HW#1

Tao He

1/25/2022

3.1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

sol.n: From Table 3.4, our three null hypotheses are that advertising budgets for "TV", "radio" or "newspaper" have no effect on sales, respectively. The mathematical expression is

$$\beta = 0$$

. The p-values corresponding to "TV" and "radio" are highly significant, and the p-value corresponding to "newspaper" is not significant; so we reject the first two original hypotheses and we do not reject the third one. Therefore, we can conclude that only the budget for newspaper advertising does not affect sales.

3.2. Carefully explain the differences between the KNN classifier and KNN regression methods.

sol.n: The KNN classifier uses the data and classifies new data points based on a similarity measure, while the classification is done by majority voting on its neighbors and then estimating the conditional probability. However, the KNN regression methods approximates the association between the independent variables and the continuous outcomes by averaging the observations in the same neighborhood in an intuitive way by alculating the average of the numerical targets in the K nearest neighbors

3.5. Consider the fitted values that result from performing linear regres- sion without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y} = x_i \hat{\beta}$$

where

$$\hat{\beta} = \frac{(\sum_{i=1}^{n} x_i y_i)}{\sum_{i'=1}^{n} x_{i'}^2}$$

Show that we can write

$$\hat{y_i} = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

What is

$$a_{i'}$$

?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

sol.n:

$$\hat{y_i} = x_i * \hat{\beta} = x_i * \frac{(\sum\limits_{i=1}^{n} x_i y_i)}{\sum\limits_{i'=1}^{n} x_{i'}^2} = \sum_{i=1}^{n} \frac{x_i y_i}{\sum\limits_{k=1}^{n} x_{k^2}} y_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

3.6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point

$$(\bar{x}, \bar{y})$$

.

sol.n: We have known

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

. The least squared line equation is

$$y = \hat{\beta}0 + \hat{\beta}1\bar{x}$$

. When

$$x = \bar{x}$$

,

$$y = \hat{\beta}0 + \hat{\beta}1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}1\bar{x} = \bar{y}$$

Therefore, a line will always pass through $(\bar{x}, \bar{y})$ when $x_i = \bar{x}$.

3.11. In this problem we will investigate the t-statistic for the null hypothesis

$$H_0 : \beta = 0$$

in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed(1)
x <- rnorm(100)
y <- 2 * x + rnorm(100)
```

a. Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate

$$\hat{\beta}$$

, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis

$$H_0 : \beta = 0$$

. Comment on these results. (You can perform regression without an intercept using the command lm(y x+0).)

```
fit1 <- lm(y ~ x + 0)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
```

```
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

We can get the estimate

$$\hat{\beta} = 1.9939$$

, with an estimate standard error is 0.1065. The t-statistic value is 18.73 and p-value is less than 2e-16. Since the small p-value, we can reject the null hypothesis

$$H_0 : \beta = 0$$

.

b. Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis

$$H_0 : \beta = 0$$

. Comment on these results.

```r
fit2 <- lm(x ~ y + 0)
summary(fit2)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

We can get the estimate

$$\hat{\beta} = 0.39111$$

, with an estimate standard error is 0.02089. The t-statistic value is 18.73 and p-value is less than 2e-16. Since the small p-value, we can reject the null hypothesis

$$H_0 : \beta = 0$$

.

c. What is the relationship between the results obtained in (a) and (b)?

sol.n: Both t-statistic value and p value in (a) and (b) are the same. And they actually create the same line. Therefore,

$$y = 2x + \epsilon$$

can be written as

$$x = 0.4(y - \epsilon)$$

.

d. For the regression of Y onto X without an intercept, the t-statistic for

$$H_0 : \beta = 0$$

takes the form

$$\hat{\beta}/SE(\hat{\beta})$$

, where

$$\hat{\beta}$$

is given by (3.38), and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2}}$$

. (These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{\sqrt{n-1}\sum_{i=1}^{n}((y_i - x_i\hat{\beta})^2}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i'=1}^{n} y_{i'}^2) - (\sum_{i'=1}^{n} x_{i'}y_{i'})^2}}$$

.

sol.n:

Show algebraically,

$$t = \frac{\frac{\sum_i x_i y_i}{\sum_j x_j^2}}{\sqrt{\frac{\sum_i (y_i - x_i\hat{\beta})^2}{(n-1)\sum_j x_j^2}}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_i (y_i - \frac{x_i \sum_j x_j y_j}{\sum_j x_j^2})^2}} = \frac{\sqrt{n-1}\sum_i x_i y_i}{\sqrt{(\sum_j x_j^2)(\sum_j y_j^2) - (\sum_j x_j y_j)}}$$

;

And confirm numerically in R,

```
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
as.numeric(t)
```

```
## [1] 18.72593
```

e. Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

sol.n: If we replace x to y, the result of equation will be the same.

f. In R, show that when regression is performed with an intercept, the t-statistic for

$$H_0 : \beta = 0$$

is the same for the regression of y onto x as it is for the regression of x onto y.

```
fit3 <- lm(y ~ x)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
fit4 <- lm(x ~y)
summary(fit4)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

We can obtain the t-statistic for

$$H_0 : \beta = 0$$

is the same for the regression of y onto x with 18.56 as it is for the regression of x onto y.

3.12. This problem involves simple linear regression without an intercept.

a. Recall that the coefficient estimate

$$\hat{\beta}$$

for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?

sol.n: The coefficient estimate for the regression of X onto Y is

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_j x_j^2}$$

;

The coefficient estimate for the regression of Y onto X is

$$\hat{\beta}' = \frac{\sum_i x_i y_i}{\sum_j y_j^2}$$

.

b. Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

```
set.seed(1)
x <- 1:100

y <- 2 * x + rnorm(100)

fit.y <- lm(y ~ x + 0)
fit.x <- lm(x ~ y + 0)

summary(fit.y)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23590 -0.62560  0.04426  0.58507  2.30926
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x 2.001514   0.001548    1293   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9005 on 99 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 1.672e+06 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
summary(fit.x)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15007 -0.29025 -0.01939  0.31486  1.11787
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## y 0.4995922  0.0003864    1293   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4499 on 99 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 1.672e+06 on 1 and 99 DF,  p-value: < 2.2e-16
```

c. Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.

```
set.seed(1)
x <- 1:100
y <- 100:1

fit.y <- lm(y ~ x + 0)
fit.x <- lm(x ~ y + 0)

summary(fit.y)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.75 -12.44  24.87  62.18  99.49
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   0.5075     0.0866    5.86 6.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.37 on 99 degrees of freedom
## Multiple R-squared:  0.2575, Adjusted R-squared:   0.25
## F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08
```

```
summary(fit.x)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -49.75 -12.44  24.87  62.18  99.49
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y   0.5075     0.0866    5.86 6.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.37 on 99 degrees of freedom
## Multiple R-squared:  0.2575, Adjusted R-squared:   0.25
## F-statistic: 34.34 on 1 and 99 DF,  p-value: 6.094e-08
```

3.13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

a. Using the rnorm() function, create a vector, x, containing 100 observations drawn from a

$$N(0, 1)$$

distribution. This represents a feature, X.

```
set.seed(1)
x <- rnorm(100)
```

b. Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a

$$N(0, 0.25)$$

distribution i.e. a normal distribution with mean zero and variance 0.25.

```
eps <- rnorm(100, sd = sqrt(0.25))
```

c. Using x and eps, generate a vector y according to the model

$$Y = -1 + 0.5X + \epsilon$$

What is the length of the vector y? What are the values of

$$\beta_0$$

and

$$\beta_1$$

in this linear model?

```
y <- -1 + 0.5 * x + eps
length(y)
```

```
## [1] 100
```

