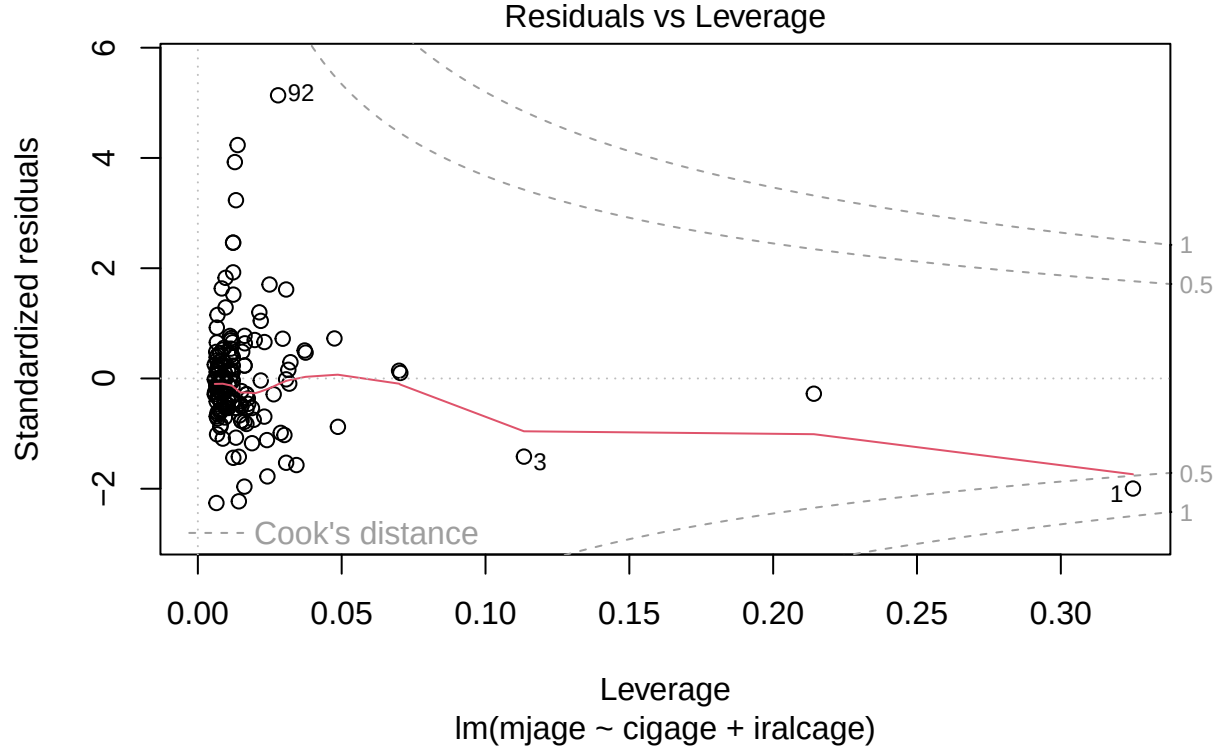
Answer:

4. Are the assumptions of the linear model satisfied?

```
plot(out)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mjage ~ cigage + iralcage)
92
9
147
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mjage ~ cigage + iralcage)

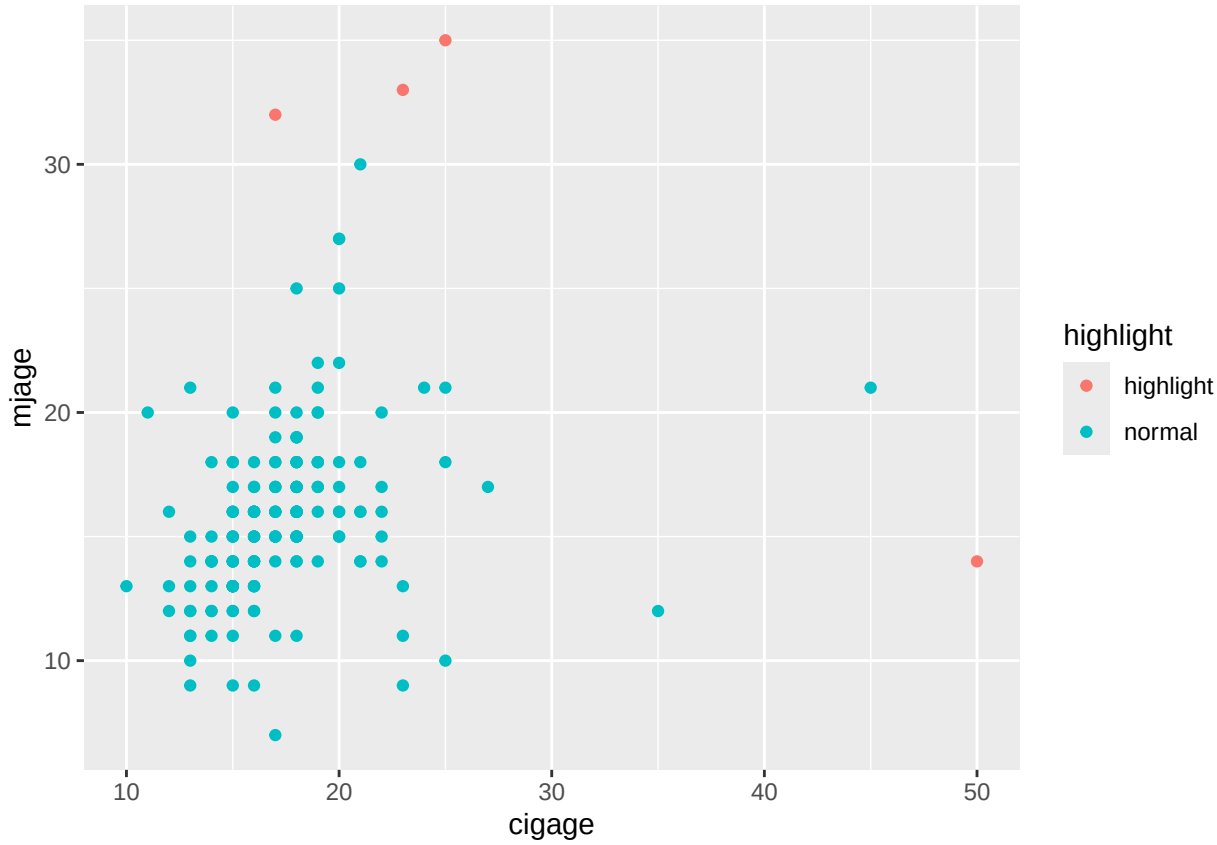

Scale−Location
92
9
147
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
12
14
16
18
20
22
Fitted values
