MA679-Homework 2

Jinfei Xue

1/28/2019

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.

The null hypotheses associated with table 3.4 are that advertising budgets of TV, radio or newspaper do not have an effect on sales. That is to say, the null hypothesis is that any of the coefficients of TV, radio or newspaper is equal to 0.

From the table 3.4, we can see that the corresponding p-values are highly significant for TV and radio but not significant for newspaper. Therefore, we can conclude that the newspaper advertising budget is not associated with sales.

3.2

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

The KNN classifier is typically used to solve classification problems (those with a qualitative response) by identifying the neighborhood of x_0 and then estimating the conditional probability $P(Y = j|X = x_0)$ for class j as the fraction of points in the neighborhood whose response values equal j. The KNN regression method is used to solve regression problems (those with a quantitative response) by again identifying the neighborhood of x_0 and then estimating $f(x_0)$ as the average of all the training responses in the neighborhood.

3.5

$$\hat{y_i} = \frac{(x_1y_1 + x_2y_2 + ... + x_ny_n)x_i}{x_1^2 + x_2^2 + ... + x_n^2} = \frac{x_1x_i}{\sum_{j=1}^n x_j^2}y_1 + \frac{x_2x_i}{\sum_{j=1}^n x_j^2}y_2 + ... + \frac{x_nx_i}{\sum_{j=1}^n x_j^2}y_1$$

$$= \sum_{i'=1}^n \frac{x_{i'}x_i}{\sum_{j=1}^n x_j^2}y_{i'}$$

$$a_{i'} = \frac{x_{i'}x_i}{\sum_{j=1}^n x_j^2}$$

3.6

The least squares line equation is $y = \hat{\beta}_0 + \hat{\beta}_1x$, so if we substitute $\bar{x}$ for x we obtain $y = \hat{\beta}_0 + \hat{\beta}_1x = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} = \bar{y}$.

Therefore, we can conclude that the least square line passes through the point $(\bar{x}, \bar{y})$.

3.11

```
#generate a predictor x and a response y
set.seed (1)
x=rnorm (100)
y=2*x+rnorm (100)
```

(a) regression without intercept (y onto x)

```
r11_a=lm(y~x+0)
summary(r11_a)
```

```
##
## 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
```

From the summary above, we can see that the coefficient of x is 1.9939, which means if x increases by 1 unit, then the value of y will approximately increase by 1.9939 units.

the t-value for null hypothesis is 18.73 and the corresponding p-value is much smaller than 0.05. Therefore, we can reject the null hypothesis and indicate that x is associated with y.

R^2=0.7798 is large enough to show that a large proportion of the variability in the response has been explained by the regression.

(b) regression without intercept (x onto y)

```
r11_b <- lm(x ~ y + 0)
summary(r11_b)
```

```
##
## 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
```

From the summary above, we can see that the coefficient of y is 0.39111, which means if y increases by 1 unit, then the value of x will approximately increase by 0.39111 units.

the t-value for null hypothesis is 18.73 (which is equal to the value in (a)) and the corresponding p-value is much smaller than 0.05. Therefore, we can reject the null hypothesis and indicate that y is associated with x.

R^2=0.7798 (which is equal to the value in (a)) is large enough to show that a large proportion of the variability in the response has been explained by the regression.

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

We obtain the same value for the t statistics and consequently the same value for the corresponding p-value, and R-squared. Both results in (a) and (b) reflect the same line created in (a).

(d) Show the t-statistic formula algebraically and confirm it numerically in R;

Given $t = \frac{\beta}{SE(\beta)}$ and $SE\left(\beta\right) = \sqrt{\frac{\sum\left(y_i - \beta x_i\right)^2}{(n-1)\sum x_i^2}}$ and $\beta = \frac{\sum x_i y_i}{\sum x_i^2}$

Substituting β into t we get: $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2 - 2\sum y_i x_i \frac{\sum x_i y_i}{\sum x_i^2} + \sum\left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 x_i^2}}$

```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
```

We can see that the t value above is exactly same as the t-statistic given in the summary of "r11_a".

(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.

From the formula in (d), if we replace x_i with y_i and replace y_i with x_i for t-statistics, the result would be the same.

(f) In R, show that when regression is performed with an intercept, the t-statistic is the same for the regression of y onto x as it is for the regression of x onto y.

```r
# the regression of y onto x
r11_f1 <- lm(y ~ x)
summary(r11_f1)
```

```
##
## 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
```

```
# the regression of x onto y
r11_f2 <- lm(x ~ y)
summary(r11_f2)
```

```
##
## 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
```

From the summary results, we can see that the t-statistic is the same for the regression of y onto x as it is for the regression of x onto y, which is equal to 18.56.

3.12

This problem involves simple linear regression without an intercept.

(a) 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?

The coefficient estimate for the regression of Y onto X is

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

The coefficient estimate for the regression of X onto Y is

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

We can see from the above two formulas the two coefficients are the same if $\sum_j x_j^2 = \sum_j y_j^2$.

(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
sum(x^2)
```

```
## [1] 338350
```

```
y <- 2 * x + rnorm(100)
sum(y^2)
```

```
## [1] 1355530
```

We can see $\sum_j x_j^2$ is different from $\sum_j y_j^2$

```
r12_b1 <- lm(y ~ x + 0)
r12_b2 <- lm(x ~ y + 0)
summary(r12_b1)
```

```
##
## 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(r12_b2)
```

```
##
## 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
```

From summary results, we can see that the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.

(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.

