Flipped Assignment 14

Group 5

2022/4/6

Input Data and Function Definition

```
setwd('C:/Users/ndong/OneDrive - Texas Tech University/IE 5344 Statistical Data Analysis/Flipped Assignm
data <- read.csv('data-table-B9.csv', header = TRUE)
head(data)
```

```
##     x1 x2   x3    x4    y
## 1 2.14 10 0.34 1.000 28.9
## 2 4.14 10 0.34 1.000 31.0
## 3 8.15 10 0.34 1.000 26.4
## 4 2.14 10 0.34 0.246 27.2
## 5 4.14 10 0.34 0.379 26.1
## 6 8.15 10 0.34 0.474 23.2
```

```
fit.summary<-function(models){
  library(broom)
  fitted<-lapply(models, lm, data=data)
  fitted<-lapply(fitted,glance)
  fitted<-as.data.frame(do.call(rbind.data.frame,fitted))
  fitted<-cbind(models,fitted)
  fitted[order(fitted$AIC), ]
}
```

Candidate Models

Here we have 15 candidate models. We omit the model that only has intercept.

```
models <- c("y ~ x1",
            "y ~ x2",
            "y ~ x3",
            "y ~ x4",
            "y ~ x1 + x2",
            "y ~ x1 + x3",
            "y ~ x1 + x4",
            "y ~ x2 + x3",
            "y ~ x2 + x4",
            "y ~ x3 + x4",
            "y ~ x1 + x2 + x3",
            "y ~ x1 + x2 + x4",
            "y ~ x1 + x3 + x4",
            "y ~ x2 + x3 + x4",
            "y ~ x1 + x2 + x3 + x4")
```

Best Model Selected by AIC

We first choose $y = \beta_0 + \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ be the best model.

```
fit.summary(models)
```

```
## Warning: package 'broom' was built under R version 4.1.3
```

```
##                   models  r.squared adj.r.squared    sigma  statistic
## 15 y ~ x1 + x2 + x3 + x4 0.69137612   0.669718301 5.013645 31.9227066
## 14      y ~ x2 + x3 + x4 0.68068003   0.664163483 5.055630 41.2119984
## 8            y ~ x2 + x3 0.66589408   0.654568456 5.127342 58.7953526
## 11      y ~ x1 + x2 + x3 0.66953663   0.652443697 5.143087 39.1703772
## 9            y ~ x2 + x4 0.62786154   0.615246678 5.411311 49.7715703
## 12      y ~ x1 + x2 + x4 0.63629330   0.617480882 5.395577 33.8230513
## 2                y ~ x2 0.61152526   0.605050681 5.482542 94.4501967
## 5            y ~ x1 + x2 0.61375821   0.600665272 5.512897 46.8770288
## 3                y ~ x3 0.03940456   0.023394638 8.621263  2.4612586
## 10           y ~ x3 + x4 0.06636259   0.034713869 8.571155  2.0968489
## 4                y ~ x4 0.02855900   0.012368314 8.669795  1.7639155
## 13      y ~ x1 + x3 + x4 0.07922153   0.031595062 8.584990  1.6633929
## 6            y ~ x1 + x3 0.04260731   0.010153320 8.679512  1.3128528
## 7            y ~ x1 + x4 0.03925649   0.006688918 8.694687  1.2053858
## 1                y ~ x1 0.00206698  -0.014565237 8.787216  0.1242757
##         p.value df    logLik      AIC      BIC deviance df.residual nobs
## 15 5.818359e-14  4 -185.3217 382.6435 395.4063 1432.788          57   62
## 14 2.145810e-14  3 -186.3779 382.7558 393.3915 1482.445          58   62
## 8  9.007060e-15  2 -187.7811 383.5622 392.0708 1551.089          59   62
## 11 5.757053e-14  3 -187.4413 384.8826 395.5182 1534.178          58   62
## 9  2.166628e-13  2 -191.1232 390.2463 398.7548 1727.655          59   62
## 12 9.055318e-13  3 -190.4127 390.8254 401.4610 1688.511          58   62
## 2  6.228541e-14  1 -192.4550 390.9100 397.2914 1803.496          60   62
## 5  6.491545e-13  2 -192.2763 392.5526 401.0611 1793.130          59   62
## 3  1.219445e-01  1 -220.5201 447.0401 453.4215 4459.570          60   62
## 10 1.319041e-01  2 -219.6376 447.2753 455.7838 4334.417          59   62
## 4  1.891671e-01  1 -220.8681 447.7362 454.1176 4509.921          60   62
## 13 1.848521e-01  3 -219.2077 448.4154 459.0511 4274.720          58   62
## 6  2.767950e-01  2 -220.4165 448.8331 457.3416 4444.701          59   62
## 7  3.068457e-01  2 -220.5248 449.0497 457.5582 4460.257          59   62
## 1  7.256780e-01  1 -221.7022 449.4044 455.7858 4632.910          60   62
```