The length of the vector y is 100. The values of

$$\beta_0$$

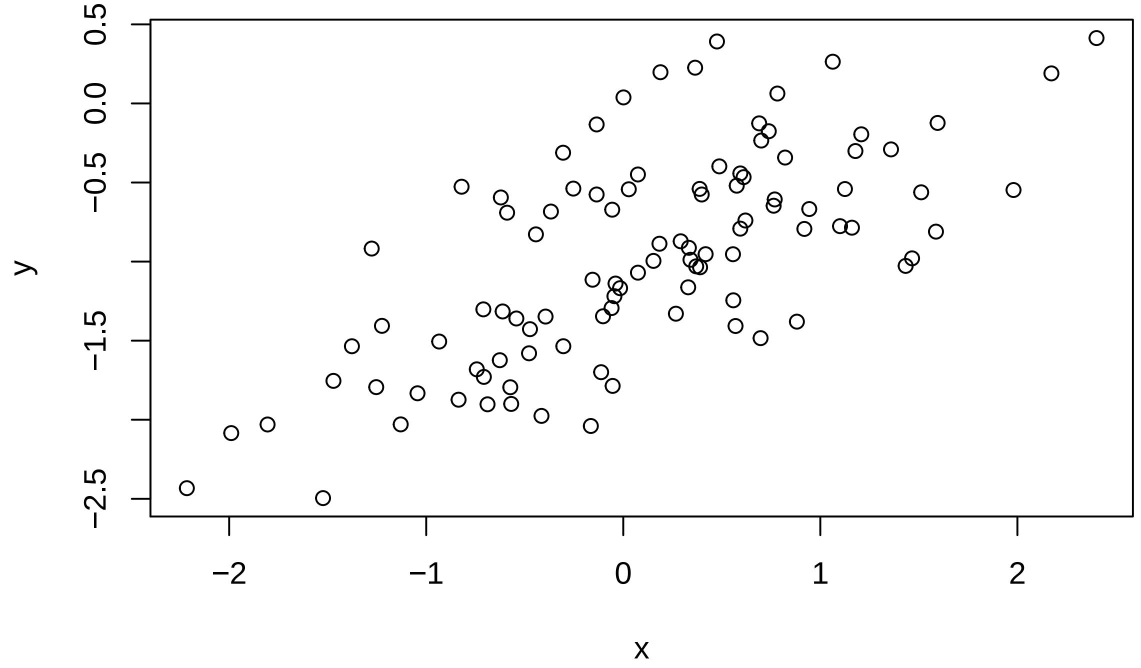
and

$$\beta_1$$

in this linear model are -1 and 0.5, respectively.

d. Create a scatter plot displaying the relationship between x and y. Comment on what you observe.

```
plot(x, y)
```



The relation between x and y looks like linear.

e. Fit a least squares linear model to predict y using x. Comment on the model obtained. How do

$$\hat{\beta}_0$$

and

$$\hat{\beta}_1$$

compare to

$$\beta_0$$

and

$$\beta_1$$

?

```
fit5 <- lm(y ~ x)
summary(fit5)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

The value of

$$\beta_0$$

and

$$\beta_1$$

compare to

$$\beta_0$$

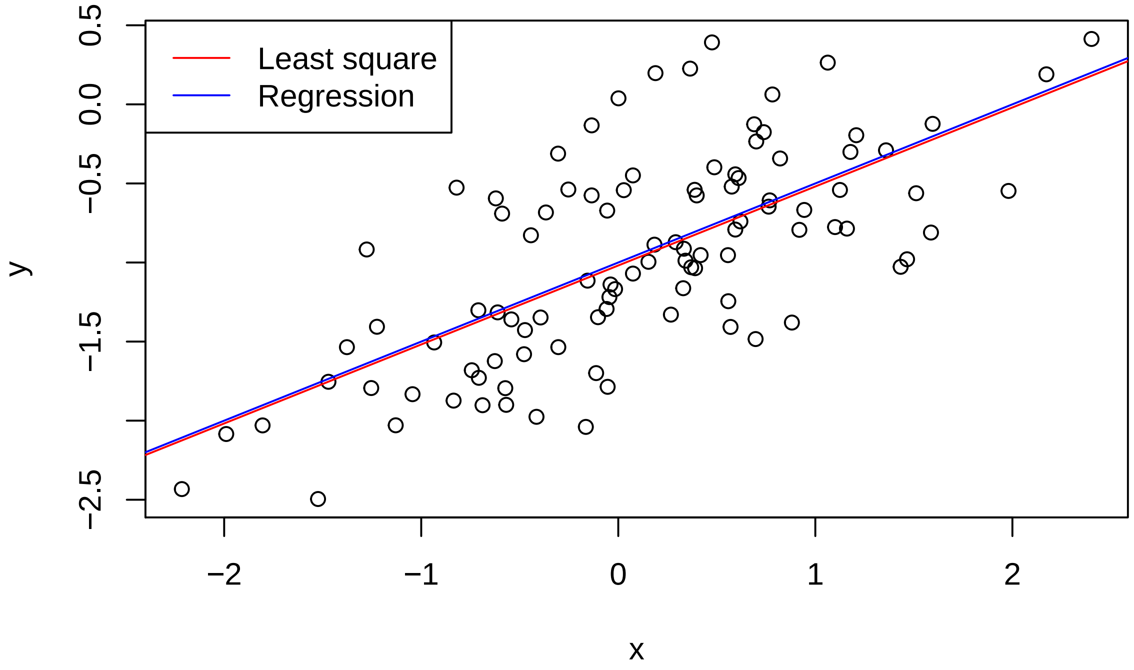
and

$$\beta_1$$

are very close.

f. Display the least squares line on the scatter plot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.

```
plot(x, y)
abline(fit5, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least square", "Regression"), col = c("red", "blue"), lty = c(1, 1))
```



g. Now fit a polynomial regression model that predicts y using x and x2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

```
fit6 <- lm(y ~ x + I(x^2))
summary(fit6)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98252 -0.31270 -0.06441  0.29014  1.13500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
## x            0.50858    0.05399   9.420  2.4e-15 ***
## I(x^2)      -0.05946    0.04238  -1.403    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.479 on 97 degrees of freedom
## Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
## F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

R^2 in linear regression model is 0.4619. R^2 in polynomial regression model is 0.4672. RSE in linear regression model is 0.4814. RSE in polynomial regression model is 0.479. Although the R^2 in polynomial regression model is slightly higher than it in linear regression model and RSE is slightly lower than it in linear regression model, the coefficient value of x^2 is not significant since the p-value is higher than 0.05. Therefore, we have not sufficient evidence to say the quadratic term improves the model fit.