```r
set.seed(1)
x <- 1:100
sum(x^2)
```

```
## [1] 338350
```

```r
y <- 100:1
sum(y^2)
```

```
## [1] 338350
```

We can see $\sum_j x_j^2$ *is the same as* $\sum_j y_j^2$

```r
r12_c1 <- lm(y ~ x + 0)
r12_c2 <- lm(x ~ y + 0)
summary(r12_c1)
```

```
## 
## 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
```

```r
summary(r12_c2)
```

```
## 
## 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
```

From summary results, we can see that the coefficient estimate for the regression of X *onto* Y *is the same as the coefficient estimate for the regression of* Y *onto* X*.*

3.13

(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.

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

(c)

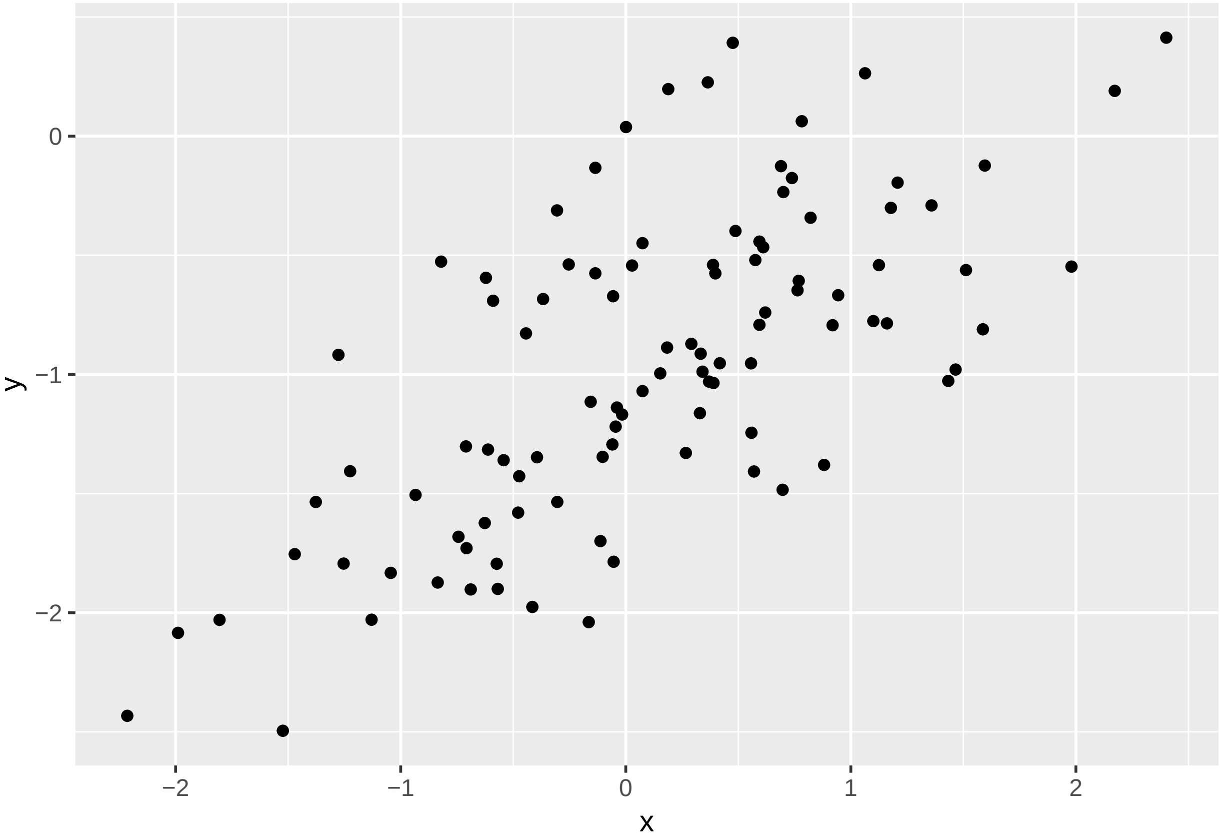
```
y <- -1 + 0.5 * x + eps
paste("the length of the vector y is", as.character(length(y)), sep = " ")
```

```
## [1] "the length of the vector y is 100"
```

In this linear model, $\beta_0 = -1$, $\beta_1 = 0.5$.

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

```
library(ggplot2)
ggplot(data = as.data.frame(y,x,eps))+
  geom_point(aes(x,y))
```



There exists a linear relationship between x and y with some noise introduced by the eps variable.

(e)

```
r13_e <- lm(y ~ x)
summary(r13_e)
```

```
##
## 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
```

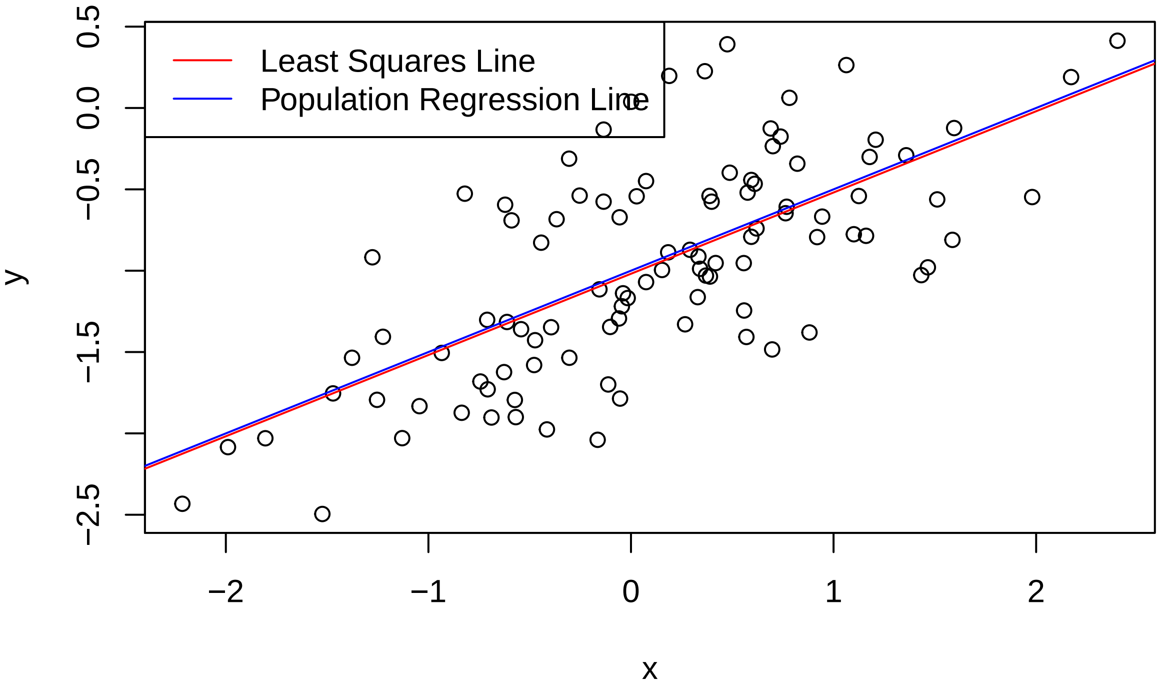
The p-values for t-statistics are pretty small, which indicates the coefficients in the linear model are significant.

$\hat{\beta}_0 = -1.01885, \hat{\beta}_1 = 0.49947$

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are pretty close to the values of β_0 and β_1.

(f)

```
plot(x, y)
abline(r13_e, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least Squares Line", "Population Regression Line"), col = c("red", "blue"), lty =
```



(g) Now fit a polynomial regression model

```
r13_g <- lm(y ~ poly(x,2,raw = TRUE))
# r13_g <- lm(y ~ x + I(x^2))
summary(r13_g)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2, raw = TRUE))
##
## 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 ***
## poly(x, 2, raw = TRUE)1  0.50858    0.05399   9.420  2.4e-15 ***
## poly(x, 2, raw = TRUE)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
```

The coefficient of x^2 is not significant, so there is not sufficient evidence to show the quadratic term can improve the model even if the R^2 is little higher.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The initial model should remain the same. Describe your results.

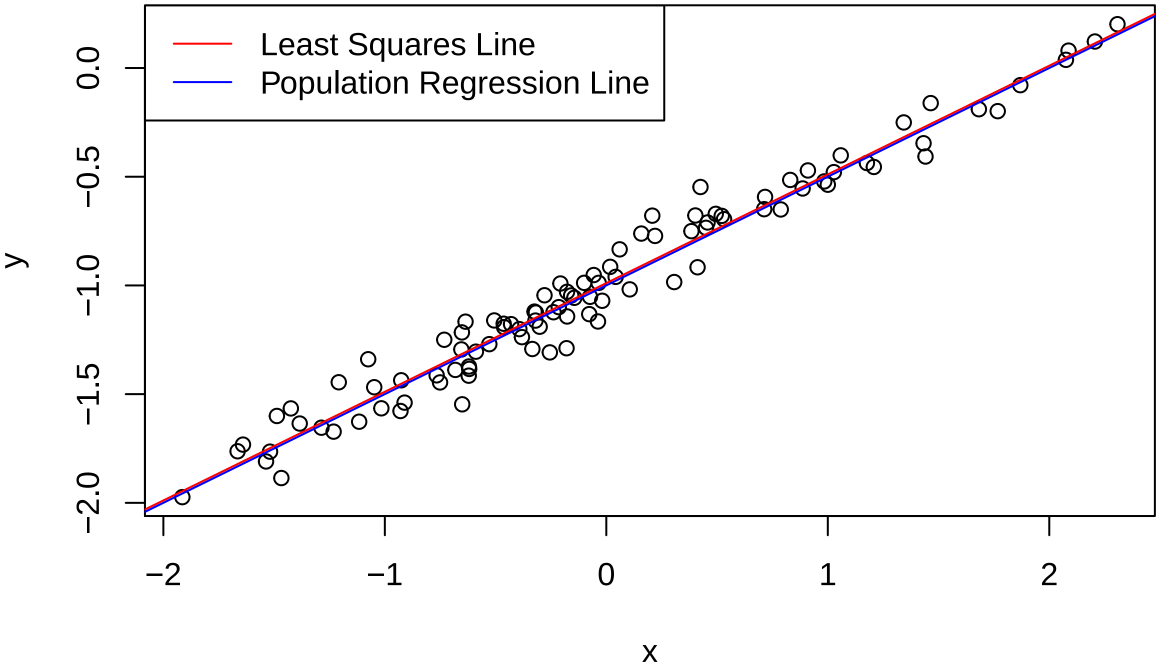
```
# generate variables
set.seed(1)
eps <- rnorm(100, sd = 0.1)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps

# regression
r13_h <- lm(y ~ x)
summary(r13_h)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.232416 -0.060361  0.000536  0.058305  0.229316
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.989115   0.009035 -109.48   <2e-16 ***
## x            0.499907   0.009472   52.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.09028 on 98 degrees of freedom
## Multiple R-squared:  0.966,  Adjusted R-squared:  0.9657
## F-statistic:  2785 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# plot the relationship
plot(x, y)
abline(r13_h, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least Squares Line", "Population Regression Line"), col = c("red", "blue"), lty =
```



We reduced the noise by decreasing the variance of the normal distribution used to generate the error term ϵ. We may see that the coefficients are very close to those in population regression. Besides, R^2 is much higher and RSE is much lower than the previous one. Moreover, the two lines are very close to each other.

(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is more noise in the data.

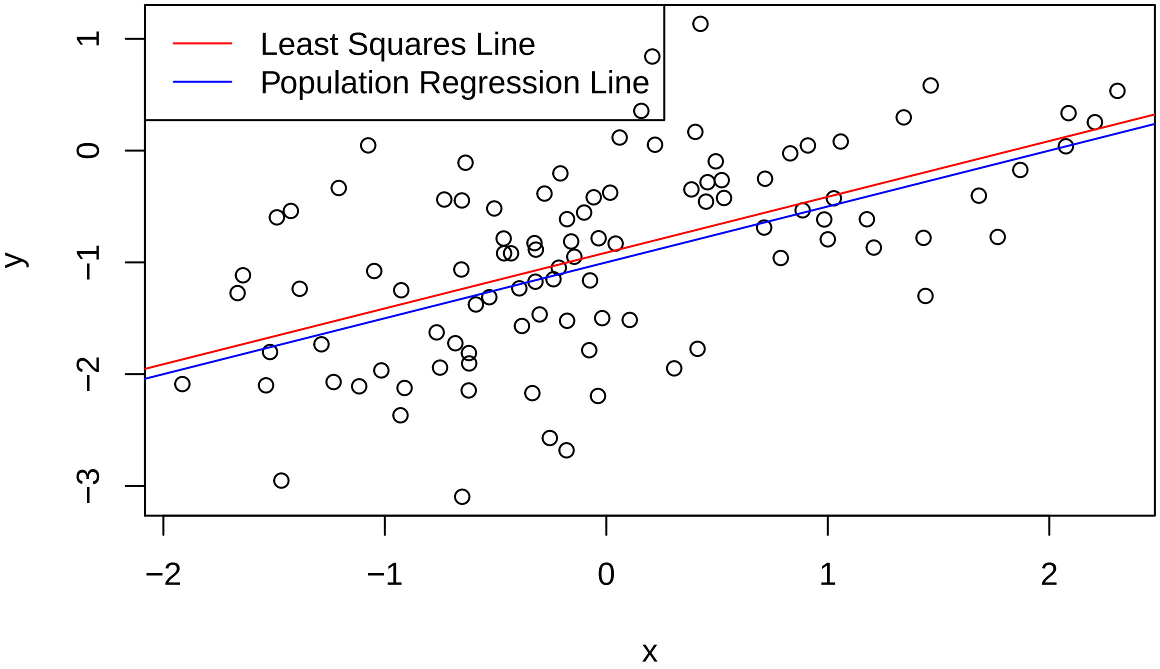
```
# generate variables
set.seed(1)
eps <- rnorm(100, sd = 0.8)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps

# regression
r13_i <- lm(y ~ x)
summary(r13_i)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85933 -0.48289  0.00429  0.46644  1.83453
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91292    0.07228 -12.631  < 2e-16 ***
## x            0.49925    0.07578   6.588 2.23e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7222 on 98 degrees of freedom
## Multiple R-squared:  0.307,  Adjusted R-squared:  0.2999
## F-statistic: 43.41 on 1 and 98 DF,  p-value: 2.235e-09
```

```r
# plot the relationship
plot(x, y)
abline(r13_i, col = "red")
abline(-1, 0.5, col = "blue")
legend("topleft", c("Least Squares Line", "Population Regression Line"), col = c("red", "blue"), lty =
```



We increased the noise by increasing the variance of the normal distribution used to generate the error term ϵ. We can see that the coefficients are close to those in population regression. However, R^2 is much lower and RSE is much higher than the previous one. Moreover, the two lines are wider apart but are still close to each other.

(j) confidence intervals

```r
# the original data set
confint(r13_e)
```

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

```r
# the less noisy data set
confint(r13_h)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0070441 -0.9711855
## x             0.4811096  0.5187039
```

```
# the noisier data set
confint(r13_i)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0563524 -0.7694842
## x            0.3488766  0.6496316
```

Larger the deviation of noise is, wider the confidence intervals are. With less noise, there is more predictability in the data set.

3.14

(a) 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?

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

The form of the linear model is:

$$Y = 2 + 2X_1 + 0.3X_2 + \epsilon$$

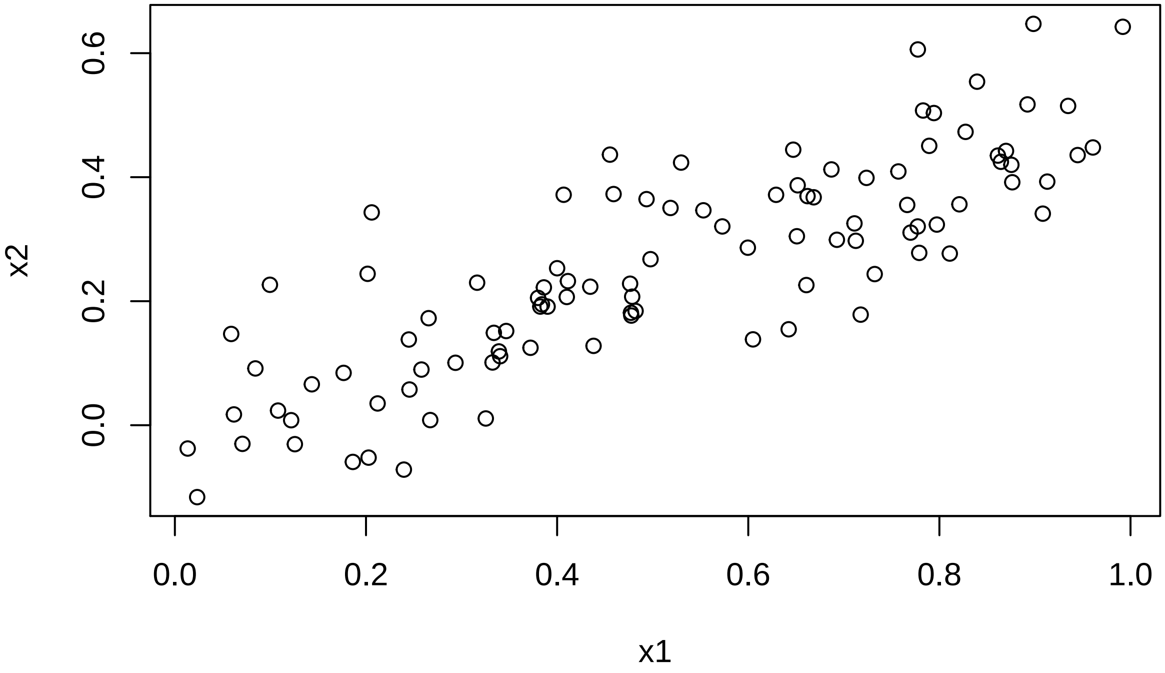
$\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$.

(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)
```