lm(mjage ~ cigage + iralcage)


Residuals vs Leverage
92
3
1
Cook's distance
1
0.5
0.5
1
Standardized residuals
6
4
2
0
−2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(mjage ~ cigage + iralcage)

Answer:

5. Are the outliers from the diagnostic plots visible in the scatterplot? This code helps you see the outliers in the scatterplot. Write down their id's here, where it says 1,2,3,4,5.

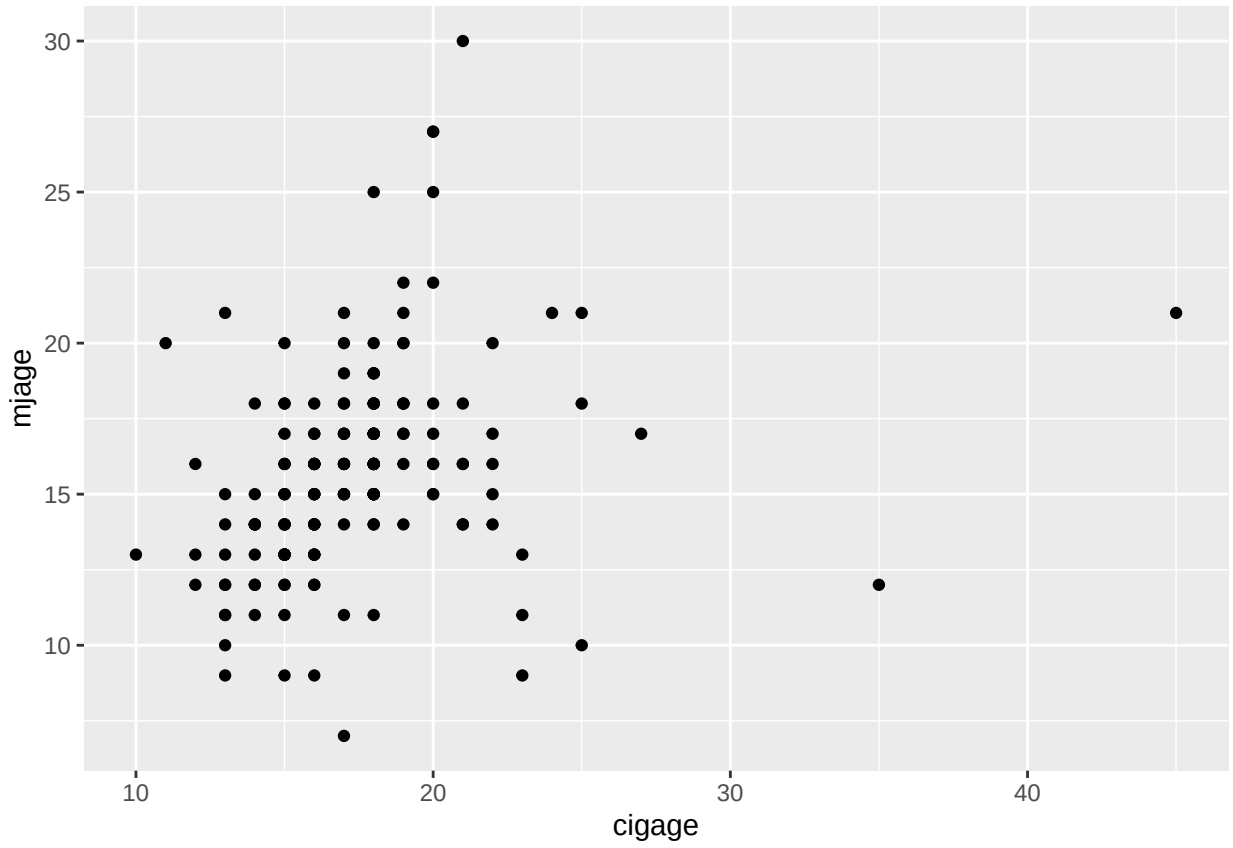
```