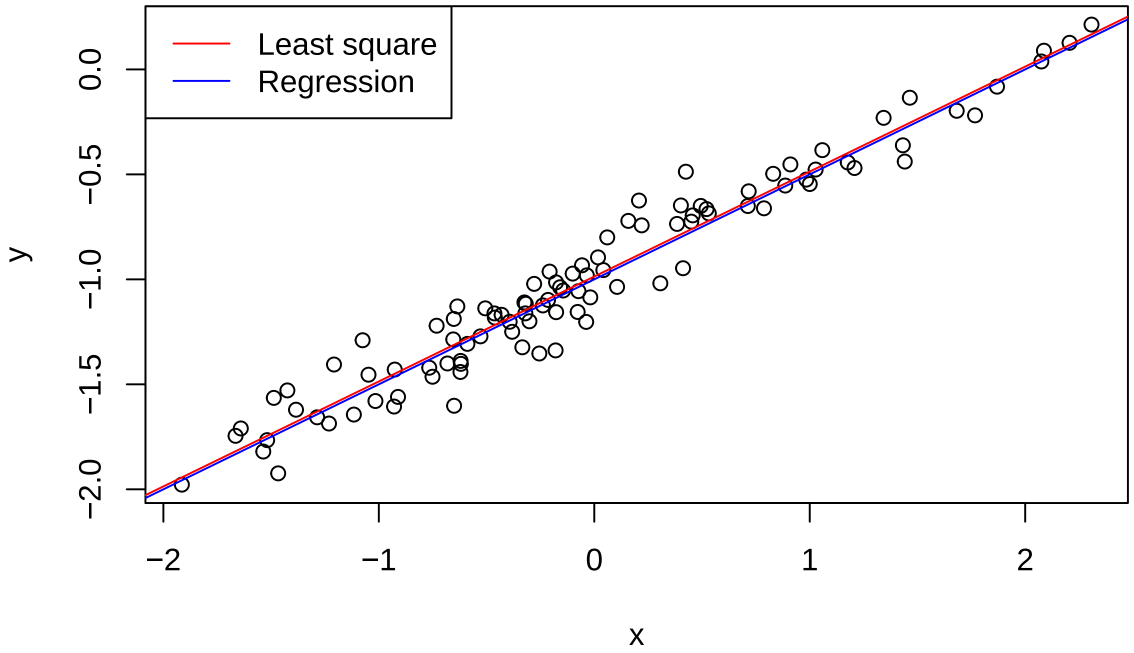
h. Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

```
set.seed(1)
eps <- rnorm(100, sd = 0.125)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps

fit7 <- lm(y ~ x)
summary(fit7)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29052 -0.07545  0.00067  0.07288  0.28664
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.98639    0.01129  -87.34   <2e-16 ***
## x            0.49988    0.01184   42.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1128 on 98 degrees of freedom
## Multiple R-squared:  0.9479, Adjusted R-squared:  0.9474
## F-statistic:  1782 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x, y)
abline(fit7, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least square", "Regression"), col = c("red", "blue"), lty = c(1, 1))
```



We try to decrease the variance of the normal distribution used to generate the error term in (b). Now, the relationship between x and y is nearly linear and we get a higher R^2 and lower RSE.

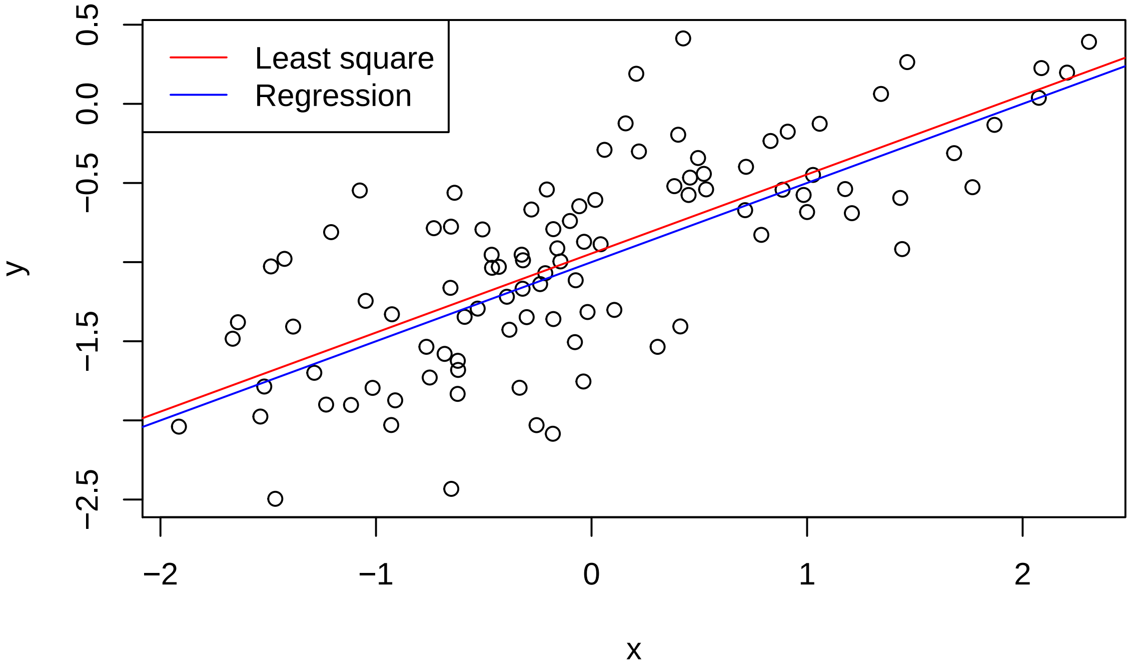
i. Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

```
set.seed(1)
eps <- rnorm(100, sd = 0.5)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps

fit8 <- lm(y ~ x)
summary(fit8)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16208 -0.30181  0.00268  0.29152  1.14658
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.94557    0.04517  -20.93   <2e-16 ***
## x            0.49953    0.04736   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4514 on 98 degrees of freedom
## Multiple R-squared:  0.5317, Adjusted R-squared:  0.5269
## F-statistic: 111.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(x, y)
abline(fit8, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least square", "Regression"), col = c("red", "blue"), lty = c(1, 1))
```



We try to increase the variance of the normal distribution used to generate the error term in (b). The relationship between x and y is not linear and we get a lower R^2 and higher RSE.

(j) What are the confidence intervals for

$$\beta_0$$

and

$$\beta_1$$

based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

```
confint(fit5)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```
confint(fit7)
```

```
##                 2.5 %     97.5 %
## (Intercept) -1.008805 -0.9639819
## x            0.476387  0.5233799
```

```
confint(fit8)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0352203 -0.8559276
## x            0.4055479  0.5935197
```

As the noise increases, the confidence intervals widen.