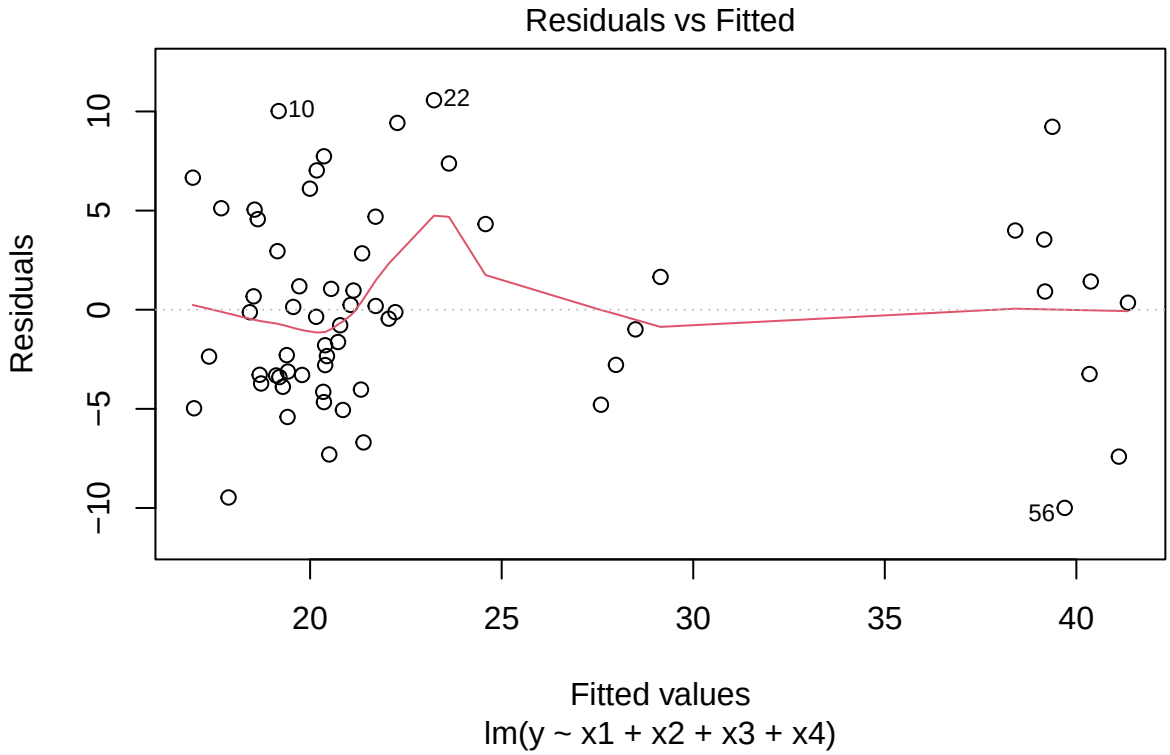
```
bestfit <- lm(y ~ x1 + x2 + x3 + x4, data)
summary(bestfit)
```