dat %>% mutate(highlight = ifelse(row_number() %in% c(1,9,92,147), "highlight", "normal")) %>%
  ggplot(aes(x=cigage, y=mjage)) + geom_point(aes(colour = highlight))
```



Answer:

6. Make a new dataframe without outliers, call it dat.no.outliers. See how the scatterplot changes for cigage and mjage in dat.no.outliers.

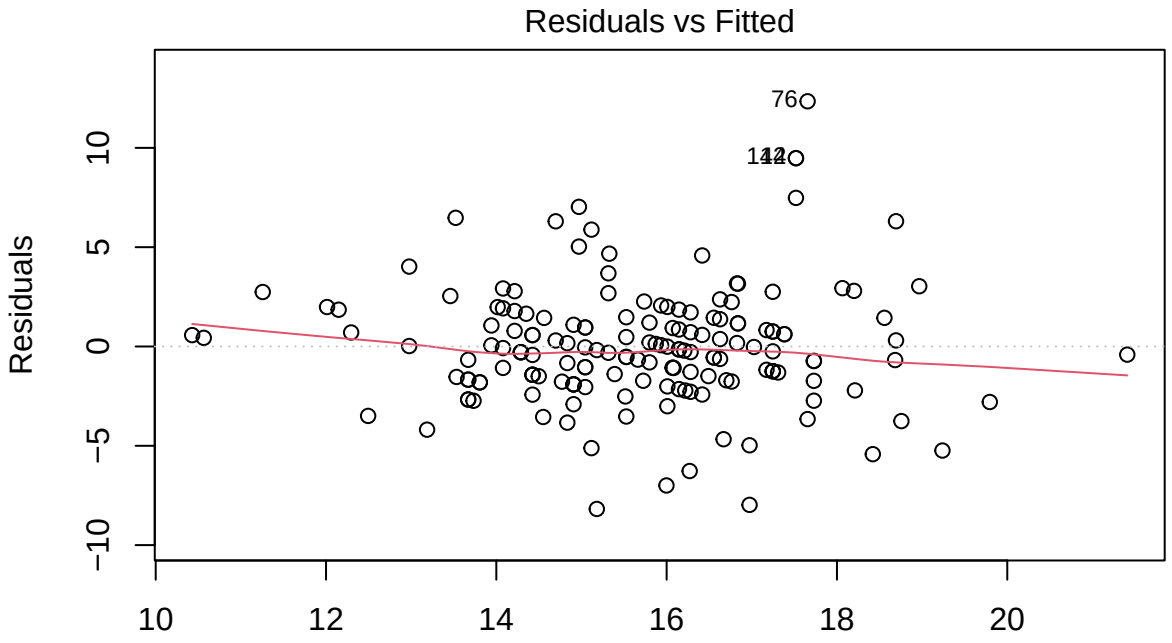
```
# make new dataset without outliers
`%not_in%` <- purrr::negate(`%in%`) # this line defines a new command not_in.