3.14. This problem focuses on the col linearity problem.

(a) Perform the following commands in R:

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100)/10
y <- 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

sol.n: The form of linear model is

$$y = 2 + 2x_1 + 0.3x2 + \epsilon$$

, with

$$\epsilon$$

~

$$N(0, 1)$$

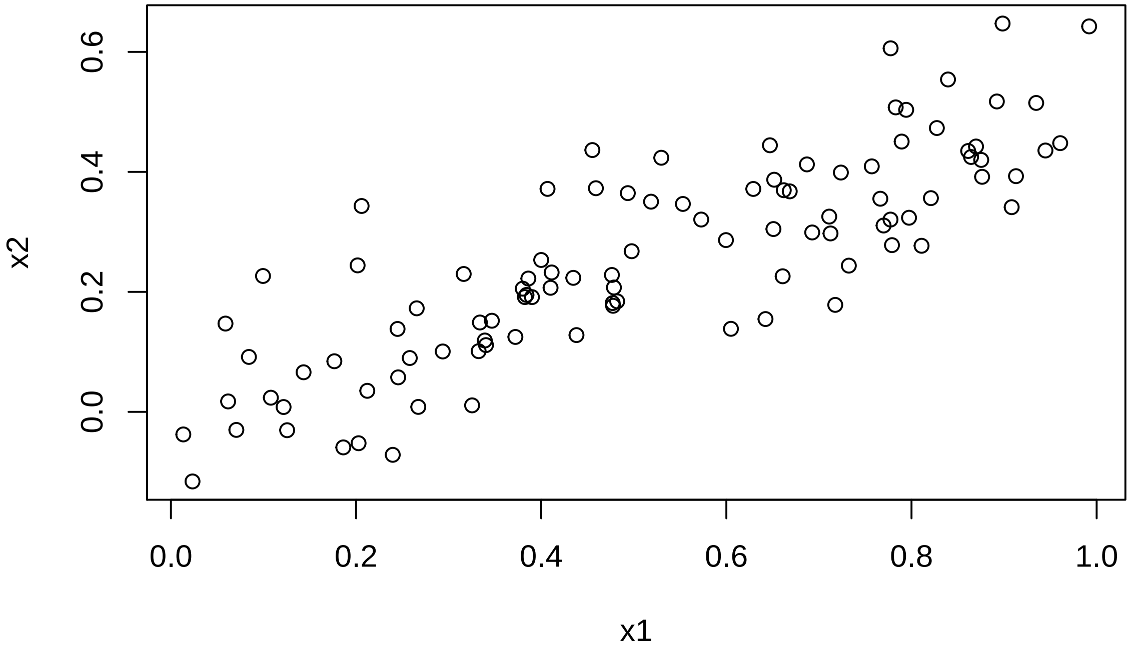
. The regression coefficients are 2, 2 and 0.3, respectively.

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
plot(x1, x2)
```



x1 and x2 are highly related.

(c) Using this data, fit a least squares regression to predict

$$y$$

using

$$x1$$

and

$$x2$$

. Describe the results obtained. What are

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

? How do these relate to the true

$$\beta_0, \beta_1, \beta_2$$

? Can you reject the null hypothesis

$$H_0 : \beta_1 = 0$$

? How about the null hypothesis

$$H_0 : \beta_2 = 0$$

?

```
fit9 <- lm(y ~ x1 + x2)
summary(fit9)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

are 2.1305, 1.4396 and 1.0097, respectively. Only

$$\hat{\beta}_0$$

is close to the true

$$\beta$$

. We can reject the null hypothesis

$$H_0 : \beta_1 = 0$$

at the 95% confidence level. However, we can not reject the null hypothesis

$$H_0 : \beta_2 = 0$$

at the 95% confidence level.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis

$$H_0 : \beta_1 = 0$$

?

```
fit10 <- lm(y ~ x1)
summary(fit10)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

The estimate coefficient of x1 in this model is different from one in (c). Now, we can reject the the null hypothesis

$$H_0 : \beta_1 = 0$$

since it has a small p-value.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis

$$H_0 : \beta_1 = 0$$

?

```
fit11 <- lm(y ~ x2)
summary(fit11)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

The estimate coefficient of x2 in this model is different from one in (c). Now, we can reject the the null hypothesis

$$H_0 : \beta_1 = 0$$

since it has a small p-value.

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

sol.n: No, it is reasonable. Since the predictor variables “x1” and “x2” are highly correlated, there exists covariance. Therefore, it is difficult to determine how each predictor variable is associated with the response separately. Since covariance reduces the accuracy of the regression coefficient estimates, it leads to an increase in the standard error of

$$\hat{\beta}_1$$

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

```
fit12 <- lm(y ~ x1 + x2)
fit13 <- lm(y ~ x1)
fit14 <- lm(y ~ x2)
summary(fit12)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
summary(fit13)
```

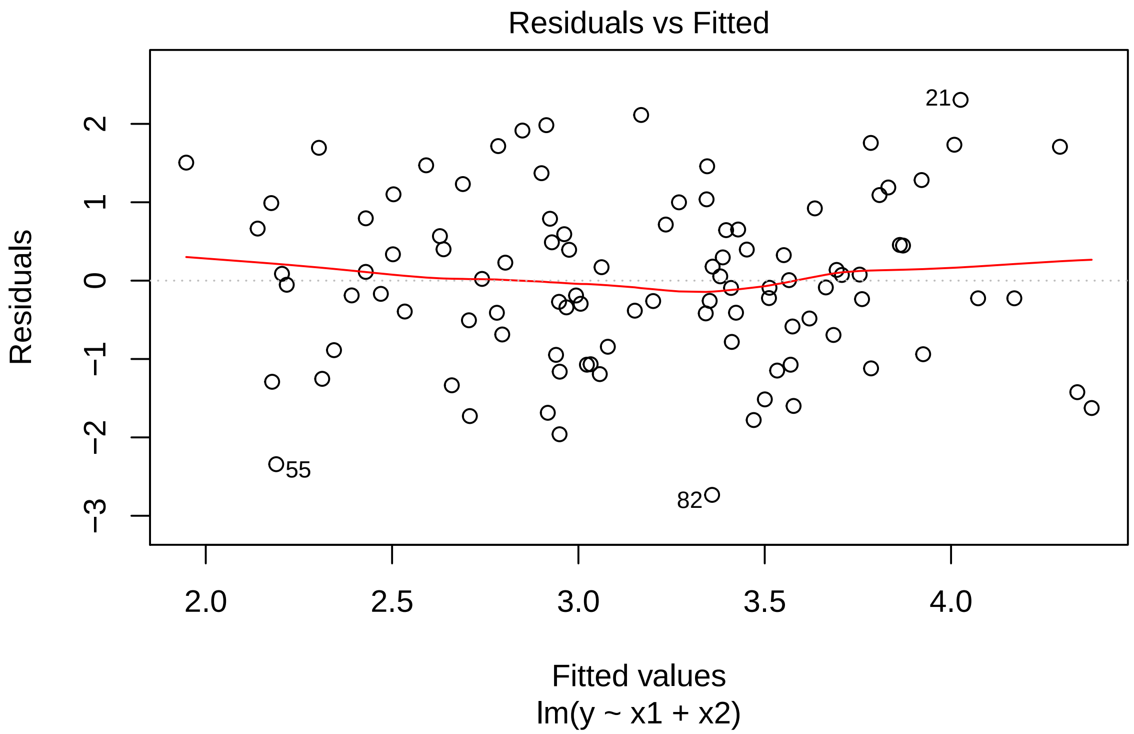
```
##
## Call:
## lm(formula = y ~ x1)
##
```

