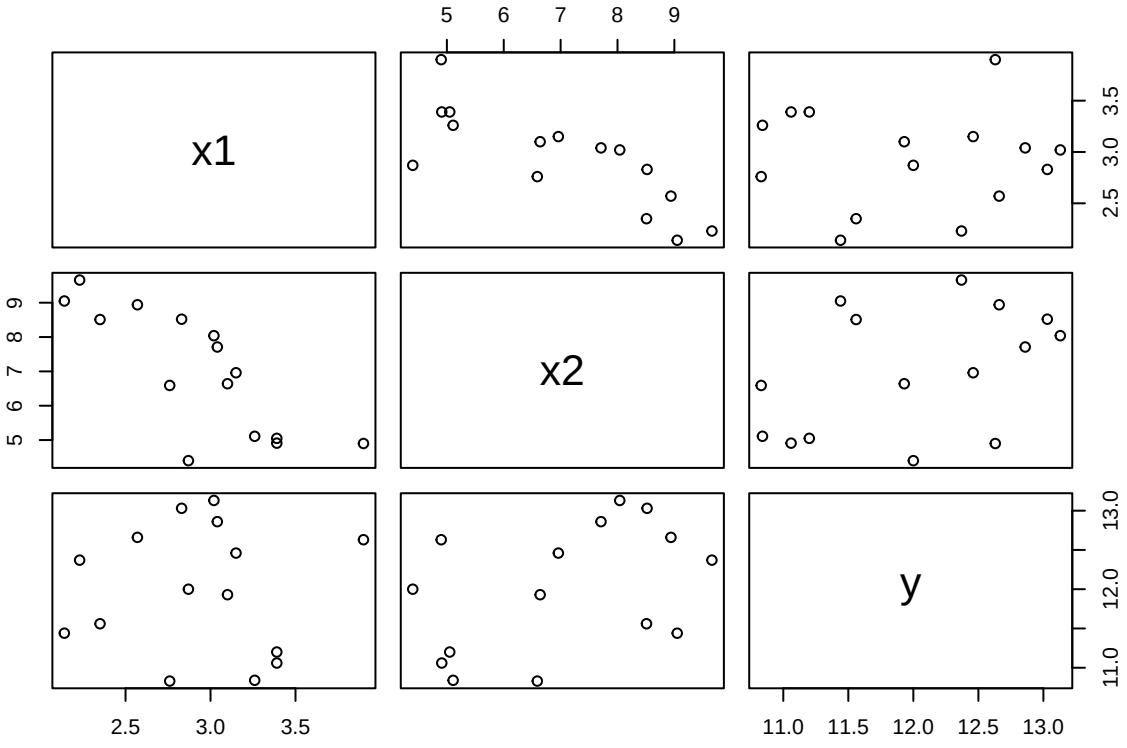
hw5

2022-11-03

##Problem 1

```
dat = data.frame(
     x1=c(2.23,2.57,2.87,3.1,3.39,2.83,3.02,2.14,3.04,3.26,3.39,2.35,
       2.76,3.9,3.15),
     x2=c(9.66,8.94,4.4,6.64,4.91,8.52,8.04,9.05,7.71,5.11,5.05,8.51,
       6.59,4.9,6.96),
     y=c(12.37,12.66,12,11.93,11.06,13.03,13.13,11.44,12.86,10.84,
       11.2,11.56,10.83,12.63,12.46))
#(a)
plot(dat)
```



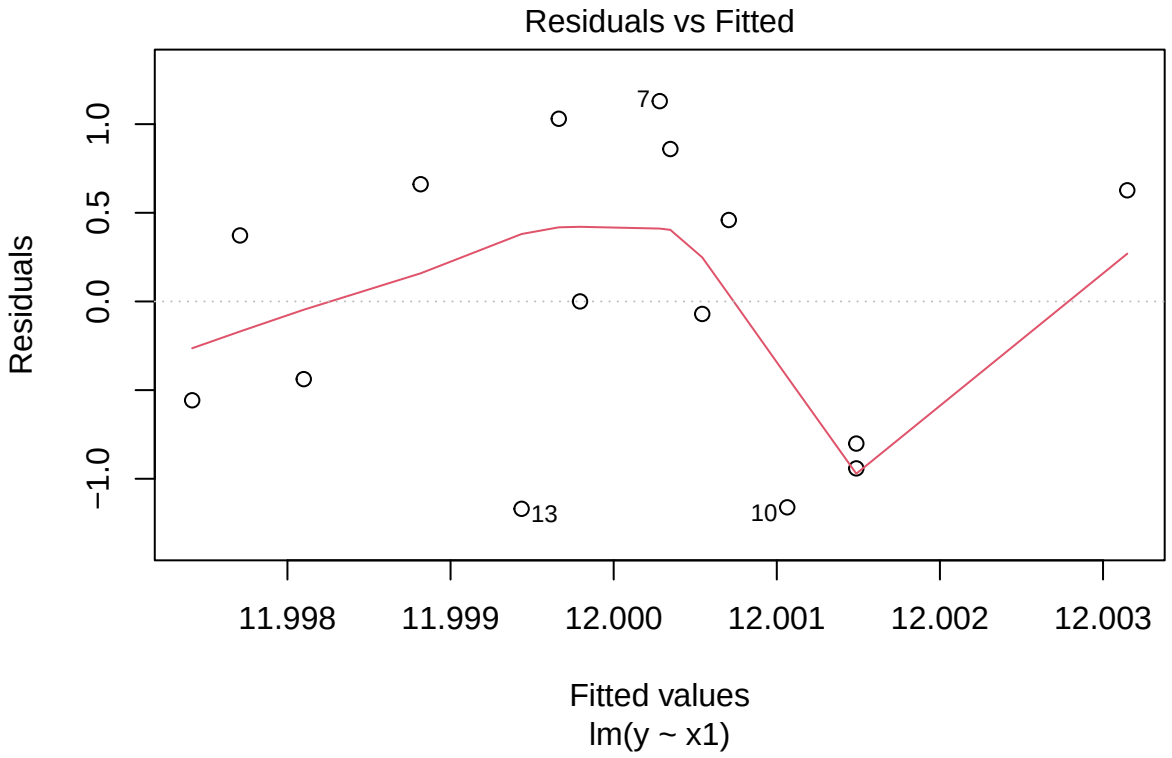