From both correlation value and scatterplot, we can see x1 and x2 is highly correlated with each other.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained.

```r
r14_c <- lm(y ~ x1 + x2)
summary(r14_c)
```

```
##
## 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
```

$\hat{\beta}_0 = 2.1305$, $\hat{\beta}_1 = 1.4396$, $\hat{\beta}_2 = 1.0097$. *Only the value of $\hat{\beta}_0$ is very close to that of β_0. Because the p-value for β_1 is smaller than 0.05, we can reject the null hypothesis. However, the p-value for β_2 is bigger than 0.05, so we cannot reject H_0.*

(d) Now fit a least squares regression to predict y using only x1. Comment on your results.

```r
r14_d <- lm(y ~ x1)
summary(r14_d)
```

```
##
## 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 coefficient for x1 in this model is very different from that in (c). In this case because the p-value for x1 is very low, we can reject H_0, which indicates that x1 is highly significant.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results.

```
r14_e <- lm(y ~ x2)
summary(r14_e)
```

```
##
## 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 coefficient for x2 in this model is very different from that in (c). In this case because the p-value for x2 is very low, we can reject H_0, which indicates that x2 is highly significant.

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

No, the results do not contradict each other. As the predictors x1 and x2 are highly correlated, there exists collinearity between x1 and x2. In this case it can be difficult to determine how each predictor separately is associated with the response. Since collinearity reduces the accuracy of the coefficients estimation. Consequently, we may fail to reject H_0 in the presence of collinearity. The importance of the x2 variable has been masked due to the presence of collinearity.

(g)

```
x1=c(x1 , 0.1)
x2=c(x2 , 0.8)
y=c(y,6)
```

```
r14_c2 <- lm(y ~ x1 + x2)
r14_d2 <- lm(y ~ x1)
r14_e2 <- lm(y ~ x2)
summary(r14_c2)
```

```
##
## 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
```

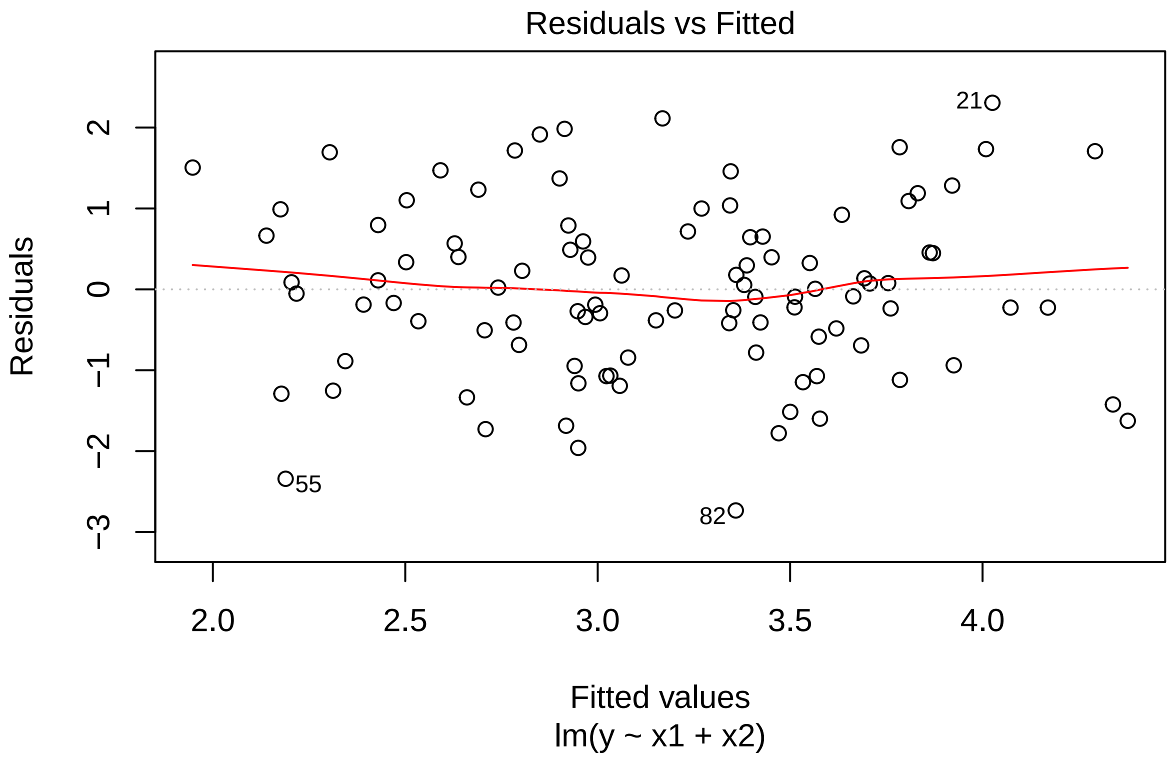
```r
summary(r14_d2)
```

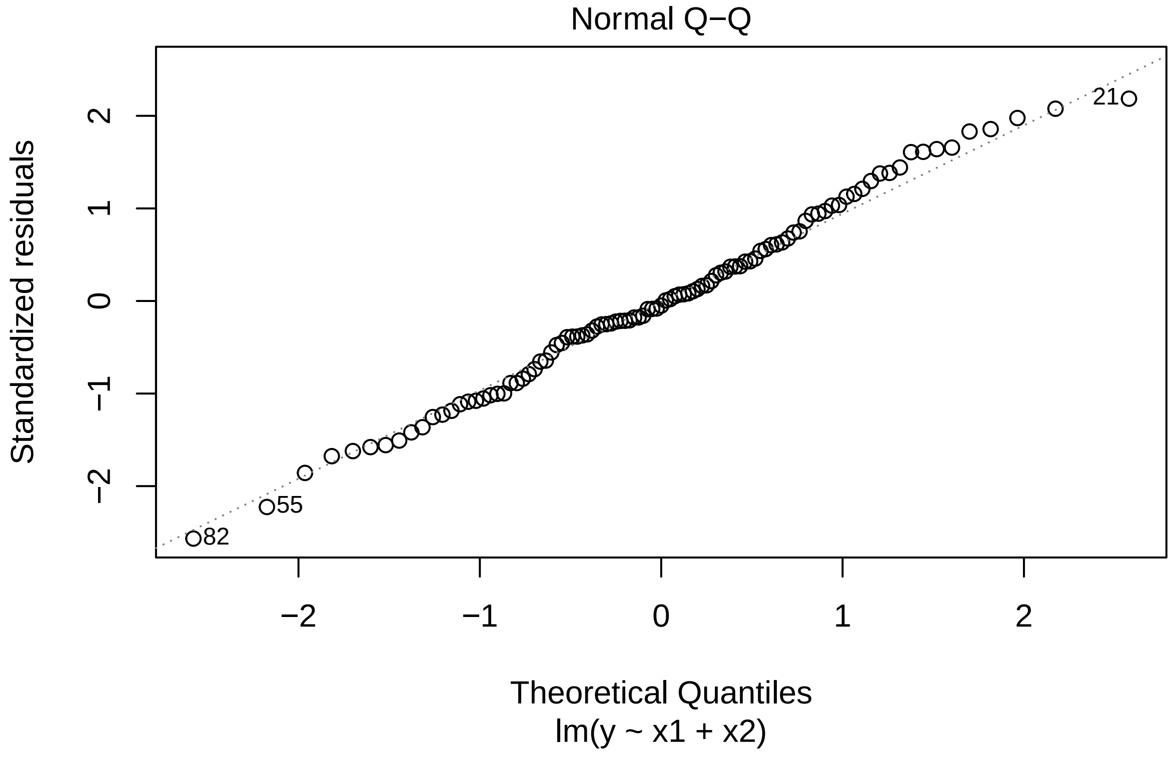
```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   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
```

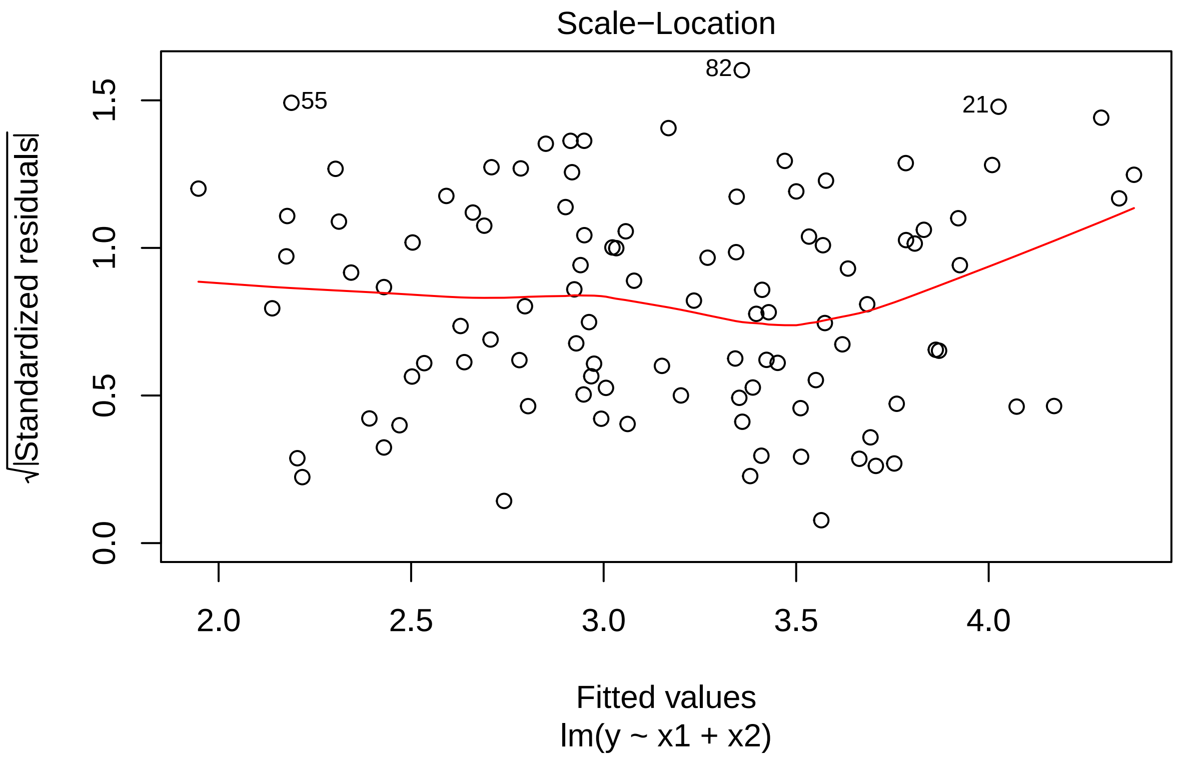
```r
summary(r14_e2)
```

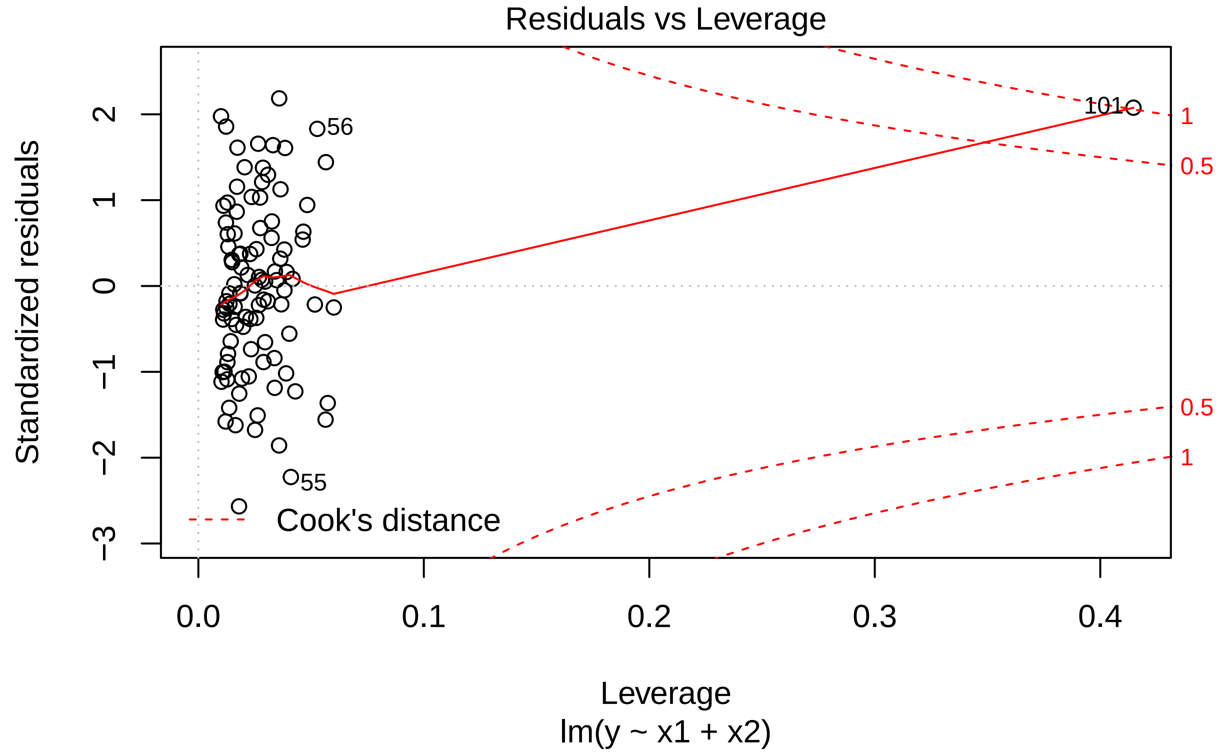
```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     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
```

```r
plot(r14_c2)
```

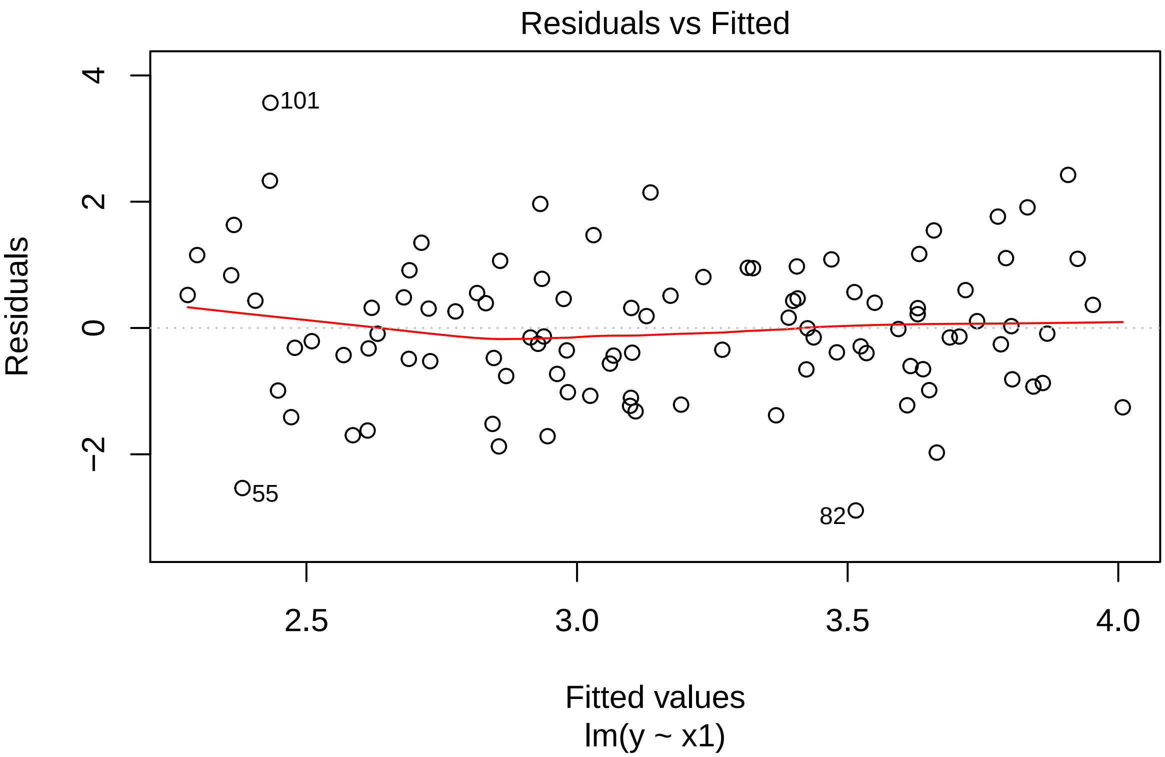

Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
21
55
82
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1 + x2)

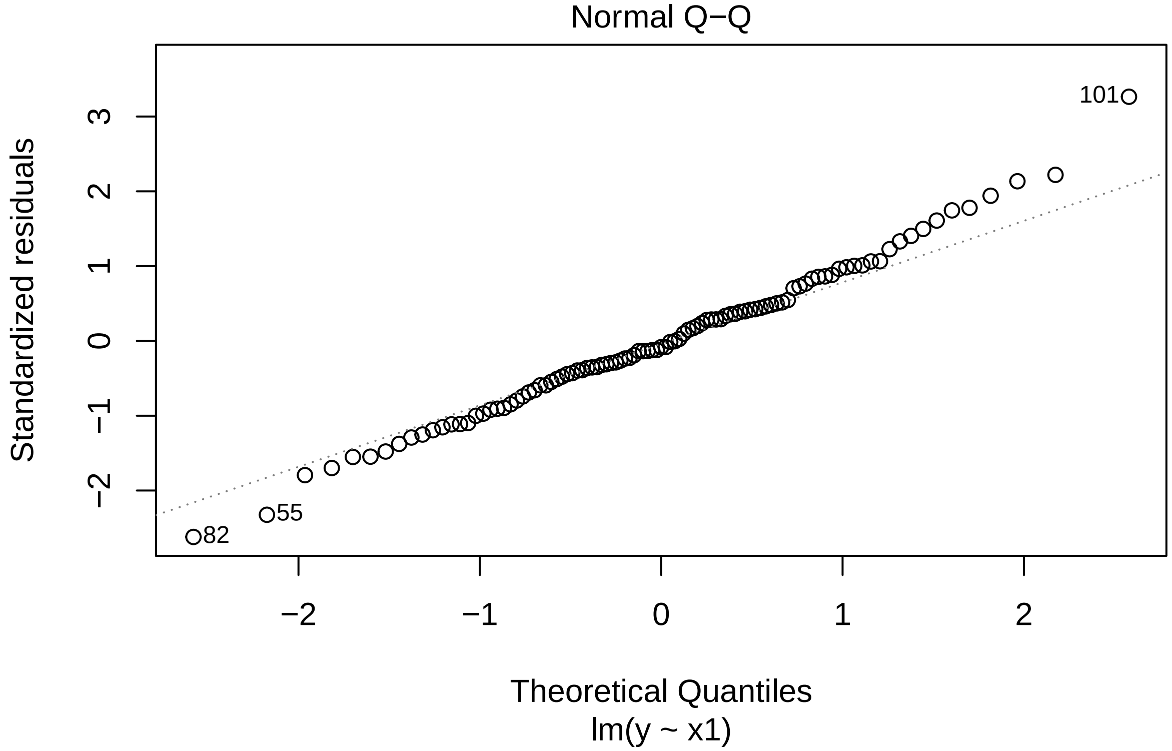

Normal Q−Q
Standardized residuals
2
1
0
−1
−2
21
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1 + x2)

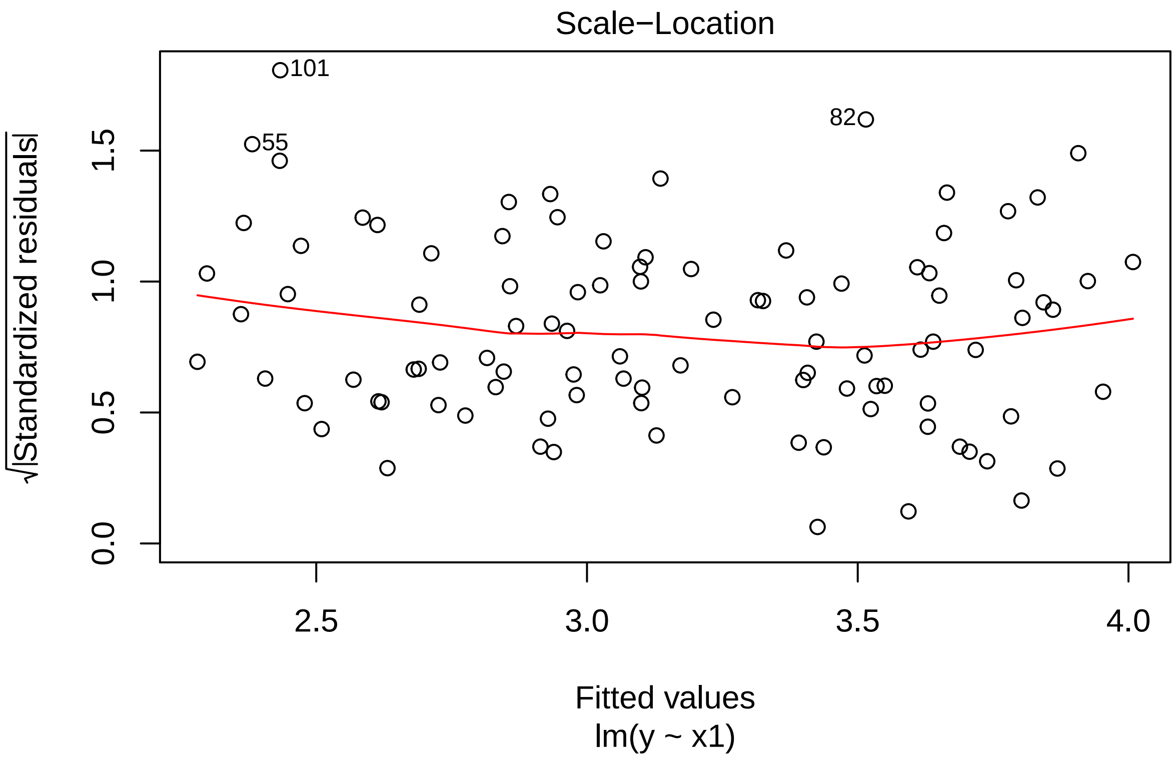

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
55
82
21
2.0
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1 + x2)

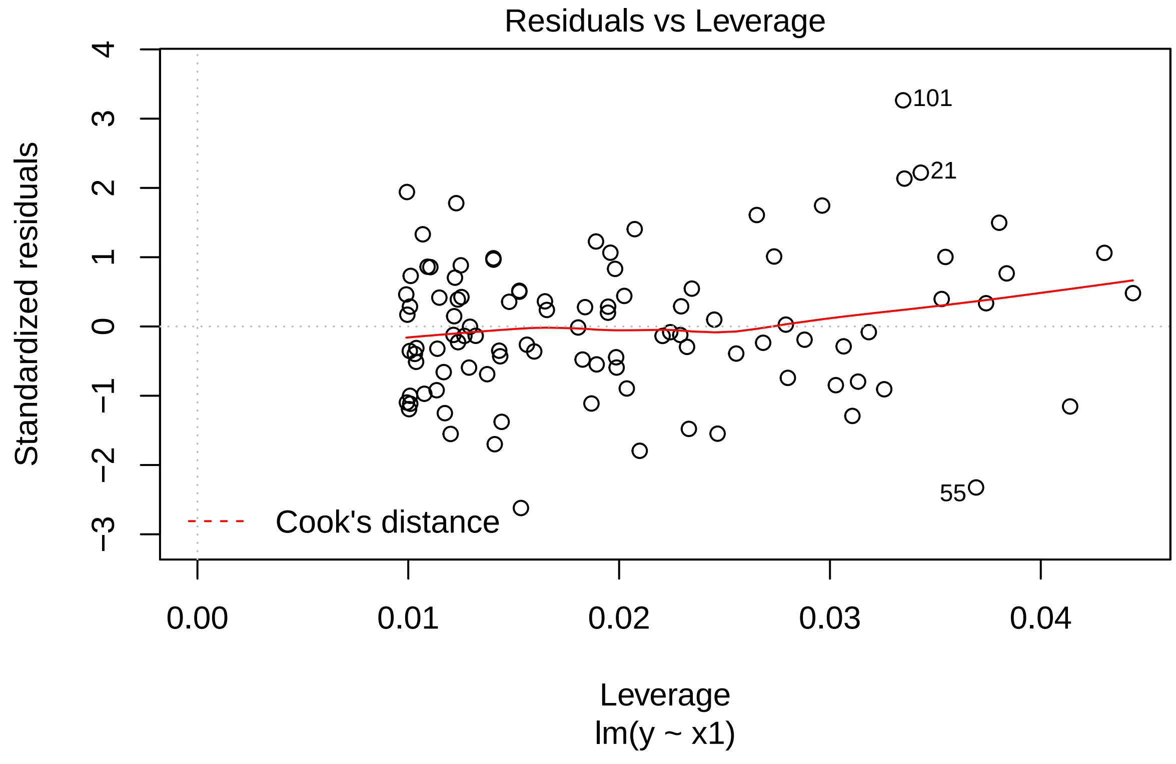

Residuals vs Leverage
Standardized residuals
−3
−2
−1
0
1
2
56
55
101
1
0.5
0.5
1
Cook's distance
0.0
0.1
0.2
0.3
0.4
Leverage
lm(y ~ x1 + x2)

```
plot(r14_d2)
```

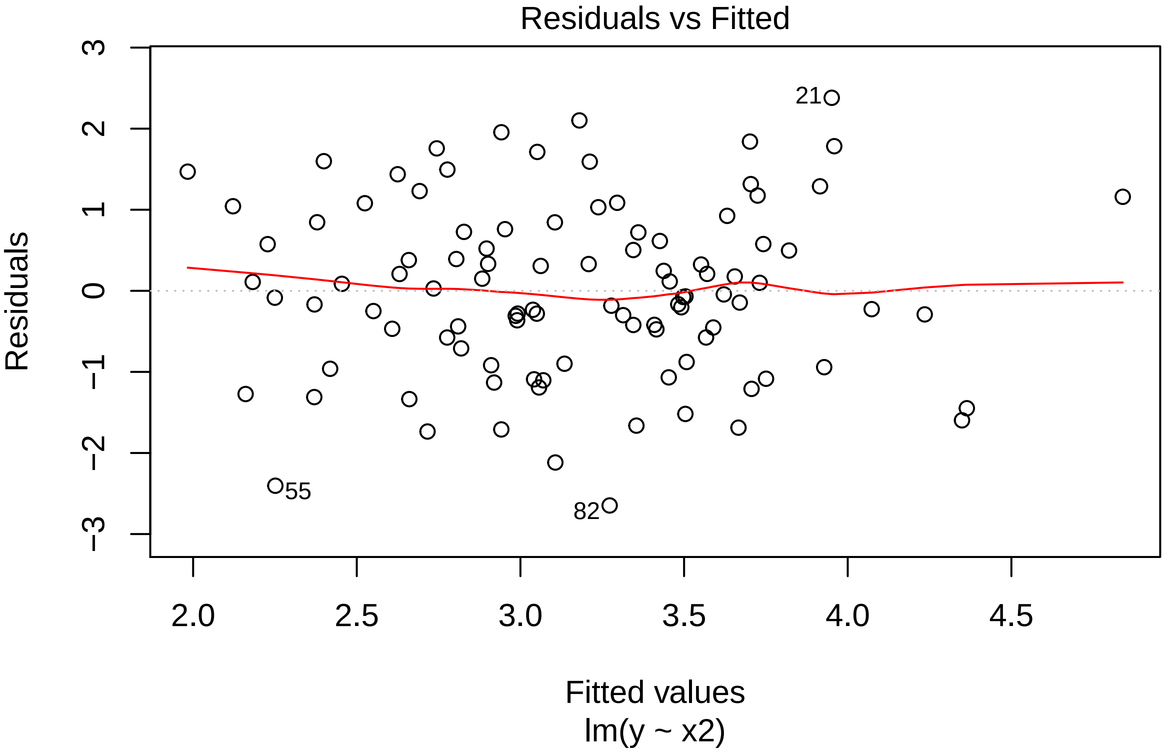

Residuals vs Fitted
Residuals
4
2
0
−2
101
55
82
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1)

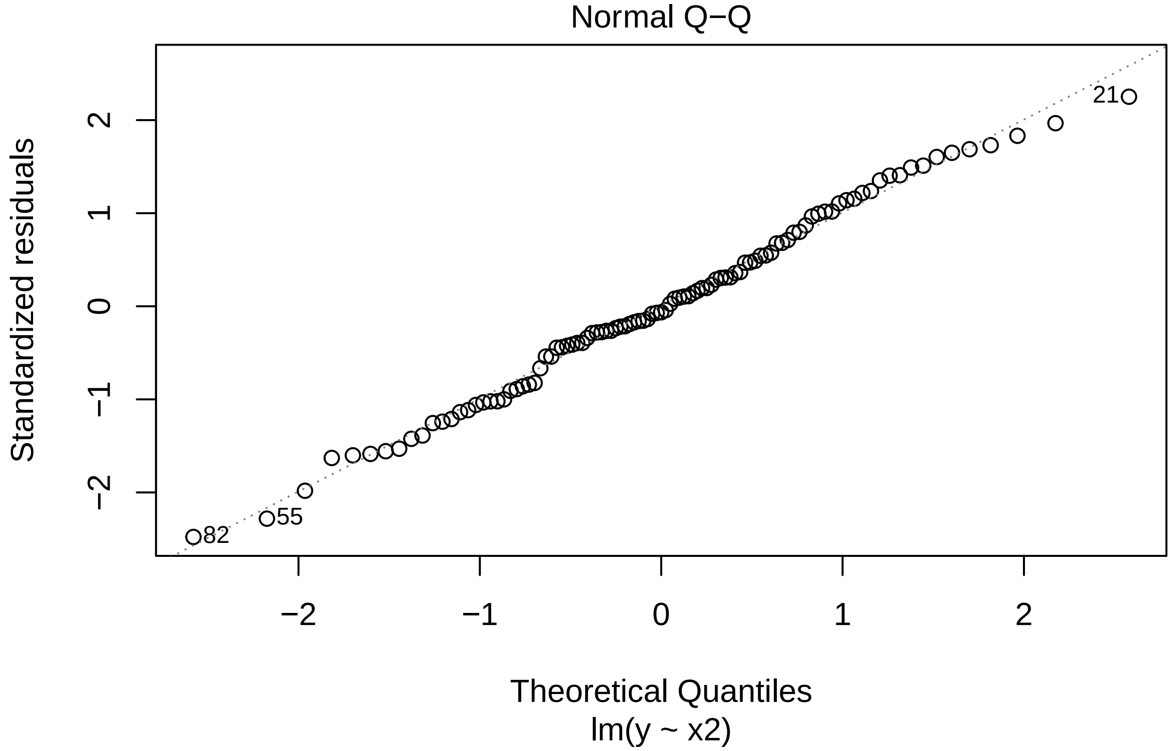

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
101
55
82
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x1)

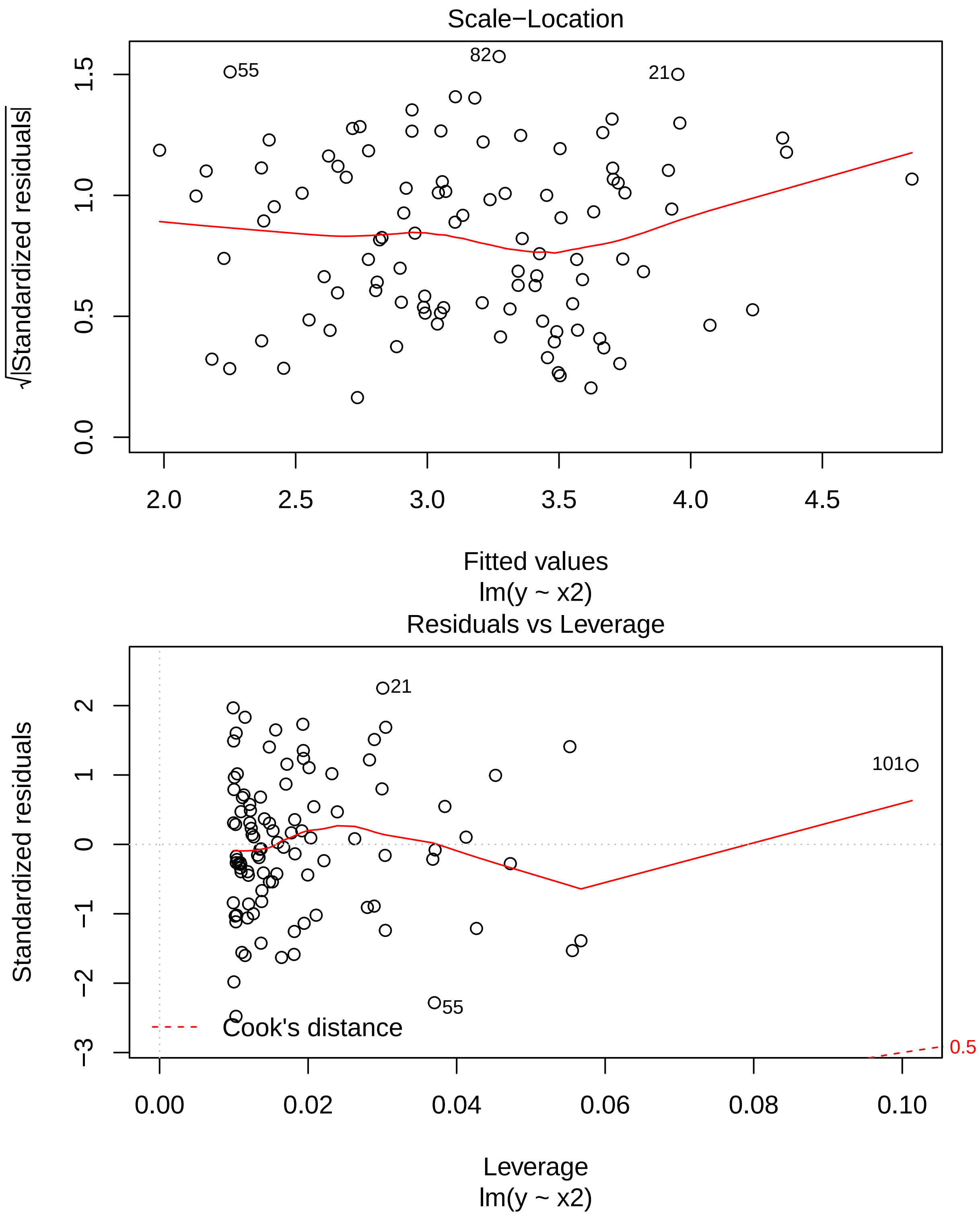

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.5
3.0
3.5
4.0
Fitted values
lm(y ~ x1)
101
55
82


Residuals vs Leverage
Standardized residuals
−3
−2
−1
0
1
2
3
4
0.00
0.01
0.02
0.03
0.04
Leverage
lm(y ~ x1)
101
21
55
Cook's distance

```
plot(r14_e2)
```


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 two predictors, the last point is a high-leverage point. In the model with x1 as sole predictor, the last point is an outlier. In the model with x2 as sole predictor, the last point is a high-leverage point.