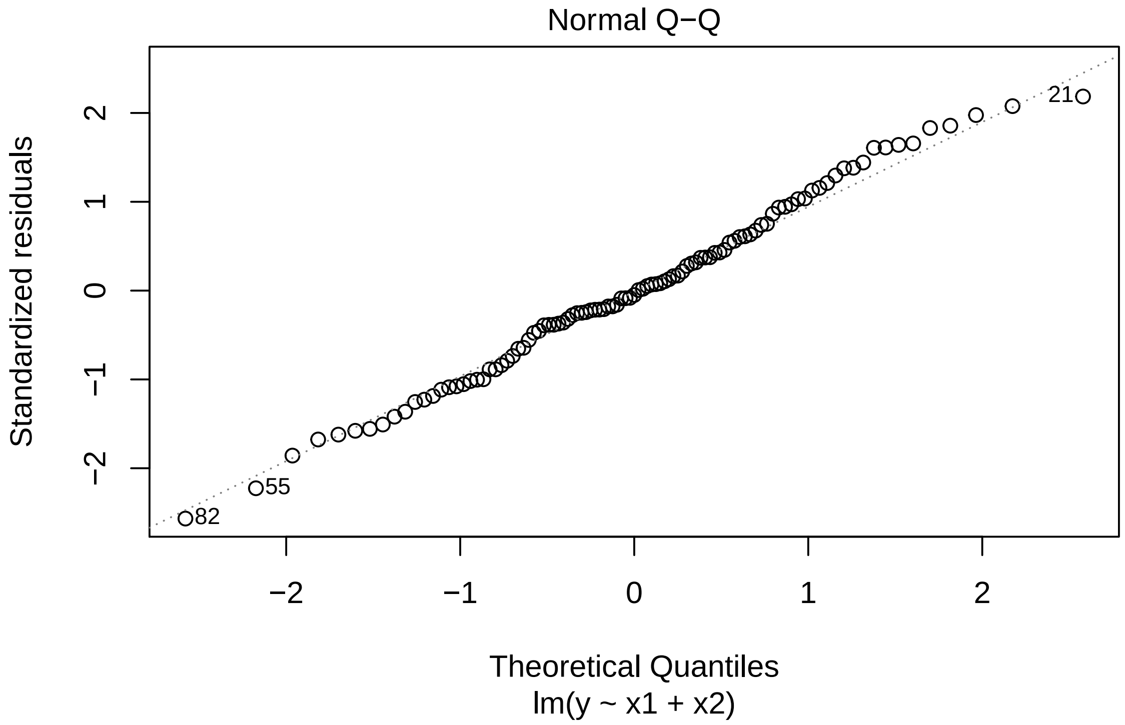
```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.8897 -0.6556 -0.0909  0.5682  3.5665 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477 
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

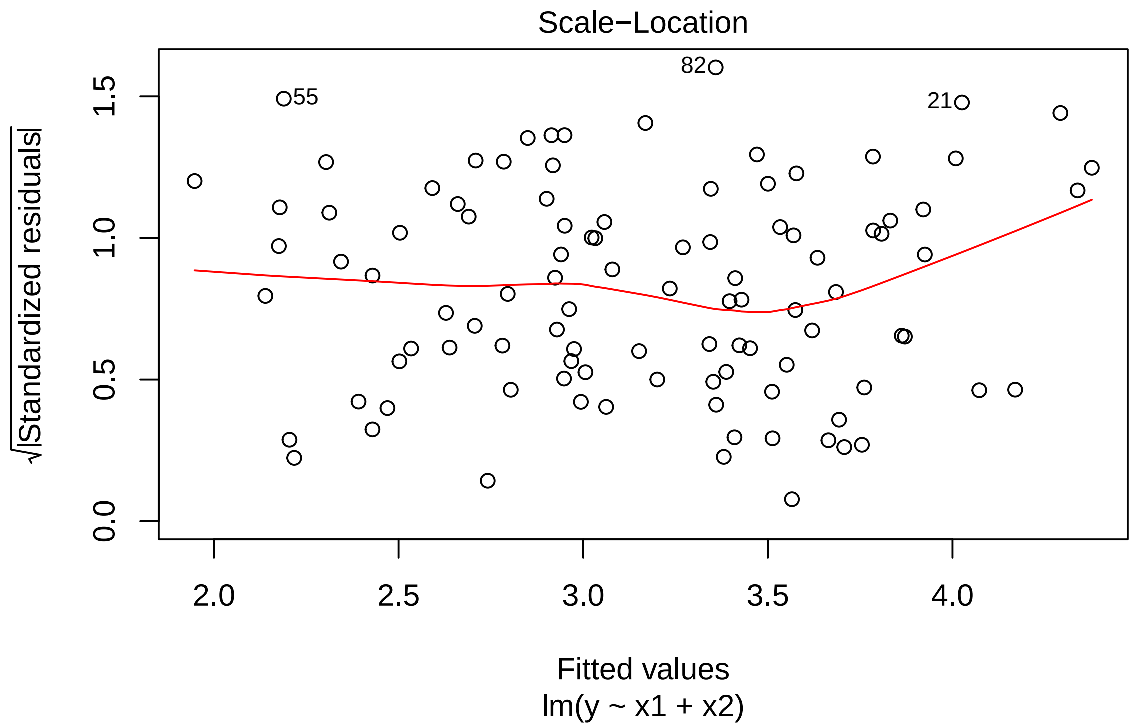
```
summary(fit14)
```

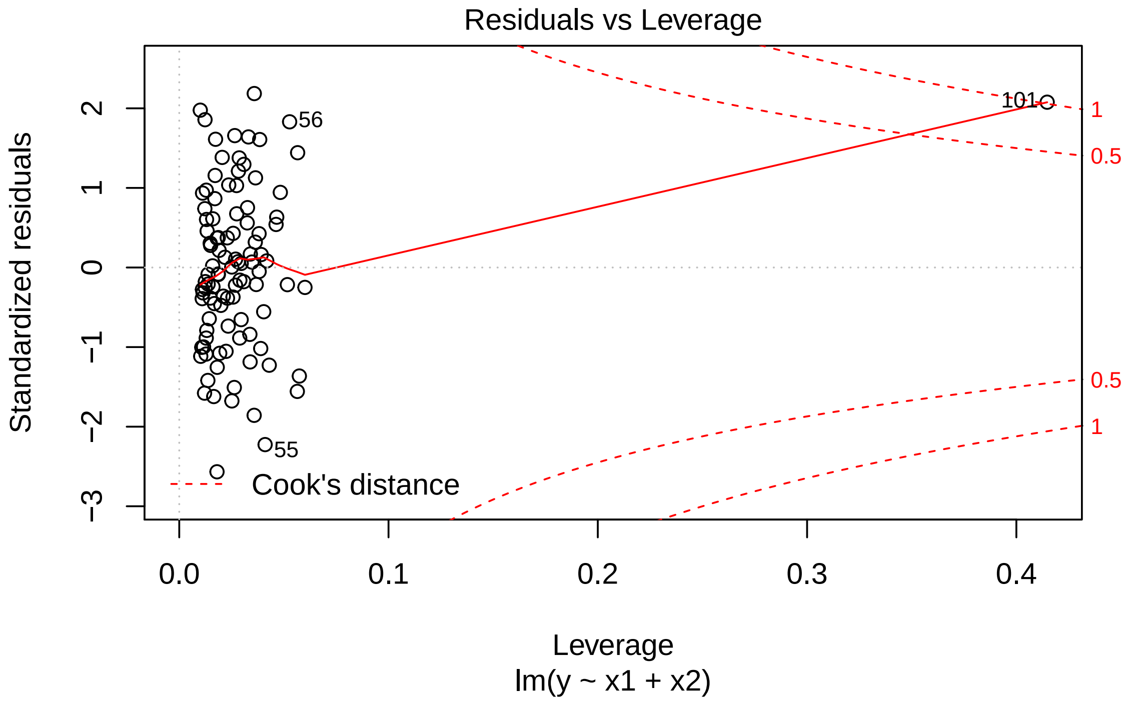
```
## 
## Call:
## lm(formula = y ~ x2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.64729 -0.71021 -0.06899  0.72699  2.38074 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042 
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```
plot(fit12)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
21
55
82
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1 + x2)