From the scatter plot, I don't see any clear association between y and x1 and between y and x2. And there is a negative association between x1 and x2.

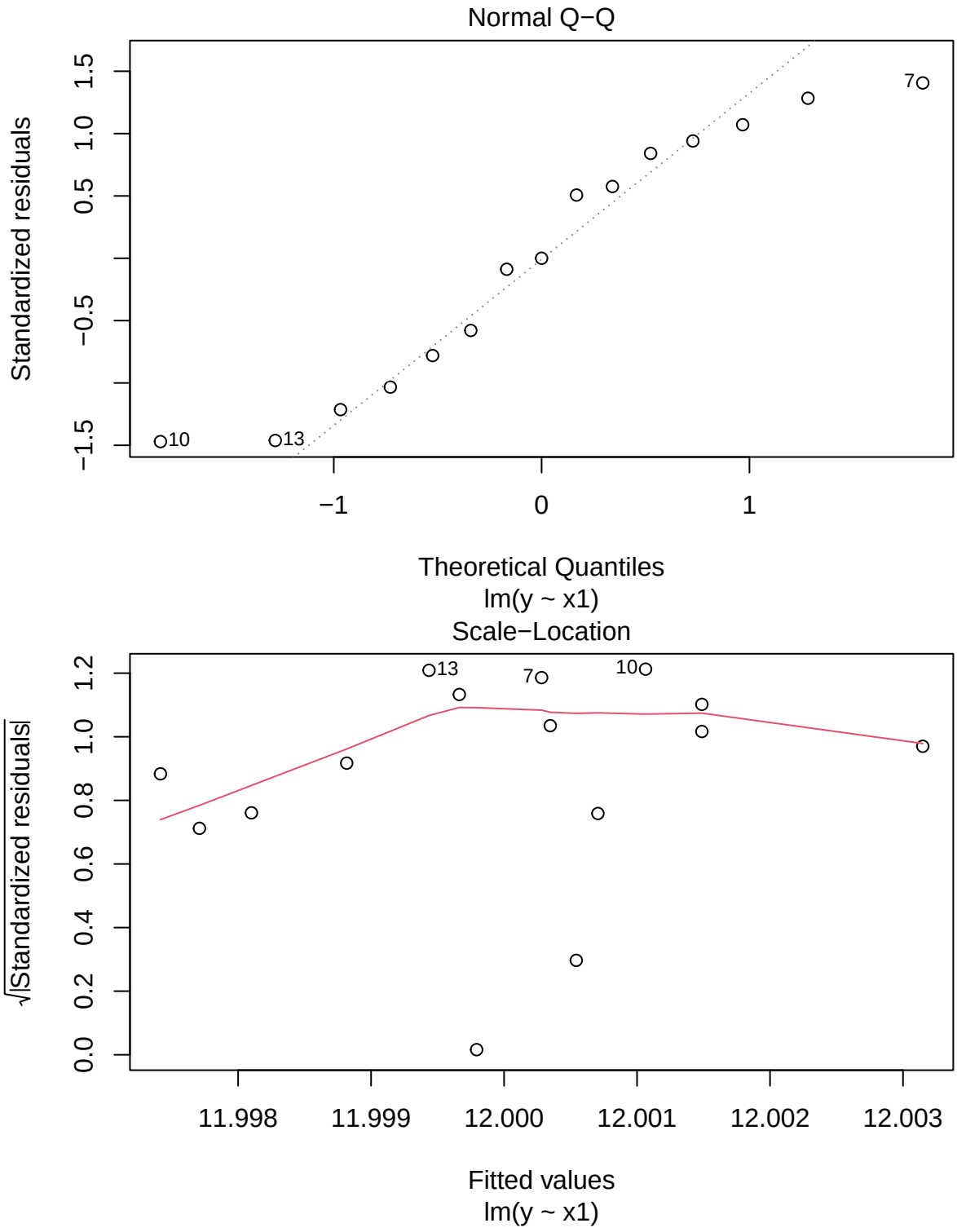
```
#(b)
fit1 <- lm(y~x1,data = dat)
summary(fit1)
```

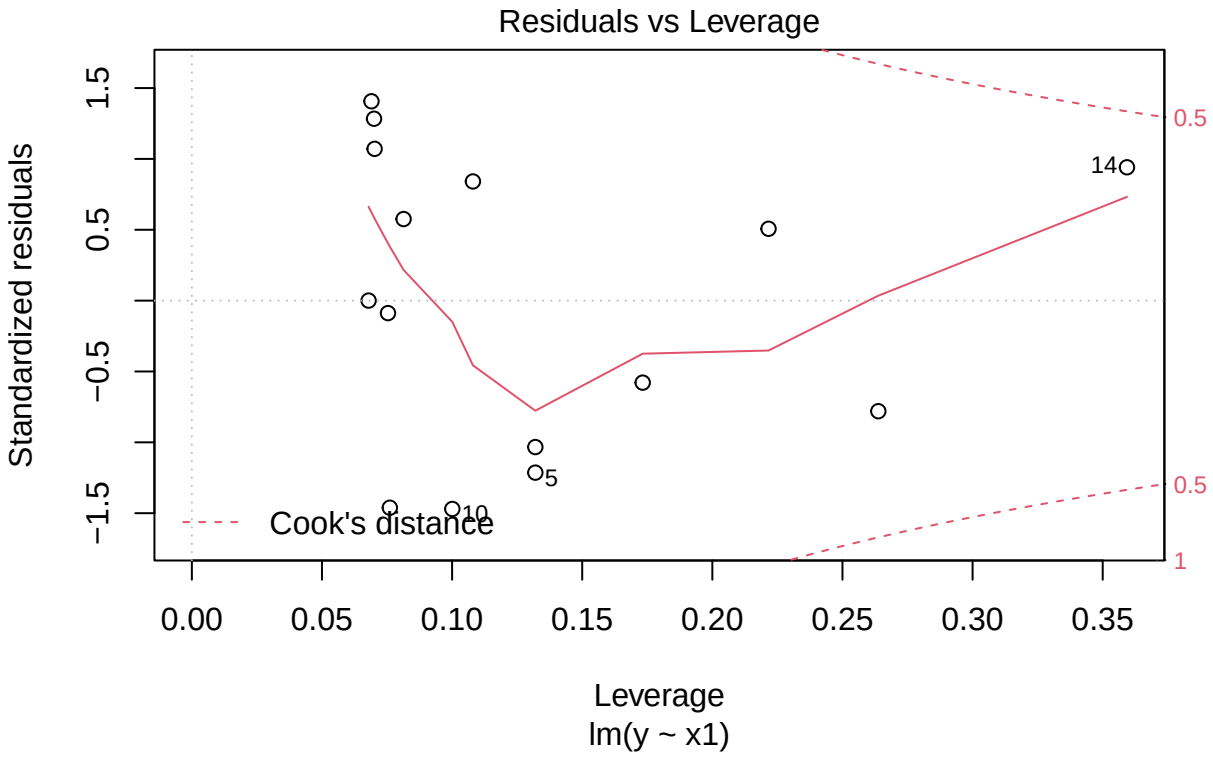
```
##
## Call:
## lm(formula = y ~ x1, data = dat)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16944 -0.67945  0.00021  