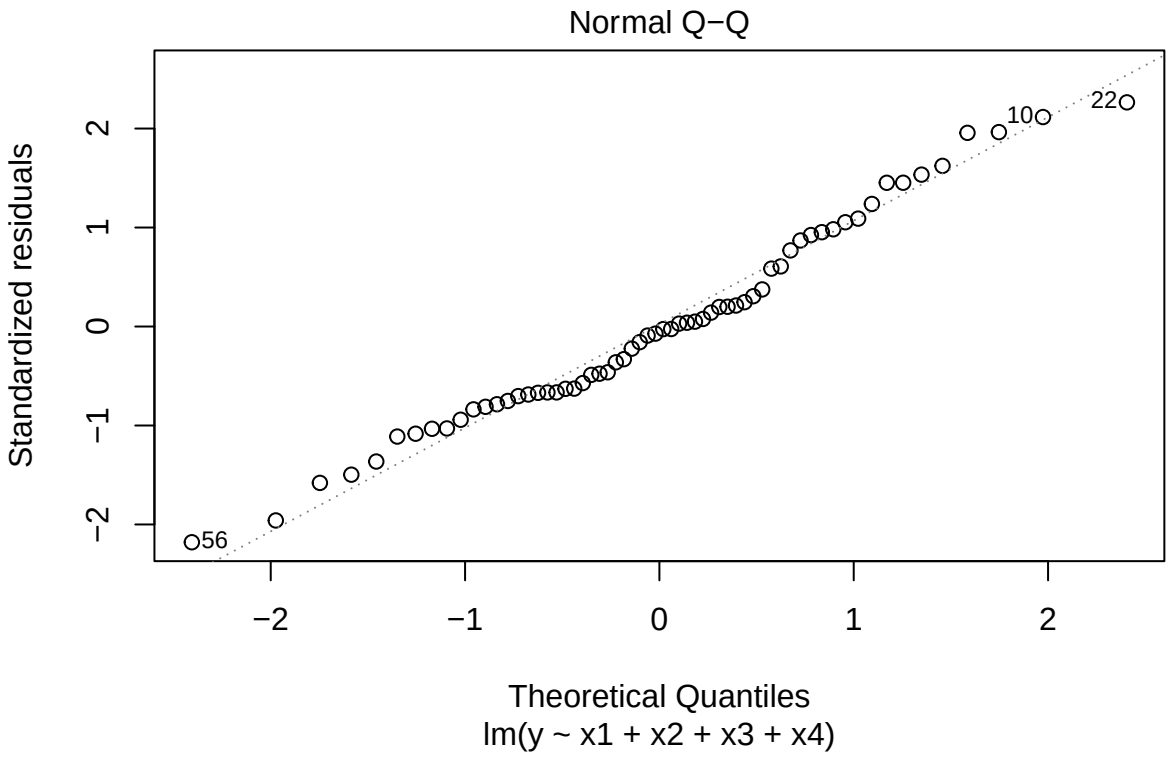
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9958 -3.3092 -0.2419  3.3924 10.5668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.89453    4.32508   1.363  0.17828
## x1          -0.47790    0.34002  -1.406  0.16530
```

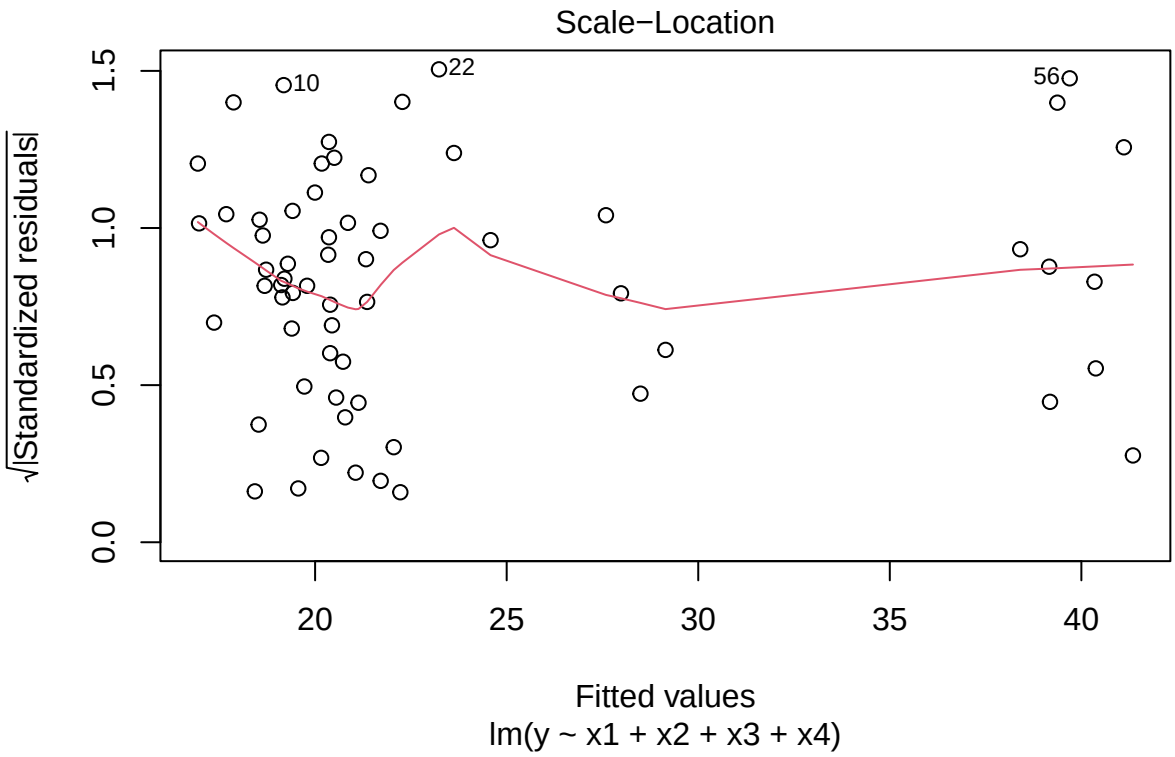
```
## x2           0.18271    0.01718  10.633 3.78e-15 ***
## x3          35.40284   11.09960   3.190  0.00232 **
## x4           5.84391    2.90978   2.008  0.04935 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.014 on 57 degrees of freedom
## Multiple R-squared:  0.6914, Adjusted R-squared:  0.6697
## F-statistic: 31.92 on 4 and 57 DF,  p-value: 5.818e-14
```

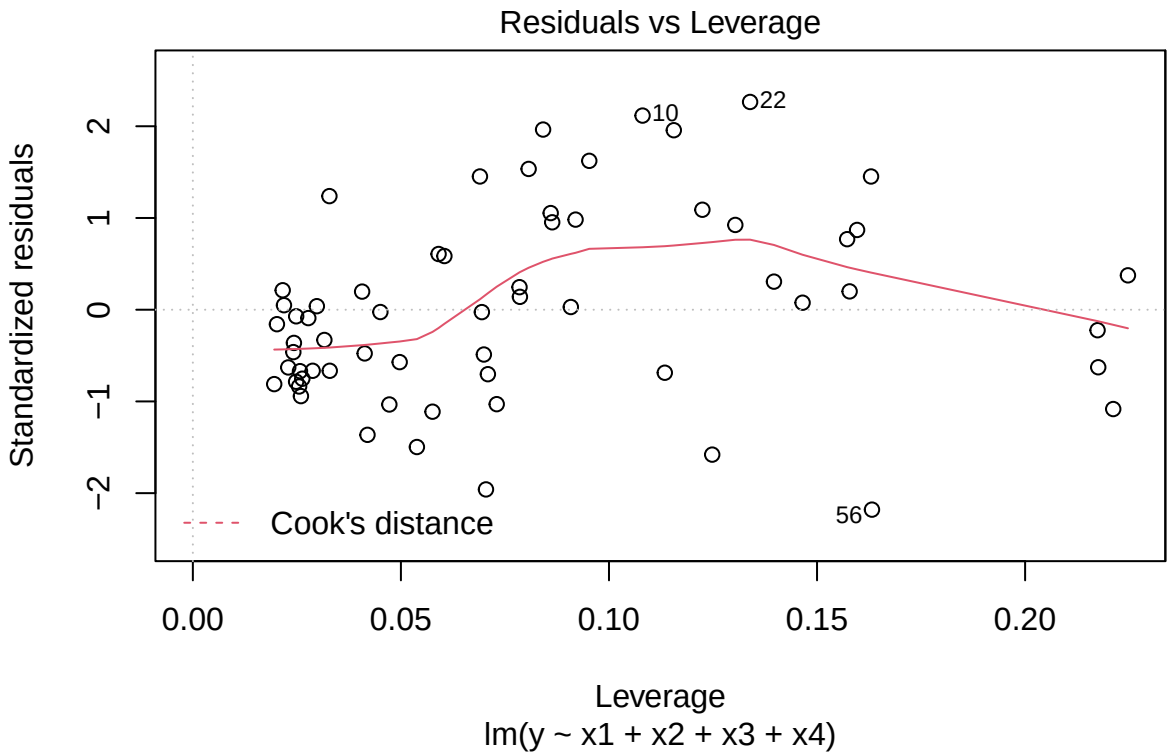
Then, we check the model adequacy:

```
plot(bestfit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + x2 + x3 + x4)
56
10
22
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
22
56
20
25
30
35
40
Fitted values
lm(y ~ x1 + x2 + x3 + x4)



This model is adequate. However, from the summary of this model, x_1 is not significant for its p-value is greater than 0.05. So we pick the models that have the second and the third smallest AIC.

```
bestfit_opt1 <- lm(y ~ x2 + x3 + x4, data)
summary(bestfit_opt1)
```

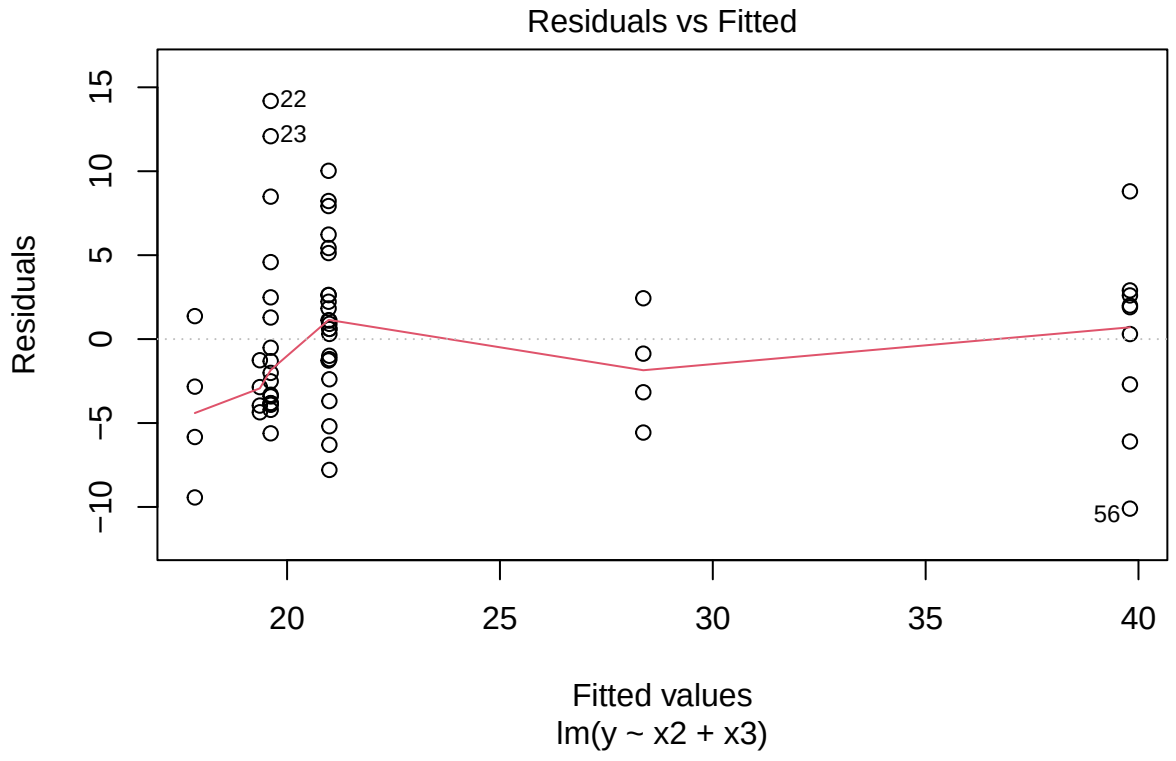
```
##
## Call:
## lm(formula = y ~ x2 + x3 + x4, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2730  -3.4598  -0.5632   2.7904  12.3370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.64065    4.26751   1.087  0.28134
## x2           0.18302    0.01733  10.563 3.92e-15 ***
## x3          34.62435   11.17861   3.097  0.00301 **
## x4           4.56878    2.78788   1.639  0.10667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.056 on 58 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6642
## F-statistic: 41.21 on 3 and 58 DF,  p-value: 2.146e-14
```

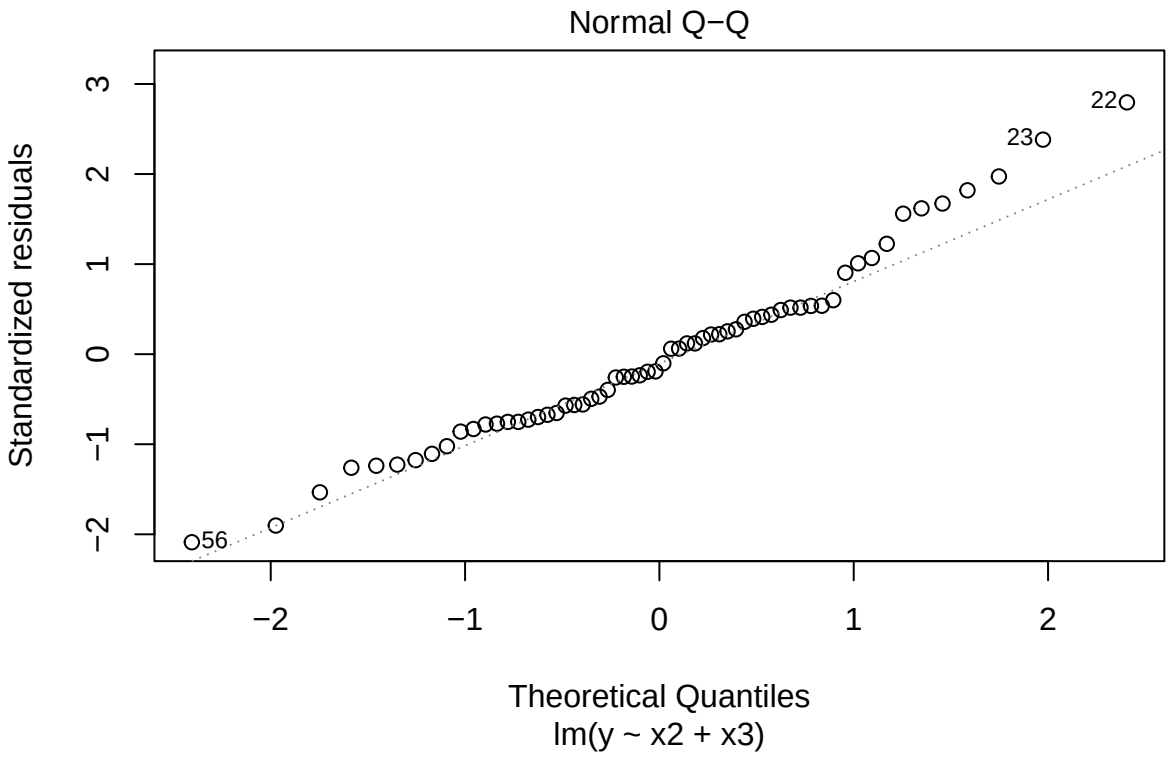
```
bestfit_opt2 <- lm(y ~ x2 + x3, data)
summary(bestfit_opt2)
```

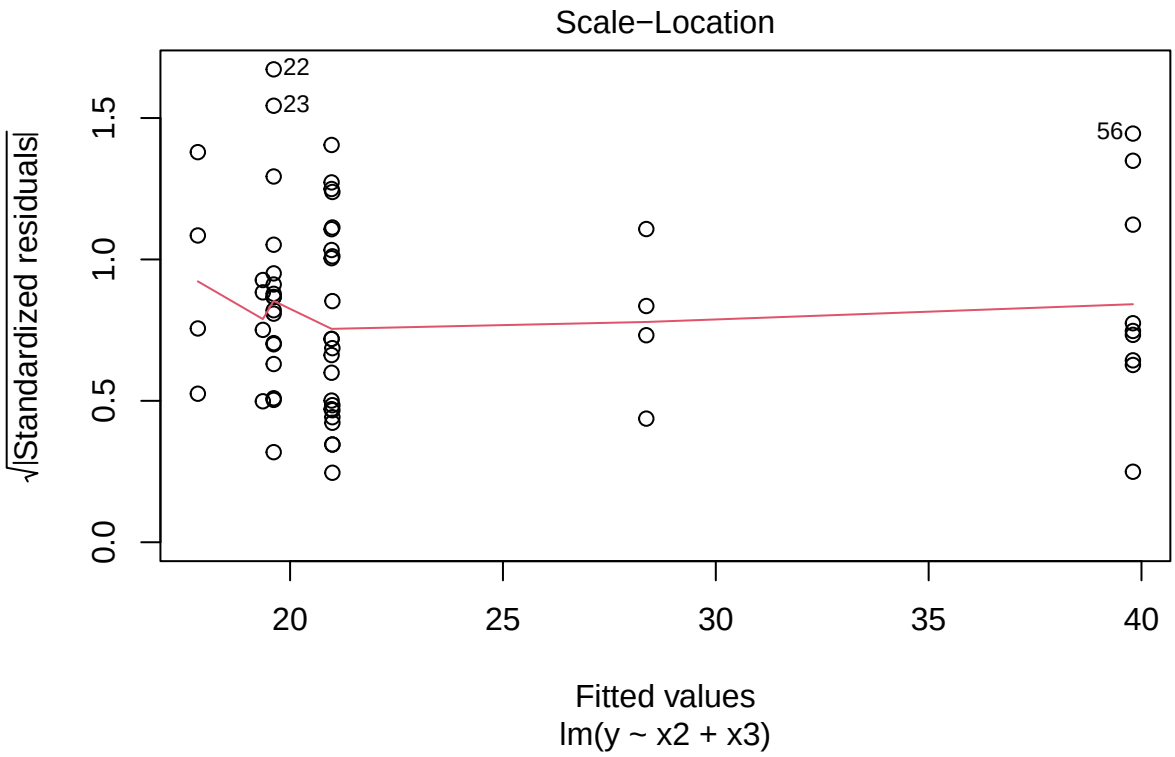
```
##
## Call:
## lm(formula = y ~ x2 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0994  -3.6236  -0.6911   2.4722  14.1854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.18971    4.03031   1.784  0.07958 .
## x2           0.18456    0.01755  10.518 3.74e-15 ***
## x3          35.11616   11.33308   3.099  0.00298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.127 on 59 degrees of freedom
## Multiple R-squared:  0.6659, Adjusted R-squared:  0.6546
## F-statistic:  58.8 on 2 and 59 DF,  p-value: 9.007e-15
```

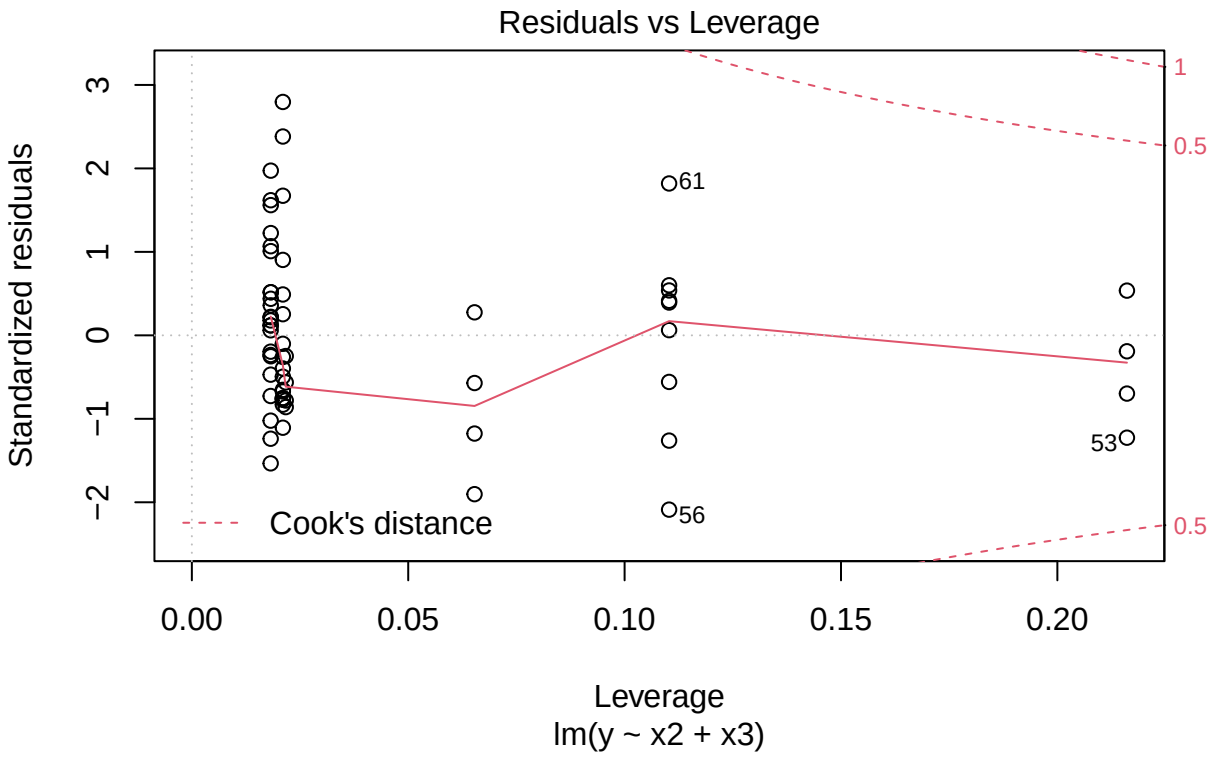
The model that has the second smallest AIC has similar issue of previous model, x_4 is not significant. The model that has the third smallest AIC could be the best model. So, we check its model adequacy.

```
plot(bestfit_opt2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x2 + x3)
22
23
56
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
22
23
56
20
25
30
35
40
Fitted values
lm(y ~ x2 + x3)



This model is adequate. So, we pick it to be our best model for the following reasons:

(1) Its AIC is 383.5622, which is slightly greater than the smallest AIC among candidate models.
(2) Its adjusted R^2 is 0.6546, which is slightly less than the largest adjusted R^2 among candidate models.
(3) Its p-value of whole regression model is the smallest among the candidate models.
(4) All of its predictor variables are significant.