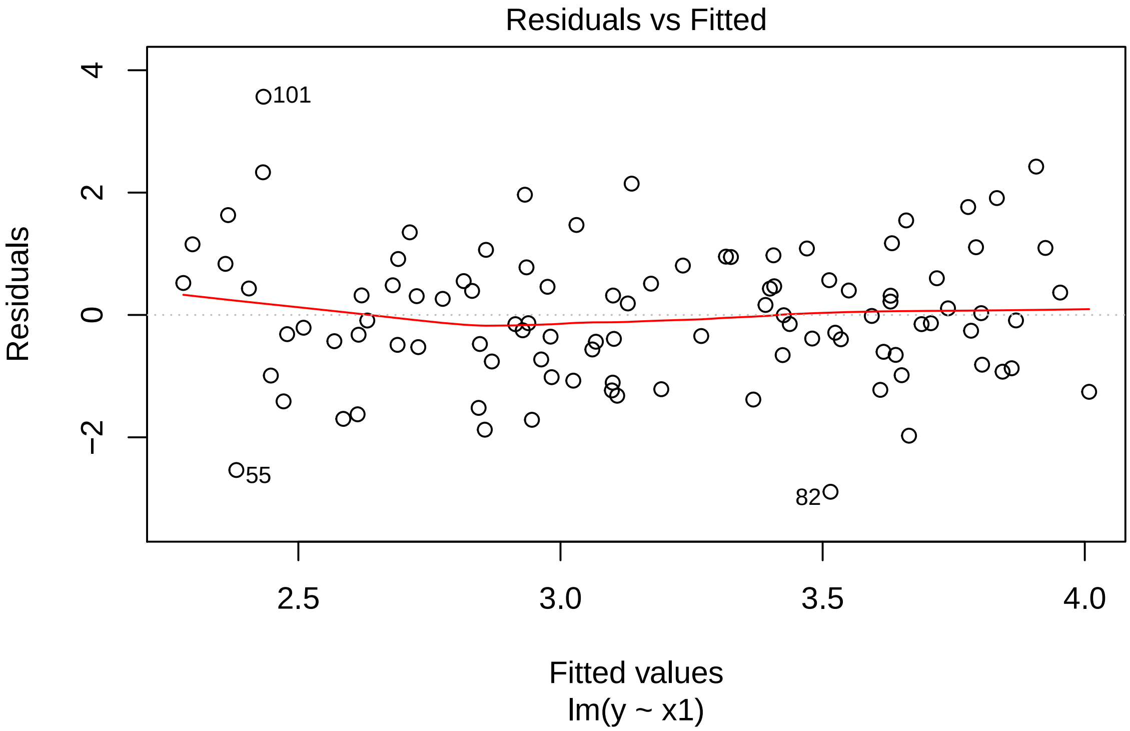

Normal Q−Q
Standardized residuals
2
1
0
−1
−2
21
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1 + x2)

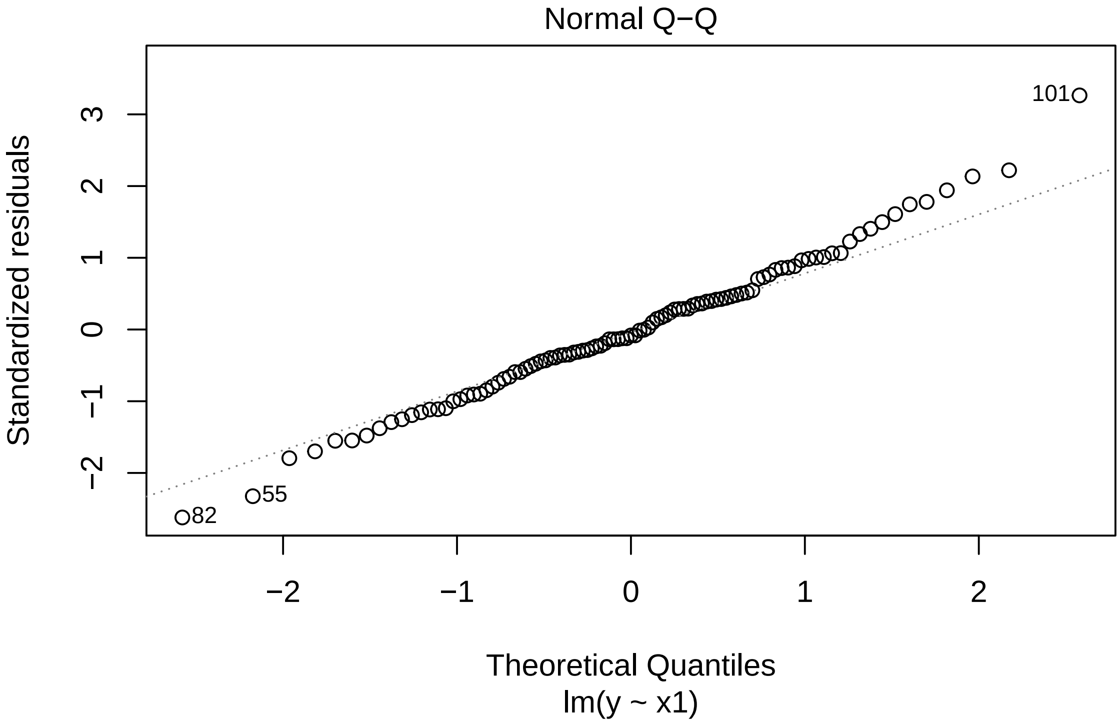

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
55
82
21
Fitted values
lm(y ~ x1 + x2)