0.64402  1.12972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.990446   1.383341   8.668  9.2e-07 ***
## x1           0.003257   0.465866   0.007    0.995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8324 on 13 degrees of freedom
## Multiple R-squared:  3.76e-06,   Adjusted R-squared:  -0.07692
## F-statistic: 4.888e-05 on 1 and 13 DF,  p-value: 0.9945
```

```
plot(fit1)
```




Normal Q−Q
Standardized residuals
−1.5
−0.5
0.5
1.0
1.5
10
13
7
−1
0
1
Theoretical Quantiles
lm(y ~ x1)
Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
13
7
10
11.998
11.999
12.000
12.001
12.002
12.003
Fitted values
lm(y ~ x1)

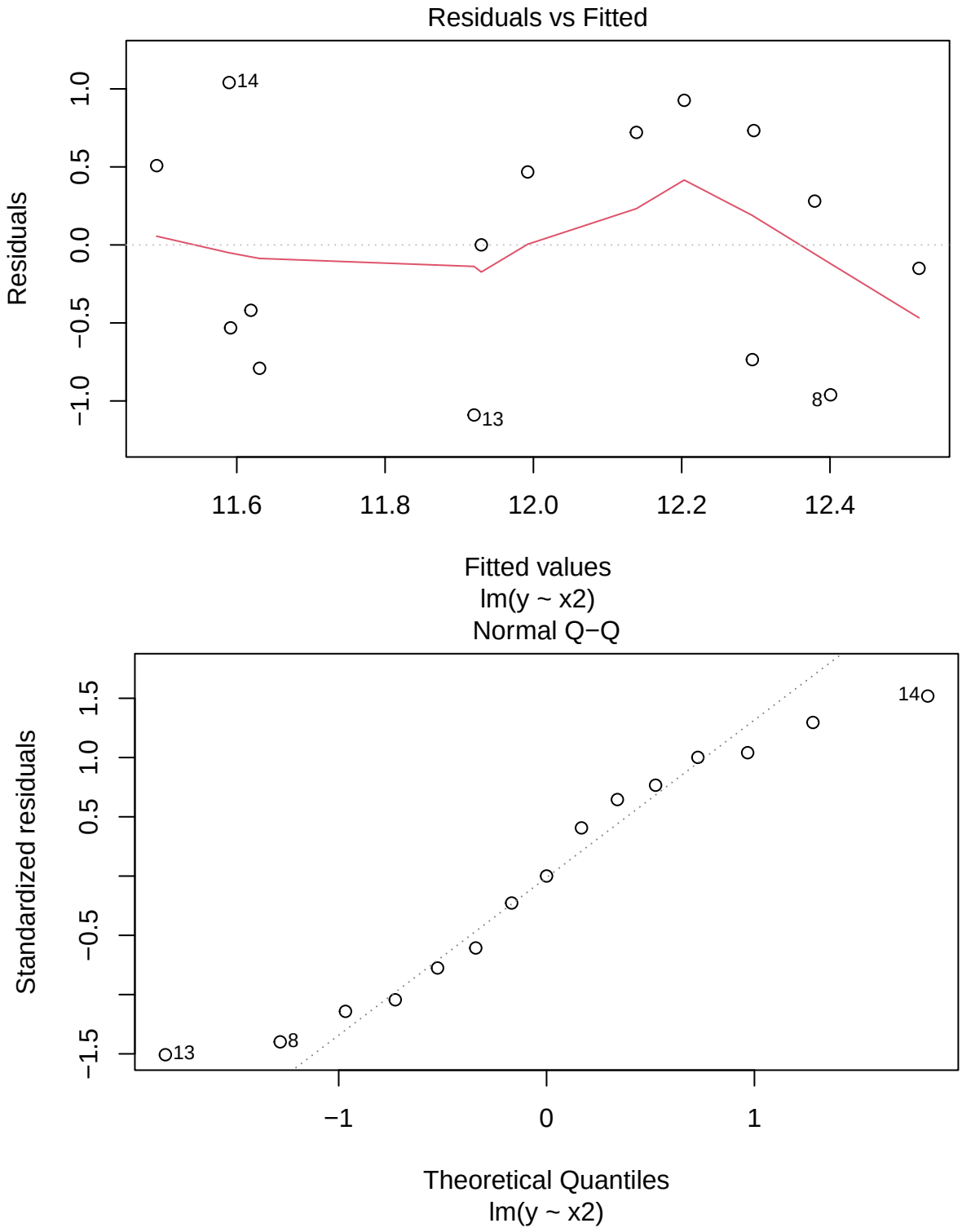


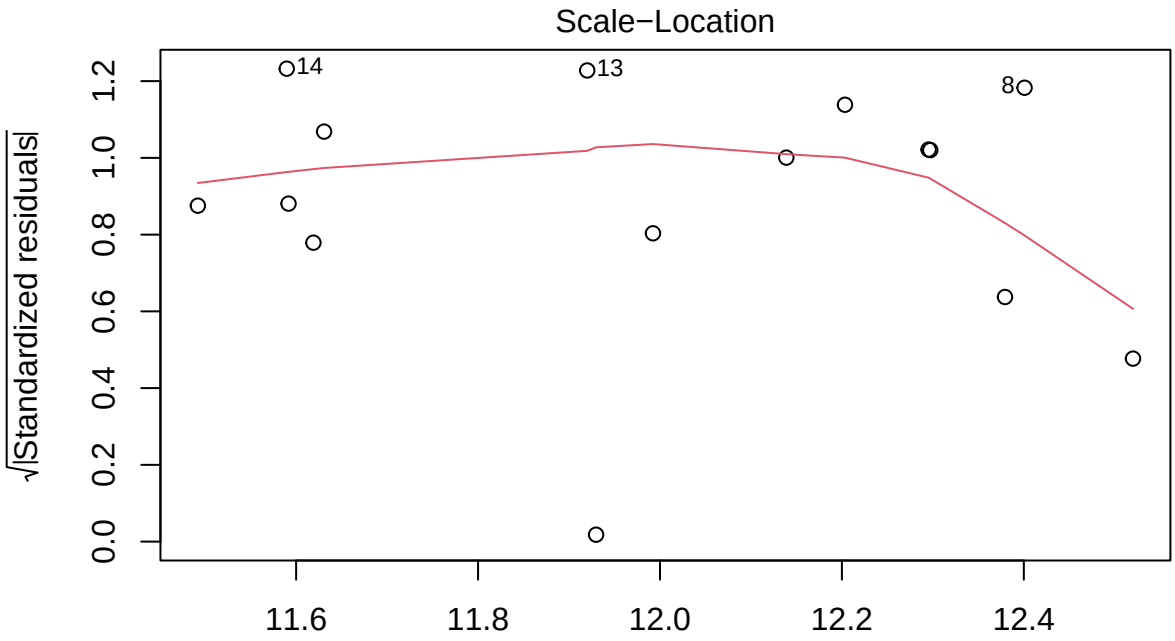
The overall model is not significant. Residuals do not distribute randomly above and below residuals = 0 line. The residual plot shows that there's no linear relationship between y and x1. And there exist heteroscedasticity. From the qq plot, we see that data is not normally distributed.

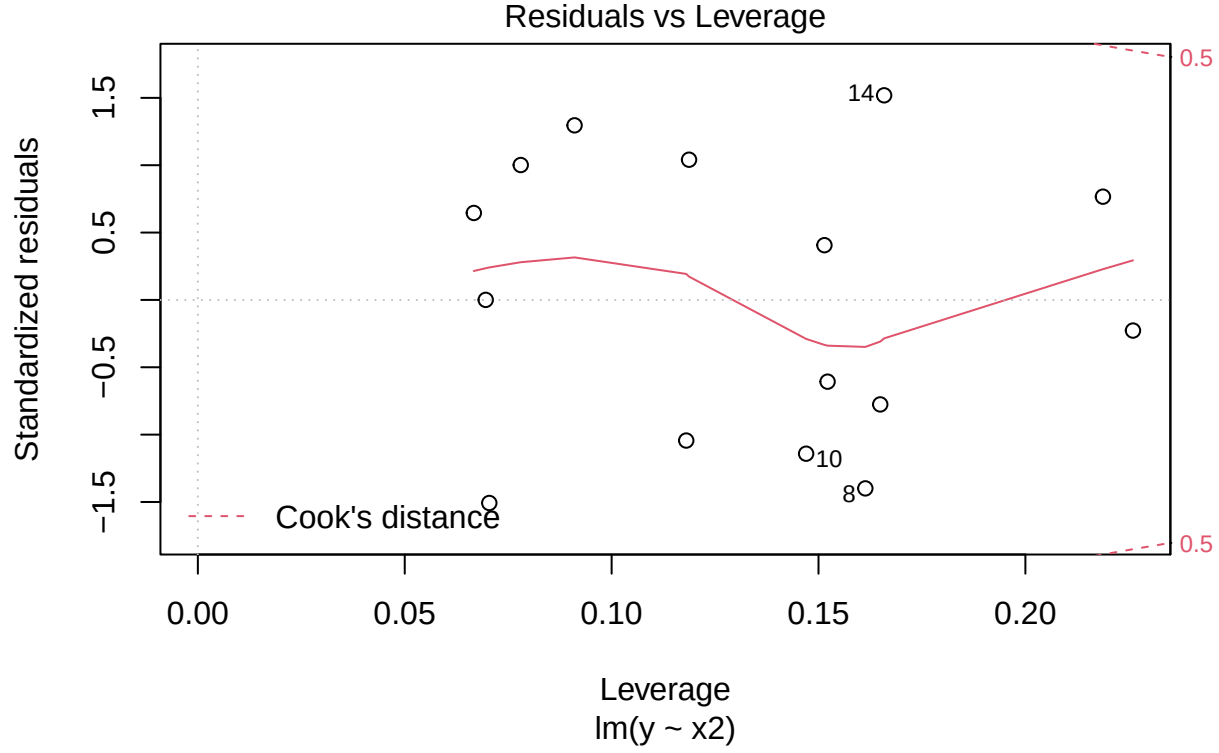
```
#(c)
fit2 = lm(y~x2, data = dat)
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ x2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08999 -0.63345  0.00023  0.61458  1.04033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.6319     0.8109  13.111 7.18e-09 ***
## x2            0.1955     0.1125   1.737    0.106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7499 on 13 degrees of freedom
## Multiple R-squared:  0.1884, Adjusted R-squared:  0.126
## F-statistic: 3.018 on 1 and 13 DF,  p-value: 0.106
```

```
plot(fit2)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
14
13
8
11.6
11.8
12.0
12.2
12.4
Fitted values
lm(y ~ x2)
Normal Q−Q
Standardized residuals
1.5
1.0
0.5
−0.5
−1.5
14
13
8
−1
0
1
Theoretical Quantiles
lm(y ~ x2)



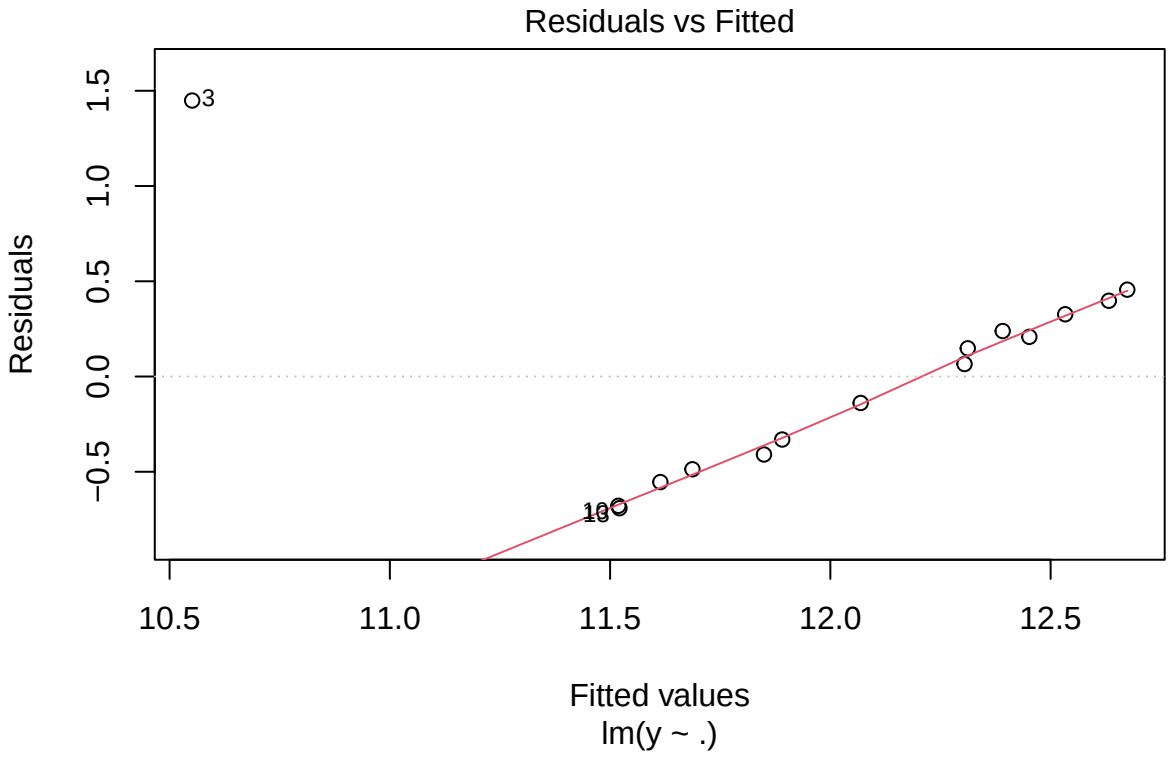


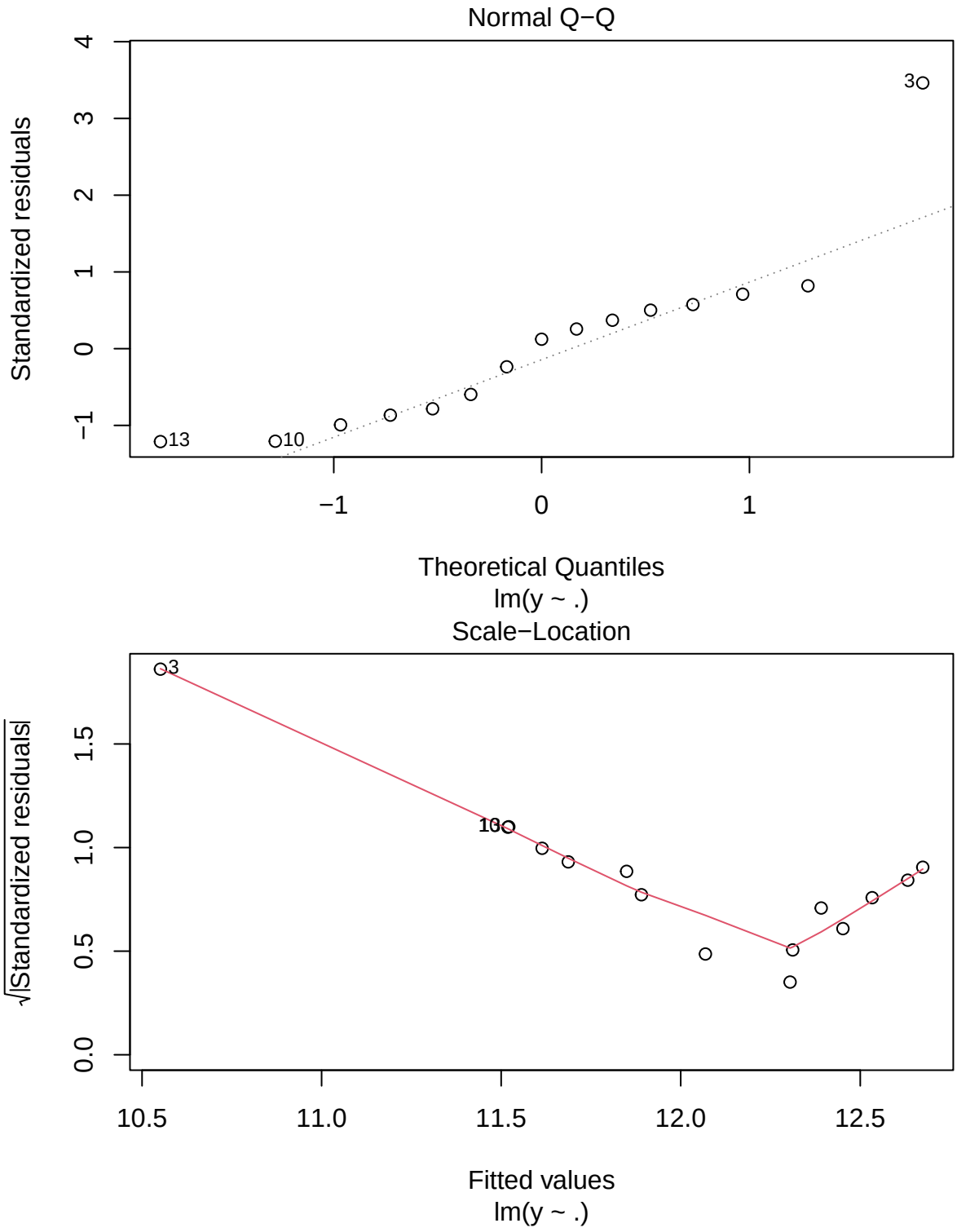
The overall model is not significant. Similarly, residuals do not distribute randomly above and below residuals = 0 line. The residual plot shows that there's no linear relationship between y and x2. Similarly, there exist heteroscedacsticity and the data is not normally distributed.