dat.no.outliers <- dat %>% filter(id %not_in% c(1,9,92,147))

dat.no.outliers %>% ggplot(aes(x=cigage, y=mjage)) + geom_point()
```

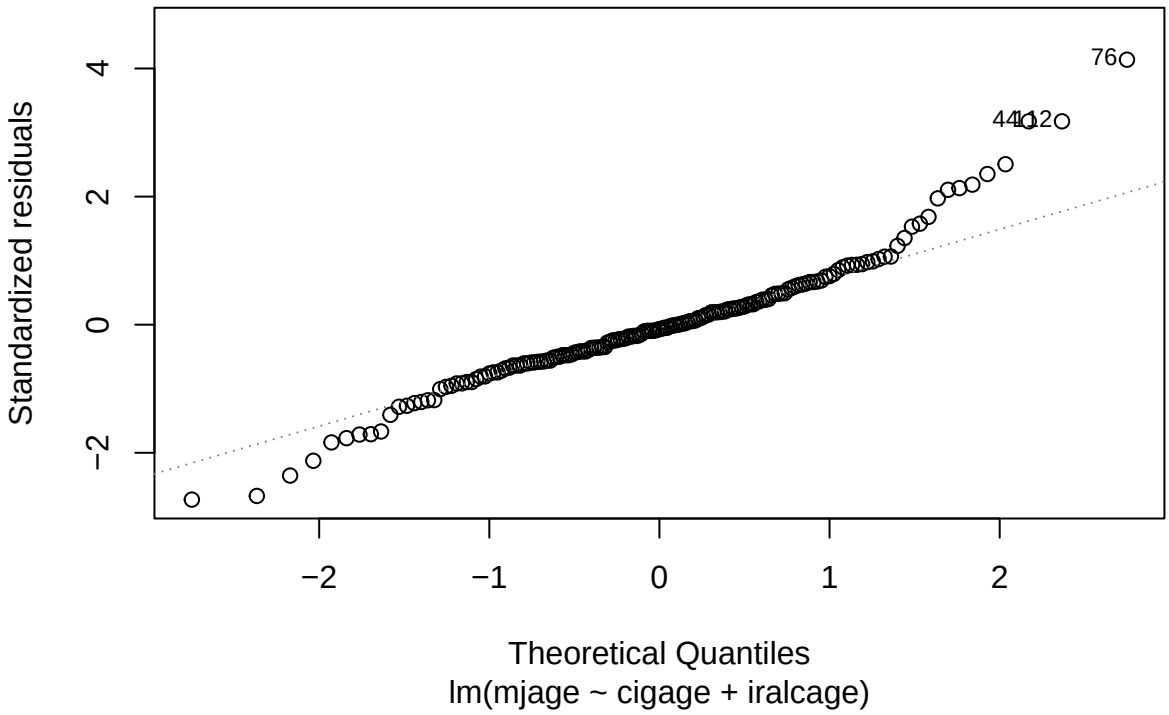


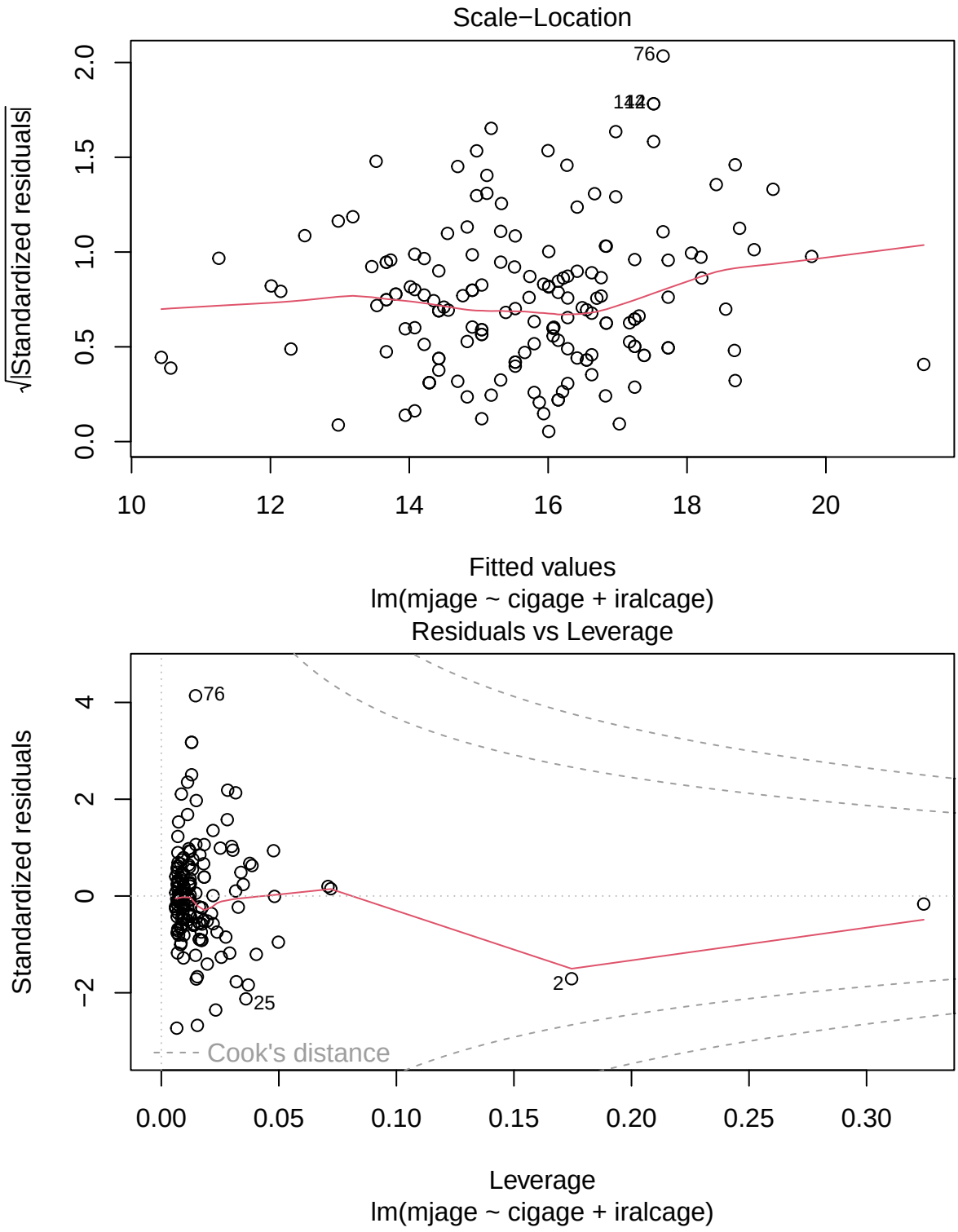
Answer:

7. How do the model diagnostics change? Did the model fit improve?

```
out.no.outliers <- lm(mjage ~ cigage + iralcage, dat.no.outliers)
plot(out.no.outliers)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mjage ~ cigage + iralcage)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mjage ~ cigage + iralcage)


Scale−Location
√|Standardized residuals|
76
142
0.0
0.5
1.0
1.5
2.0
10
12
14
16
18
20
Fitted values
lm(mjage ~ cigage + iralcage)
Residuals vs Leverage
Standardized residuals
76
25
2
Cook's distance
1
0.5
0.5
1
−2
0
2
4
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(mjage ~ cigage + iralcage)

Answer:

10. Compare the two model outputs. Do the coefficients change? What about R^2, the coefficient of determination?

```
summary(out.no.outliers)
```

```
##
## Call:
## lm(formula = mjage ~ cigage + iralcage, data = dat.no.outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1799 -1.6853 -0.2077  1.4058 12.3448
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.10734    1.37190   4.452 1.57e-05 ***
## cigage       0.13632    0.06432   2.119   0.0356 *
## iralcage     0.48251    0.08144   5.925 1.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.005 on 164 degrees of freedom
## Multiple R-squared:  0.2473, Adjusted R-squared:  0.2381
## F-statistic: 26.94 on 2 and 164 DF,  p-value: 7.671e-11
```

```
summary(out)
```

```
##
## Call:
## lm(formula = mjage ~ cigage + iralcage, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4195 -1.9189 -0.4195  1.1879 18.9370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.99942    1.63902   3.660 0.000337 ***
## cigage       0.14286    0.06451   2.215 0.028136 *
## iralcage     0.49940    0.09838   5.076 1.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.739 on 168 degrees of freedom
## Multiple R-squared:  0.1882, Adjusted R-squared:  0.1786
## F-statistic: 19.48 on 2 and 168 DF,  p-value: 2.465e-08
```

Answer: