PE I: Multivariable Regression

Outliers
(Chapter 4.9)

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

Multiple Regression: Model Assumptions

Much like in the case of the univariate regression with one independent variable, the multiple regression model has a number of required assumptions:

(MR.1): Linear Model The Data Generating Process (**DGP**), or in other words, the population, is described by a linear (*in terms of the coefficients*) model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \tag{MR.1}$$

(MR.2): Strict Exogeneity Conditional expectation of ε, given all observations of the explanatory variable matrix $\boldsymbol{X}$, is zero:

$$\mathbb{E}\left(\varepsilon|\boldsymbol{X}\right) = \mathbf{0} \tag{MR.2}$$

This assumption also implies that $\mathbb{E}(\varepsilon) = \mathbb{E}\left(\mathbb{E}(\varepsilon|\boldsymbol{X})\right) = \mathbf{0}$, $\mathbb{E}(\varepsilon\boldsymbol{X}) = \mathbf{0}$ and $\mathbb{C}\mathrm{ov}(\varepsilon, \boldsymbol{X}) = \mathbf{0}$. Furthermore, this property implies that: $\mathbb{E}\left(\boldsymbol{Y}|\boldsymbol{X}\right) = \boldsymbol{X}\beta$

(MR.3): Conditional Homoskedasticity The variance-covariance matrix of the error term, conditional on $\boldsymbol{X}$ is constant:

$$\mathbb{V}\text{ar}\left(\boldsymbol{\varepsilon}|\boldsymbol{X}\right) = \begin{bmatrix} \mathbb{V}\text{ar}(\epsilon_1) & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_1, \epsilon_N) \\ \mathbb{C}\text{ov}(\epsilon_2, \epsilon_1) & \mathbb{V}\text{ar}(\epsilon_2) & ... & \mathbb{C}\text{ov}(\epsilon_2, \epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}(\epsilon_N, \epsilon_1) & \mathbb{C}\text{ov}(\epsilon_N, \epsilon_2) & ... & \mathbb{V}\text{ar}(\epsilon_N) \end{bmatrix} = \sigma_\epsilon^2 \boldsymbol{I} \qquad \text{(MR.3)}$$

(MR.4): Conditionally Uncorrelated Errors The covariance between different error term pairs, conditional on $\boldsymbol{X}$, is zero:

$$\mathbb{C}\text{ov}\left(\epsilon_i, \epsilon_j|\boldsymbol{X}\right) = 0, \quad i \neq j \qquad \text{(MR.4)}$$

This assumption implies that all error pairs are uncorrelated. For cross-sectional data, this assumption implies that there is no spatial correlation between errors.

(MR.5) There exists no exact linear relationship between the explanatory variables. This means that:

$$c_1 X_{i1} + c_2 X_{i2} + ... + c_k X_{ik} = 0, \ \forall i = 1, ..., N \iff c_1 = c_2 = ... = c_k = 0 \quad \text{(MR.5)}$$

This assumption is violated if there exists some $c_j \neq 0$.
Alternatively, this requirement means that:

$$\text{rank}\,(\boldsymbol{X}) = k + 1$$

or, alternatively, that:

$$\det\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \neq 0$$

This assumption is important, because a linear relationship between independent variables means that we cannot separately estimate the effects of changes in each variable separately.

(MR.6) (optional) The residuals are normally distributed:

$$\varepsilon | \boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I}\right) \quad \text{(MR.6)}$$

Outliers

An **outlier** is an observation which is significantly different from other values in a random sample from a population.

If we collect all of the various problems that can arise - we can rank them in terms of severity:

$$outliers > non - linearity > heteroscedasticity > non - normality$$

Outlier Causes

Outliers can be cause by:

- measurement errors;
- being from a different process, compared to the rest of the data;
- not having a representative sample (e.g. measuring a single observation from a different city, when the remaining observations are all from one city);

Outlier Consequences

Outliers can lead to misleading results in parameter estimation and hypothesis testing. This means that a *single outlier* can make it seem like:

- a *non-linear* model may be better suited to the data sample, as opposed to a linear model;
- the residuals are *heteroskedastic*, when in fact only a residual has a larger variance, which is different from the rest;
- the distribution is skewed (i.e. *non-normal*), because of a single observation/residual, which is significantly different form the test.

```
set.seed(123)
#
N <- 100
x <- rnorm(mean = 8, sd = 2, n = N)
y <- 4 + 5 * x + rnorm(mean = 0, sd = 0.5, n = N)
y[N] <- -max(y)
```

Outlier Detection

The broad definition of outliers means that the decision whether an observation should be considered an outlier is left to the econometrician/statistician/data scientist.
Nevertheless, there are a number of different methods, which can be used to identify abnormal observations.
Specifically, for regression models, outliers are also detected by comparing the true and fitted values. Assume that our true model is the linear regression:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

Then, assume that we estimate $\widehat{\beta}$ via OLS. Consequently, we can write the fitted values as:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is called the **hat matrix** (or the **projection matrix**), which is the orthogonal projection that maps the vector of the response values, $\mathbf{Y}$, to the vector of fitted/predicted values, $\widehat{\mathbf{Y}}$. It describes the *influence* that each response value has on each fitted value, which is why $\mathbf{H}$ is sometimes also referred to as the **influence matrix**.

To understand the projection matrix a bit better do not treat the fitted values as something that is *separate* from the true values.

- Instead assume that you have two sets of values: $\mathbf{Y}$ and $\widehat{\mathbf{Y}}$.
- Ideally, we would want $\widehat{\mathbf{Y}} = \mathbf{Y}$.
- Assuming that the linear relationship, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, holds, this will generally not be possible because of the random shocks ε

However, the closest approximation would be the conditional expectation of $\mathbf{Y}$, given a design matrix $\mathbf{X}$, since we know that the conditional expectation is the best predictor from the proof in **Ch. 3.7**.

The Conditional Expectation is The Best Predictor (Ch. 3.7)

We begin by outlining the main properties of the conditional moments, which will be useful (assume that X and Y are random variables):

- *Law of total expectation*: $\mathbb{E}[\mathbb{E}(h(Y)|X)] = \mathbb{E}[h(Y)]$;
- *Conditional variance*: $\mathbb{V}\mathrm{ar}(Y|X) := \mathbb{E}\left((Y - \mathbb{E}[Y|X])^2|X\right) = \mathbb{E}(Y^2|X) - (\mathbb{E}[Y|X])^2$;
- *Variance of conditional expectation*: $\mathbb{V}\mathrm{ar}(\mathbb{E}[Y|X]) = \mathbb{E}\left[(\mathbb{E}[Y|X])^2\right] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 = \mathbb{E}\left[(\mathbb{E}[Y|X])^2\right] - (\mathbb{E}[Y])^2$;
- *Expectation of conditional variance*: $\mathbb{E}[\mathbb{V}\mathrm{ar}(Y|X)] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2\right] = \mathbb{E}\left[\mathbb{E}\left[Y^2|X\right]\right] - \mathbb{E}\left[(\mathbb{E}[Y|X])^2\right] = \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[(\mathbb{E}[Y|X])^2\right]$;
- Adding the third and fourth properties together gives us: $\mathbb{V}\mathrm{ar}(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 = \mathbb{V}\mathrm{ar}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathbb{V}\mathrm{ar}(Y|X)]$.

For simplicity, assume that we are interested in the prediction of $\mathbf{Y}$ via the conditional expectation:

$$\mathbf{Y} = \mathbb{E}(\mathbf{Y}|\mathbf{X})$$

We will show that, in general, the conditional expectation is the *best* predictor of $\mathbf{Y}$.

Assume that the best predictor of Y (a single value), given $\mathbf{X}$ is some function $g(\cdot)$, which minimizes the expected squared error:

$$\text{argmin}_{g(\mathbf{X})}\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right].$$

Using the conditional moment properties, we can rewrite $\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ as:

$$\begin{aligned}
&\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] = \mathbb{E}\left[(Y + \mathbb{E}[Y|\mathbf{X}] - \mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X}))^2\right] \\
&\quad = \mathbb{E}\left[(Y - \mathbb{E}[Y|\mathbf{X}])^2 + 2(Y - \mathbb{E}[Y|\mathbf{X}])(\mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X})) + (\mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X}))^2\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left((Y - \mathbb{E}[Y|\mathbf{X}])^2|\mathbf{X}\right)\right] + \mathbb{E}\left[2(\mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X}))\mathbb{E}\left[Y - \mathbb{E}[Y|\mathbf{X}]|\mathbf{X}\right] + \mathbb{E}\left[(\mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X}))^2|\mathbf{X}\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{V}\text{ar}(Y|X)\right] + \mathbb{E}\left[(\mathbb{E}[Y|\mathbf{X}] - g(\mathbf{X}))^2\right].
\end{aligned}$$

Taking $g(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$ minimizes the above equality to the expectation of the conditional variance of Y given $\mathbf{X}$:

$$\mathbb{E}\left[(Y - \mathbb{E}[Y|\mathbf{X}])^2\right] = \mathbb{E}\left[\mathbb{V}\text{ar}(Y|X)\right].$$

Thus, $g(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$ is the best predictor of Y.

Going back to out **projection matrix** ...

Using the OLS definition of $\widehat{\beta}$, the best predictor (i.e. the conditional expectation) maps the values of $\mathbf{Y}$ to the values of $\widehat{\mathbf{Y}}$ via the projection matrix $\mathbf{H}$.

The projection matrix can be utilized when calculating leverage scores and Cook's distance, which are used to identify *influential* observations.

Leverage Score of Observations

Leverage measures how far away an observation of a predictor variable, $\mathbf{X}$, is from the mean of the predictor variable.

For the linear regression model, the *leverage score* for the i-th observation is defined as the i-th diagonal element of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$, which is equivalent to taking a partial derivative of $\widehat{Y}_i$ with respect to Y_i:

$$h_{ii} = \frac{\partial \widehat{Y}_i}{\partial Y_i} = (\mathbf{H})_{ii}$$

Defining the leverage score via the partial derivative allows us to interpret the leverage score as the observation self-influence, which describes how the *actual value*, Y_i, influences the *fitted value*, $\widehat{Y}_i$.

The leverage score h_{ii} is bounded:

$$0 \leq h_{ii} \leq 1$$

Proof.

Noting that $\mathbf{H}$ is symmetric and the fact that it is an idempotent matrix:

$$\mathbf{H}^2 = \mathbf{HH} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top = \mathbf{XI}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top = \mathbf{H}$$

we can examine the diagonal elements of the equality $\mathbf{H}^2 = \mathbf{H}$ to get the following bounds of H_{ii}:

$$h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2 \geq 0$$

$$h_{ii} \geq h_{ii}^2 \Longrightarrow h_{ii} \leq 1$$

□

We can also relate the residuals to the leverage score:

$$\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$$

Examining the variance-covariance matrix of the regression errors we see that:

$$\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\varepsilon}}) = \mathbb{V}\mathrm{ar}((\mathbf{I} - \mathbf{H})\,\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\,\mathbb{V}\mathrm{ar}(\mathbf{Y})\,(\mathbf{I} - \mathbf{H})^{\top} = \sigma^2\,(\mathbf{I} - \mathbf{H})\,(\mathbf{I} - \mathbf{H})^{\top} = \sigma^2\,(\mathbf{I} - \mathbf{H})\,,$$

where we have used the fact that $(\mathbf{I} - \mathbf{H})$ is idempotent and $\mathbb{V}\mathrm{ar}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Since the diagonal elements of the variance-covariance matrix are the variances of each observation, we have that $\mathbb{V}\mathrm{ar}(\widehat{\epsilon}_i) = (1 - h_{ii})\sigma^2$.