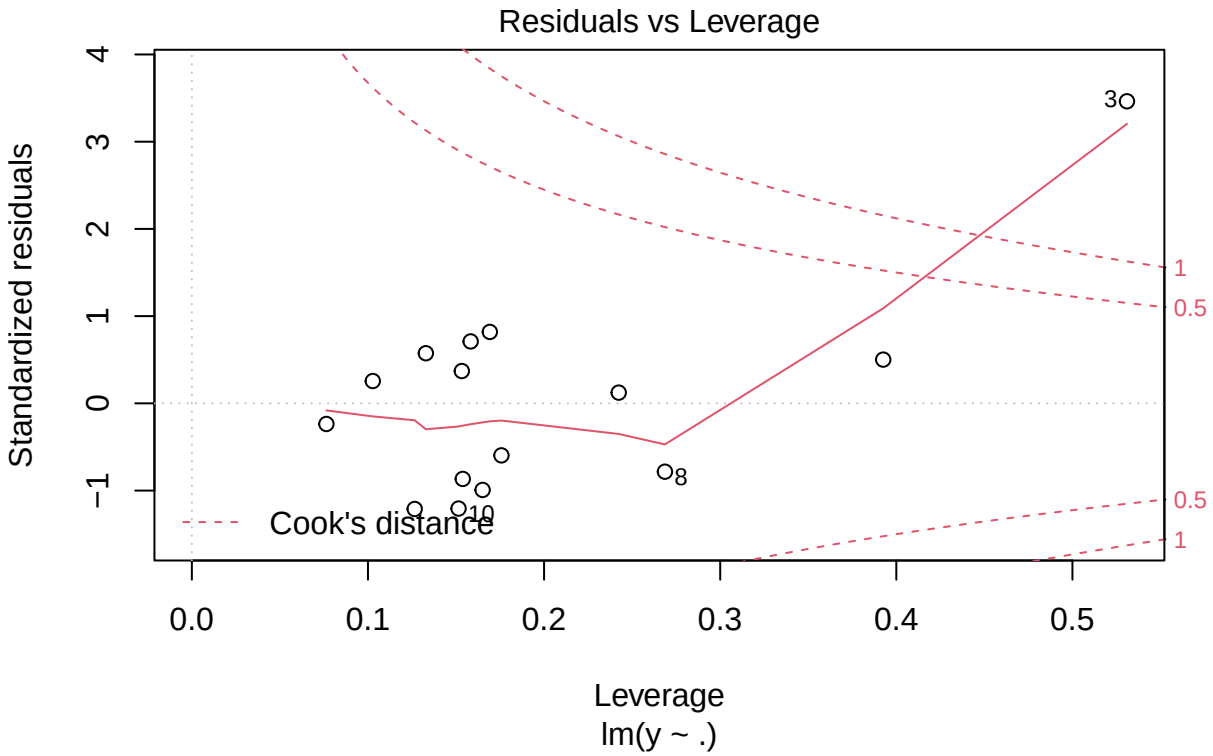
```
#(d)
fit3 <- lm(y~., data = dat)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69127 -0.44813  0.06541  0.28281  1.44873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8610     2.5440   1.518   0.1550
## x1            1.5339     0.5566   2.756   0.0174 *
## x2            0.5200     0.1492   3.485   0.0045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6108 on 12 degrees of freedom
## Multiple R-squared:  0.503,  Adjusted R-squared:  0.4202
## F-statistic: 6.073 on 2 and 12 DF,  p-value: 0.01507
```

```
plot(fit3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ .)
Scale−Location
√|Standardized residuals|
Fitted values
lm(y ~ .)



The overall model is significant since the p value of f test is $0.01507 < 0.05$. Although the f test shows the overall model is significant, the residuals plot does not show constant variance or linear relationship. The residuals lie on a straight line and do not distribute randomly above and below residuals = 0 line. There's a clear pattern in the residual plot.

##(e) The problem tells me that forward variable selection and backward variable selection do not necessarily generate the same result. They may not select the same variables. If we implement forward variable selection technique, no variable will be selected because neither x1 nor x2 is significant to y, as illustrated in (a) and (b). However, if we implement backward variable selection, both x1 and x2 will be selected and included in the model because they both have p-value < 0.5 in (d). And no variable will be dropped.