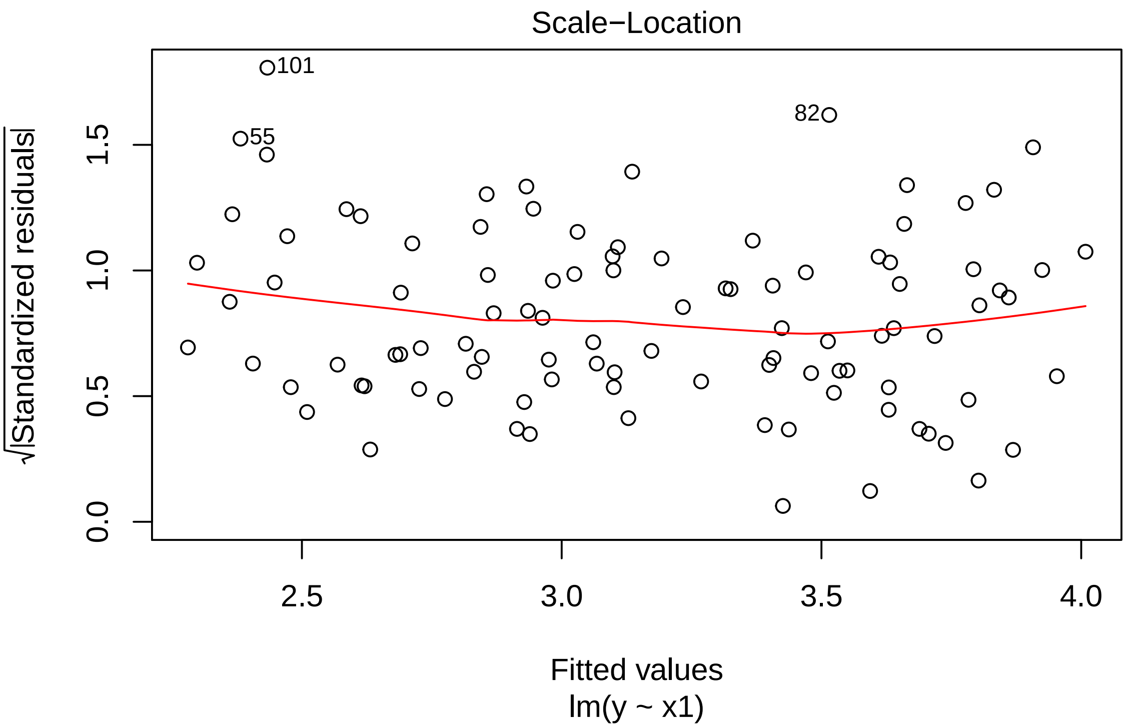

Residuals vs Leverage
Standardized residuals
−3
−2
−1
0
1
2
0.0
0.1
0.2
0.3
0.4
56
55
101
1
0.5
0.5
1
Cook's distance
Leverage
lm(y ~ x1 + x2)

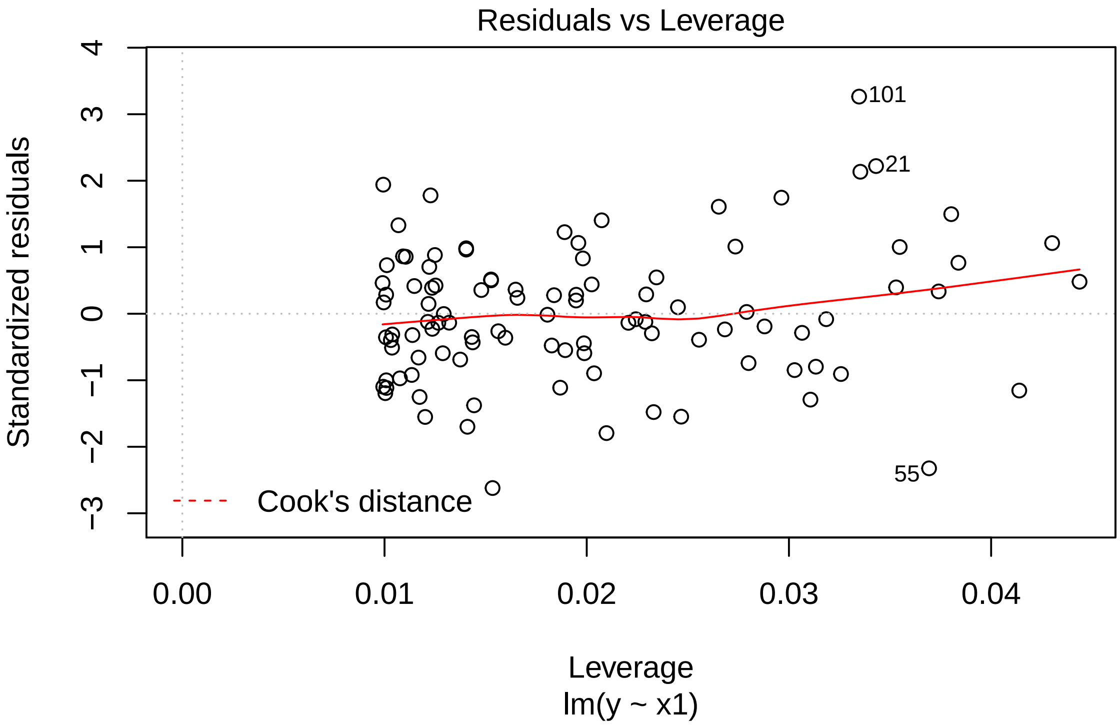
In the model with x1 and x2, the last point is the high leverage point.

```
plot(fit13)
```



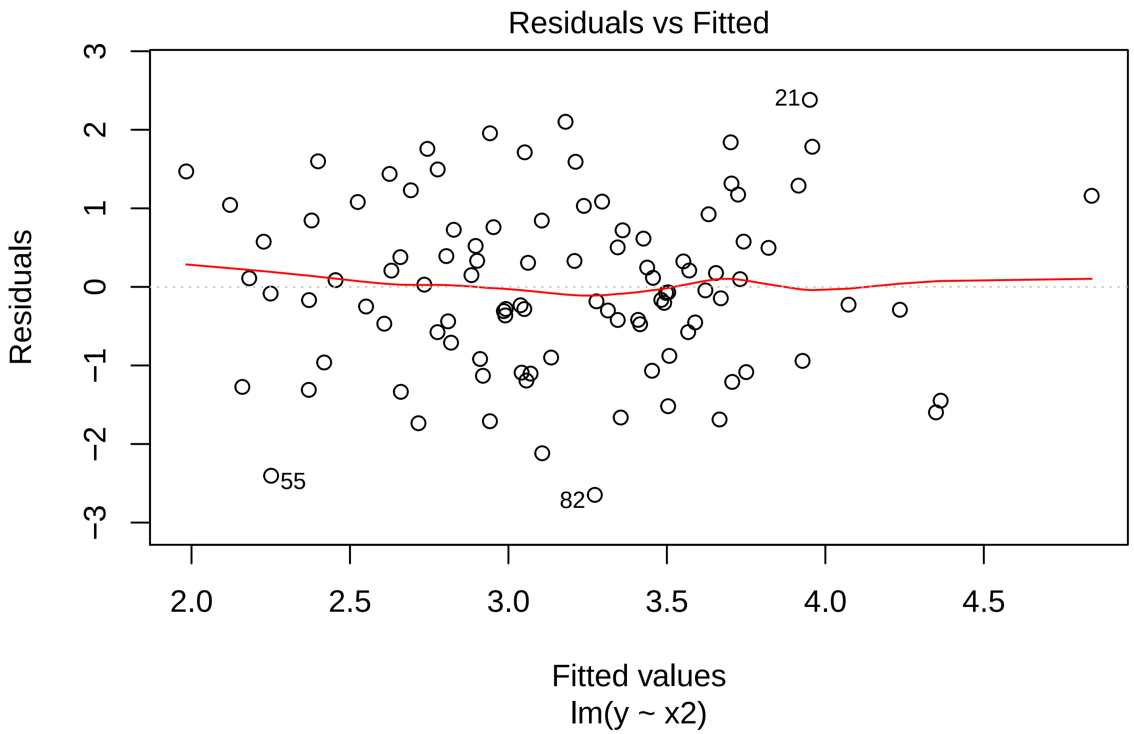


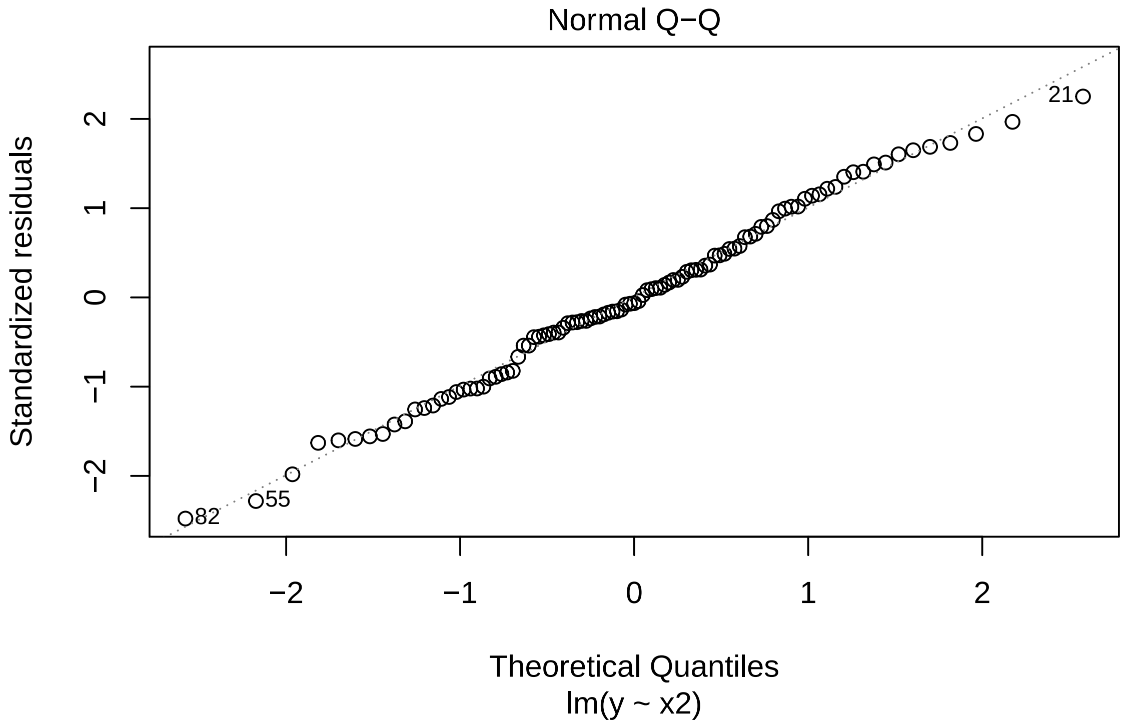


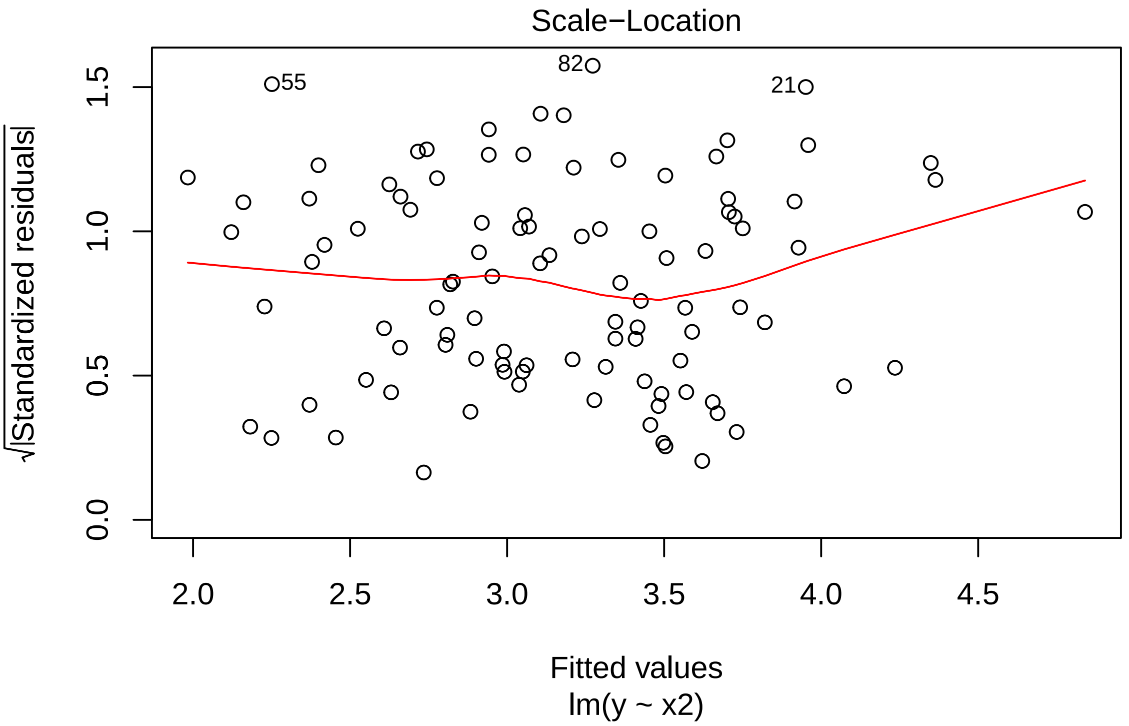


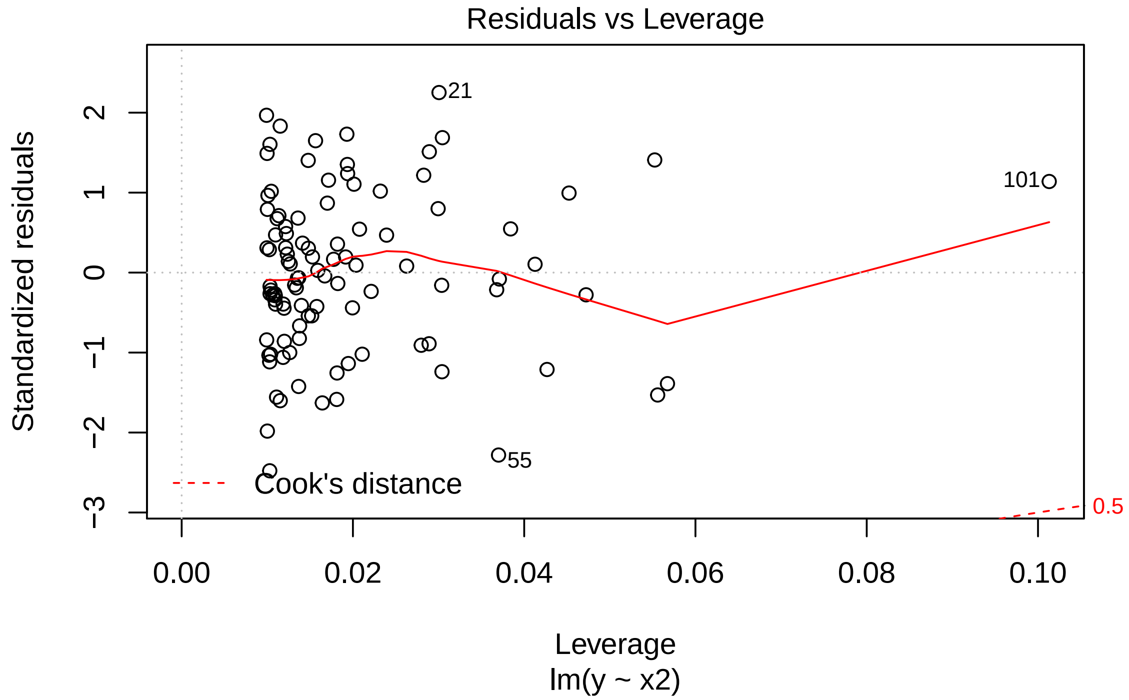
In the model with only x1, the last point is outlier.

```
plot(fit14)
```


Residuals vs Fitted
Residuals
3
2
1
0
−1
−2
−3
21
55
82
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
lm(y ~ x2)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
21
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x2)





In the model with only x2, the last point is the high leverage point.