STAT-310-Lab 6-Outliers and Influential Points

Prabh Talwar

2022-11-24

```
# loading libraries

library(tidyverse)
library(faraway)
library(olsrr)
```

Load the 'savings' data set from the faraway package.

```
savings <- faraway::savings

View(savings)
```

(a) Fit a model for the savings rate against all other variables.

```
model <- lm(sr ~ ., data = savings)
summary(model)
```

```
##
## Call:
## lm(formula = sr ~ ., data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

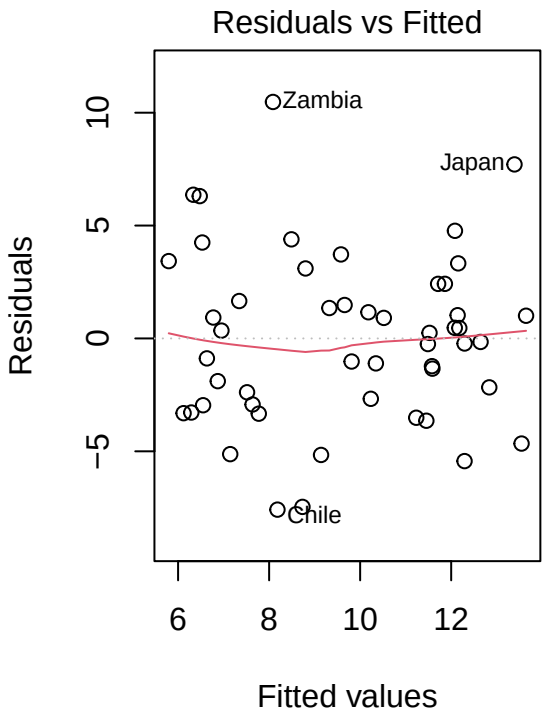
(b) Remove any nonsignificant features and rerun the model with the remaining features.

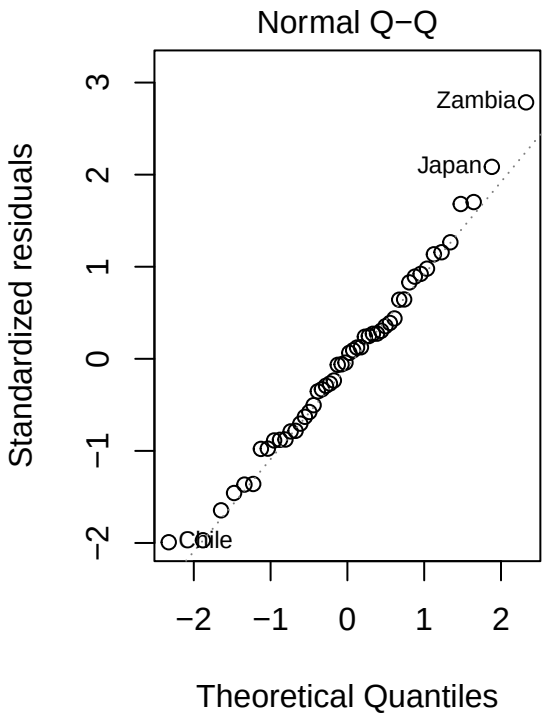
```
model_1 <- lm(sr ~ pop15 + ddpi, data = savings)
summary(model_1)
```

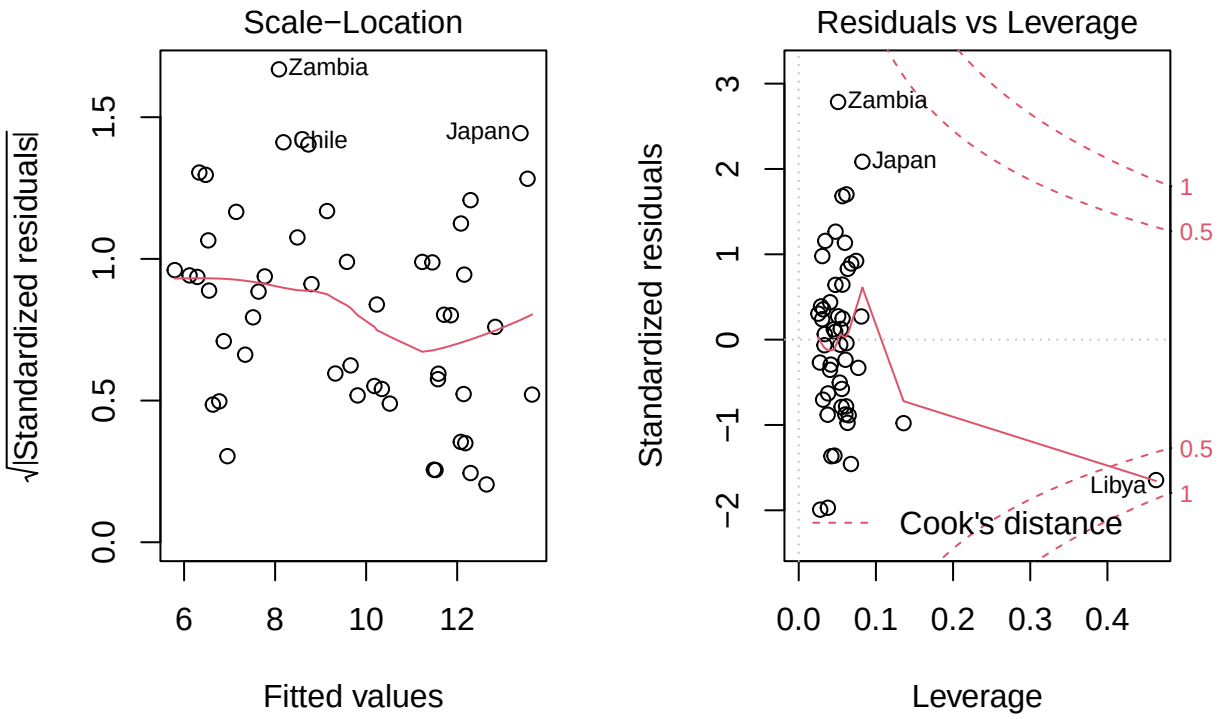
```
##
## Call:
## lm(formula = sr ~ pop15 + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5831 -2.8632  0.0453  2.2273 10.4753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
## pop15       -0.21638    0.06033  -3.586 0.000796 ***
## ddpi         0.44283    0.19240   2.302 0.025837 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.861 on 47 degrees of freedom
## Multiple R-squared:  0.2878, Adjusted R-squared:  0.2575
## F-statistic: 9.496 on 2 and 47 DF,  p-value: 0.0003438
```

(c) Use the plot() function on the model in part (b) and comment on the four graphs with respect to what they're used for.

```
par(mfrow=c(1,2))
plot(model_1)
```


Residuals vs Fitted
Residuals
Fitted values
Zambia
Japan
Chile
10
5
0
−5
6
8
10
12


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Zambia
Japan
Chile
3
2
1
0
−1
−2
−2
−1
0
1
2



Plot 1: The Residual v/s Fitted plot is used to check the linearity and the constant variance assumption. Here, the mean of the residuals are centered at zero so the linearity assumption is valid. The spread of the residuals is same so the constant variance assumption is valid as well.

Plot 2: The Q-Q plot is used to check the normality of the errors. Here, the points closely follows a straight line that suggests that the data comes from a normal distribution.

Plot 3: The Scale Location plot shows if the residuals are spread equally along the range of predictors. It checks the assumption of constant variance. Here, the residuals appears to be spread randomly.

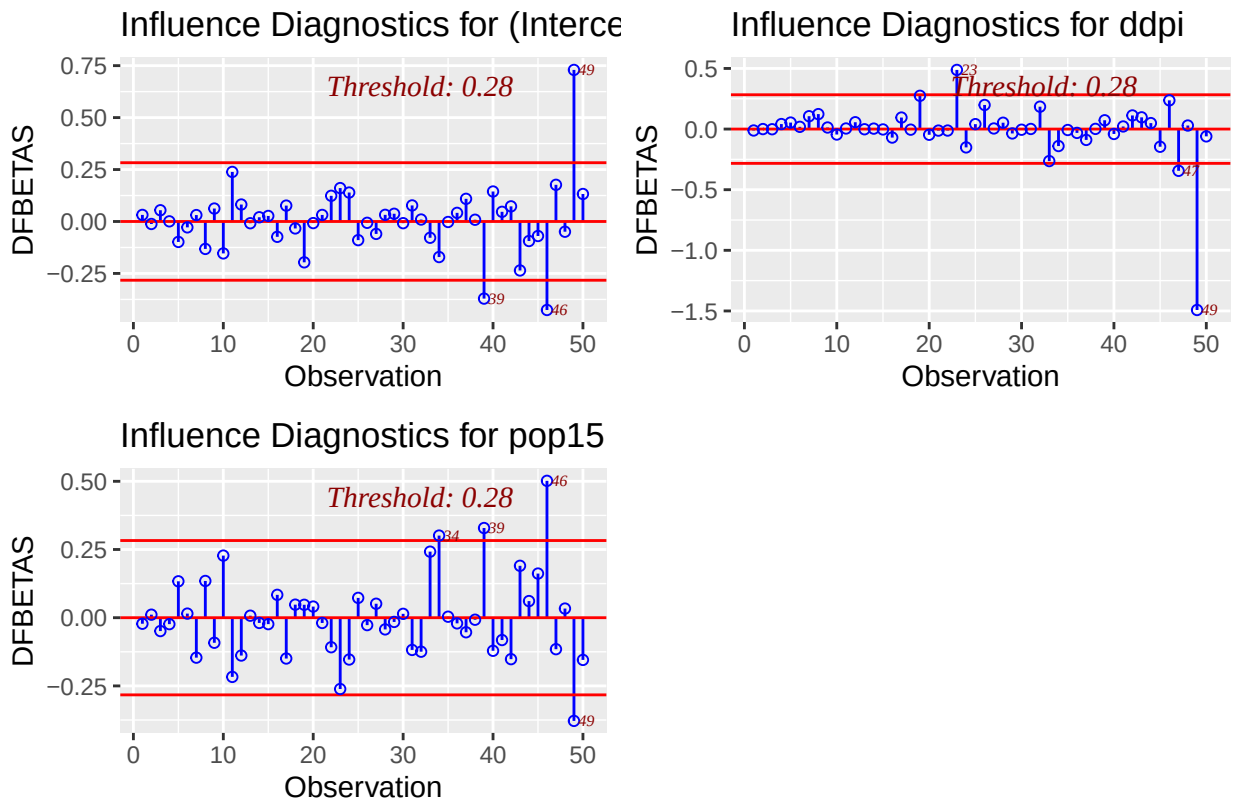
Plot 4: The Residual v/s Leverage plot helps to find the influential points in a data set. If any data point falls outside of the cook's distance the it is considered to be an influential point. Here, observation "Libya"in the bottom right corner falls outside of the re dashed line and this indicates that it is a an influential point.

(d) Use the tools in this lab to identify any outliers or influential values. Justify your answers.

Looking for Influential Values with DFBETAS

```
dfbetas <- as.data.frame(dfbetas(model_1))
```

```
ols_plot_dfbetas(model_1)
```

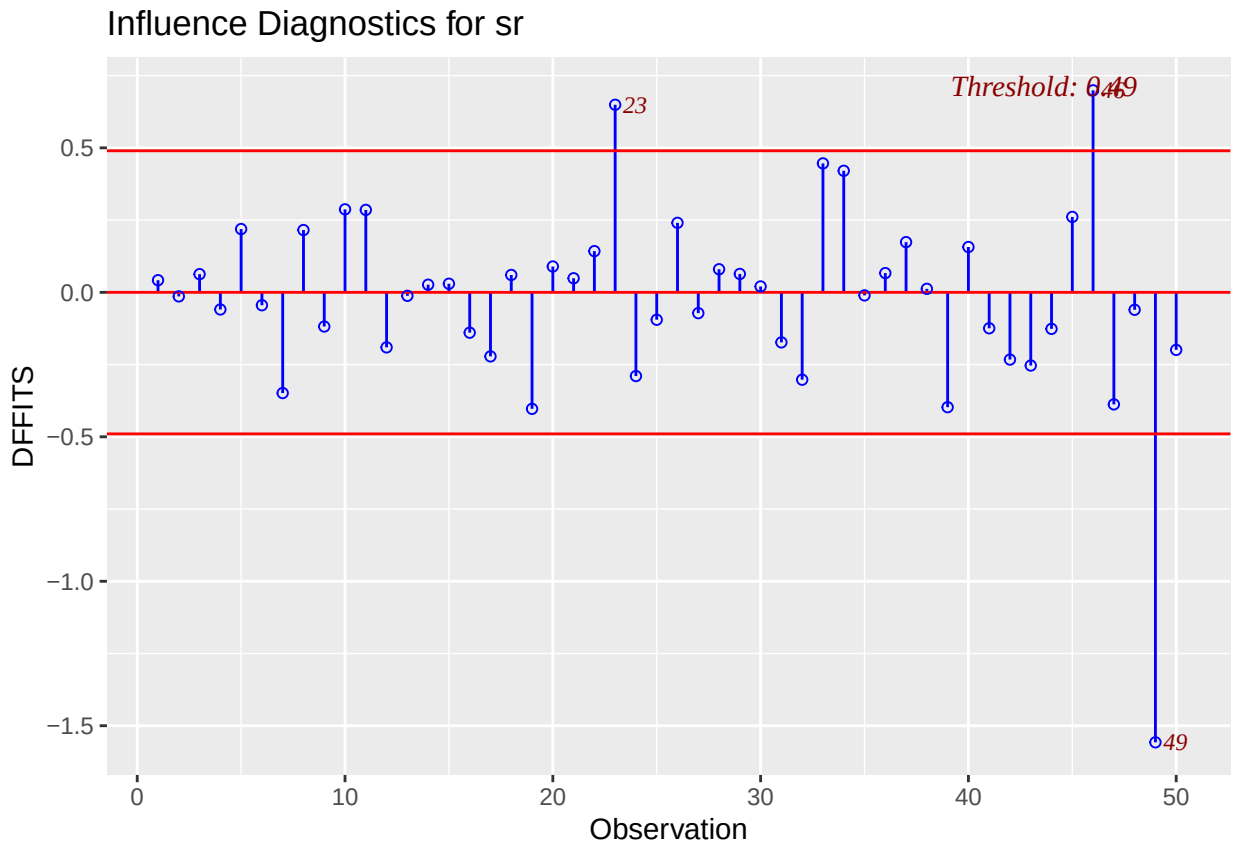


From the first plot we can see that three observations(39,46,49) exceed the absolute value of the threshold of 0.28, in the second plot we can see that two observations(23,49) exceed the absolute value of the threshold and in the third plot we can see that two observations(46,49) exceed the absolute value of the threshold.

Looking for Influential Values with DFFITS

```
dffits <- as.data.frame(dffits(model_1))
```

```
ols_plot_dffits(model_1)
```

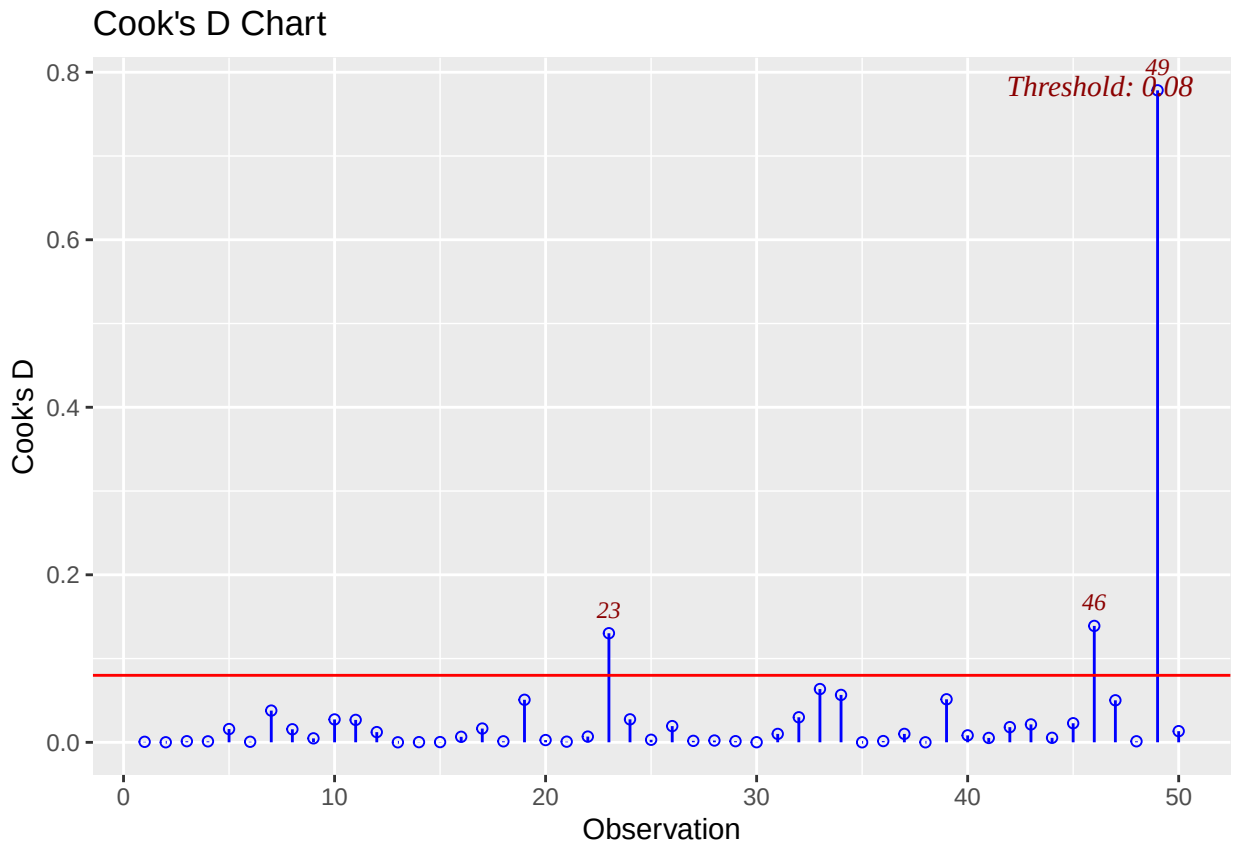


From the above plot we can see that three observations(23,46,49) exceed the absolute value of the threshold of 0.49.

Looking for Influential Values with Cook's Distance

```
cooks.d <- as.data.frame(cooks.distance(model_1))
```

```
ols_plot_cooksd_chart(model_1)
```



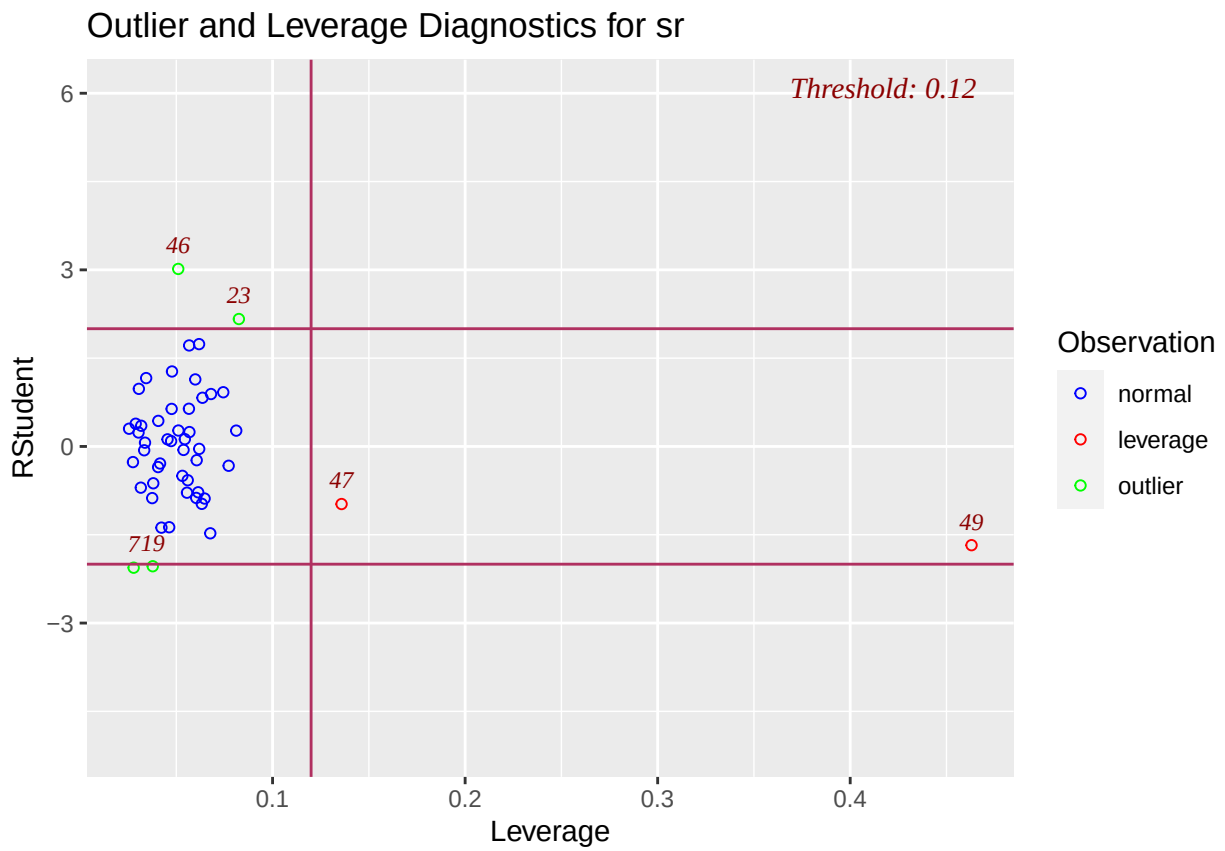
From the above plot we can see that three observations(23,46,49) exceed the absolute value of the threshold of 0.08.

From the above tools, we can notice that the three values are shown continuously that are 23, 46 and 49.

Detecting Outliers

Studentized Residuals vs Leverage Plot

```
ols_plot_resid_lev(model_1)
```



From the above plot we can see that three observations(47,49) exceed the absolute value of the threshold of 0.12. These two observations seem to be potential leverage points.

(e) Remove any outliers or influential values from your data and rerun the model in part (b).

Removing the observations 23,46,49.

```
savings_new <- savings[-23,-46,-49]
```

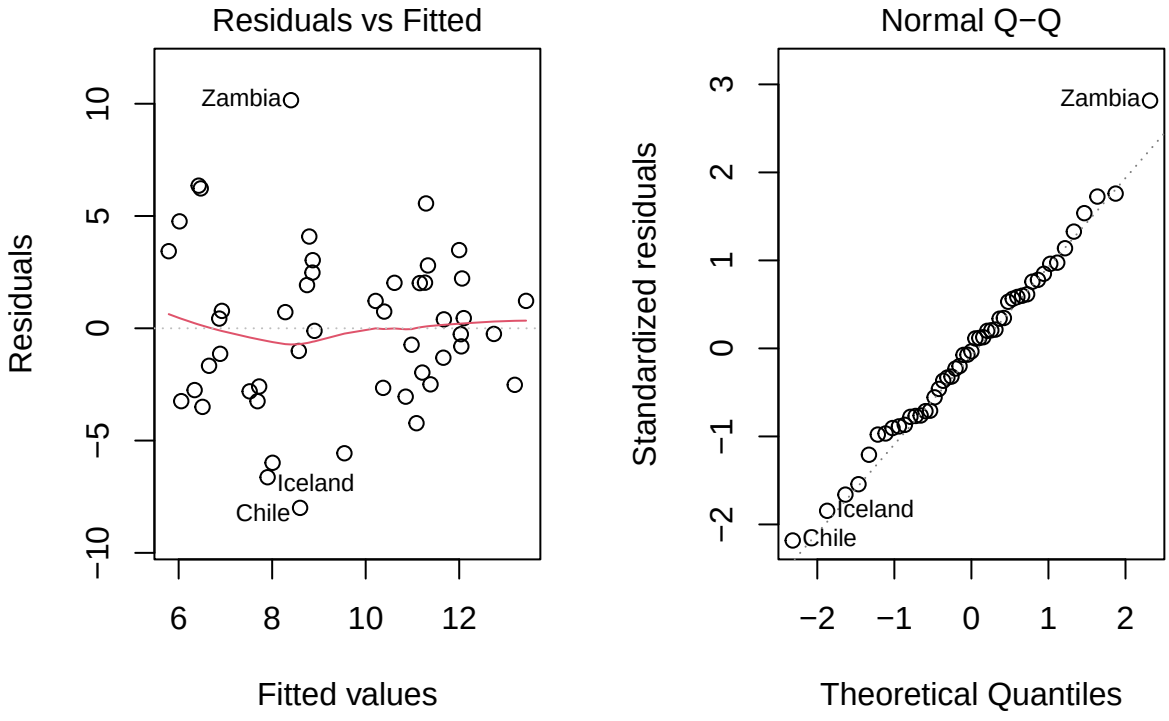
```
model_new <- lm(sr ~ ., data = savings_new)
summary(model_new)
```

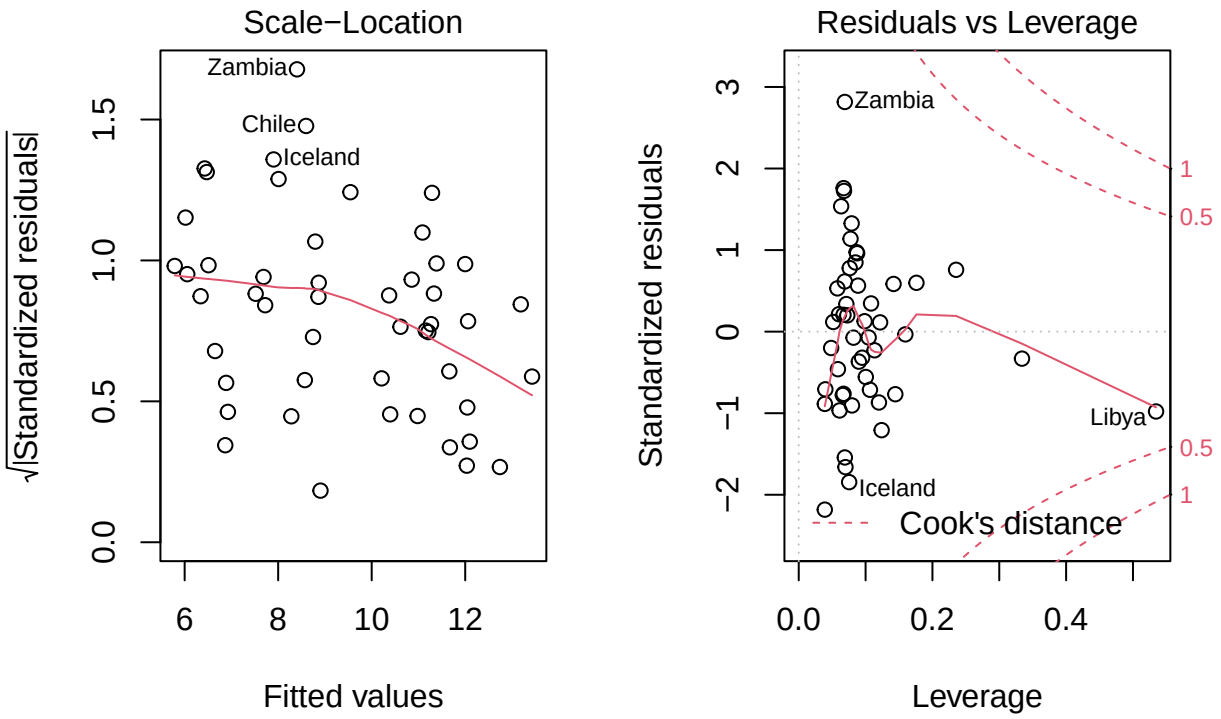
```
##
## Call:
## lm(formula = sr ~ ., data = savings_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.997 -2.592 -0.115  2.032 10.157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.9401714  7.7839968   3.076  0.00361 **
## pop15       -0.3679015  0.1536296  -2.395  0.02096 *
```

```
## pop75       -0.9736743  1.1554502  -0.843  0.40397
## dpi         -0.0004706  0.0009191  -0.512  0.61116
## ddpi         0.3347486  0.1984457   1.687  0.09871 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.738 on 44 degrees of freedom
## Multiple R-squared:  0.277,  Adjusted R-squared:  0.2113
## F-statistic: 4.214 on 4 and 44 DF,  p-value: 0.005649
```

(f) Use the plot() function again on the model and comment on the four graphs.

```
par(mfrow=c(1,2))
plot(model_new)
```





Plot 1: The Residual v/s Fitted plot is used to check the linearity and the constant variance assumption. Here, the mean of the residuals are centered at zero so the linearity assumption is valid. The spread of the residuals is same so the constant variance assumption is valid as well.

Plot 2: The Q-Q plot is used to check the normality of the errors. Here, the points closely follows a straight line that suggests that the data comes from a normal distribution.

Plot 3: The Scale Location plot shows if the residuals are spread equally along the range of predictors. It checks the assumption of constant variance. Here, the residuals appears to be spread randomly, however, the line is bending downwards here.

Plot 4: The Residual v/s Leverage plot helps to find the influential points in a data set. If any data point falls outside of the cook's distance the it is considered to be an influential point. all the observations are inside of the dashed line and this indicates that now there are no influential points.