Thus, we can see that a leverage score of $h_{ii} \approx 0$ would indicate that the i-th observation has no influence on the error variance, which would mean that its variance close to the true (unobserved) variance σ^2.

Observations with leverage score values larger than $2(k+1)/N$ are considered to be potentially highly influential.

Assume that we estimate the model via OLS:

```
mdl_1_fit <- lm(y ~ 1 + x)
```

Studentized Residuals

The **studentized residuals** are related to the **standardized residuals**, as they are defined as:

$$t_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}}$$

The main distinction comes from the calculation of $\widehat{\sigma}$, which can be calculated in two ways:

- **Standardized residuals** calculate the **internally studentized** residual variance estimate:

$$\widehat{\sigma}^2 = \frac{1}{N - (k+1)} \sum_{j=1}^{N} \widehat{\epsilon}_j^2$$

- If we suspect that the i-th residual of being *improbably large* (i.e. it cannot be from the same normal distribution as the remaining of the residuals) - we exclude it from variance estimation by calculating the **externally studentized** residual variance estimate:

$$\widehat{\sigma}^2_{(i)} = \frac{1}{N - (k+1) - 1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \widehat{\epsilon}_j^2$$

If the residuals are independent and $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, then the distribution of the studentized residuals depends on the calculation of the variance estimate:

- If the residuals are **internally studentized** - they have a *tau distribution*:

$$t_i \sim \frac{\sqrt{v} \cdot t_{v-1}}{\sqrt{t_{v-1}^2 + v - 1}}, \text{ where } v = N - (k+1)$$

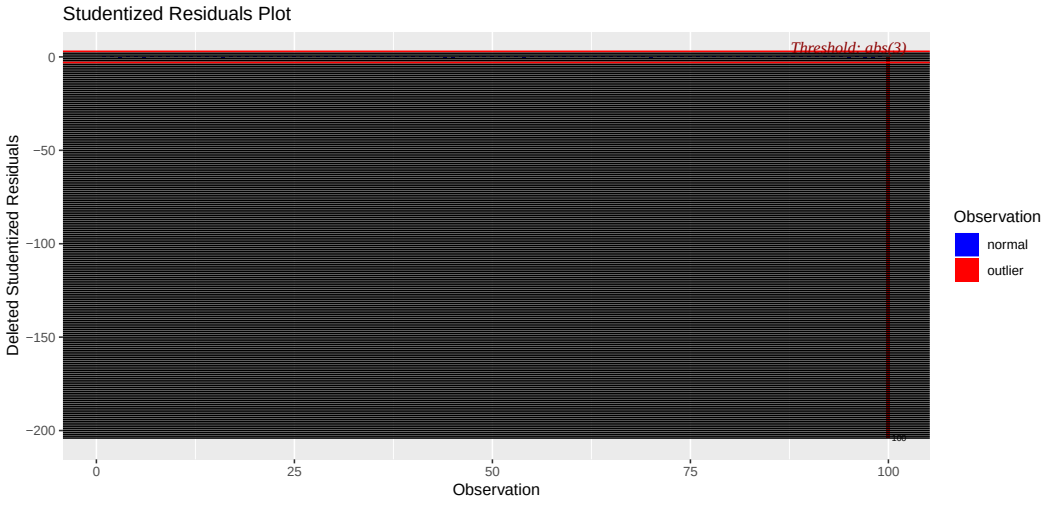
- If the residuals are **externally studentized** - they have a *Student's t-distribution* (we will also refer to them as $t_{i(i)}$):

$$t_i = t_{i(i)} \sim t_{(N-(k-1)-1)}$$

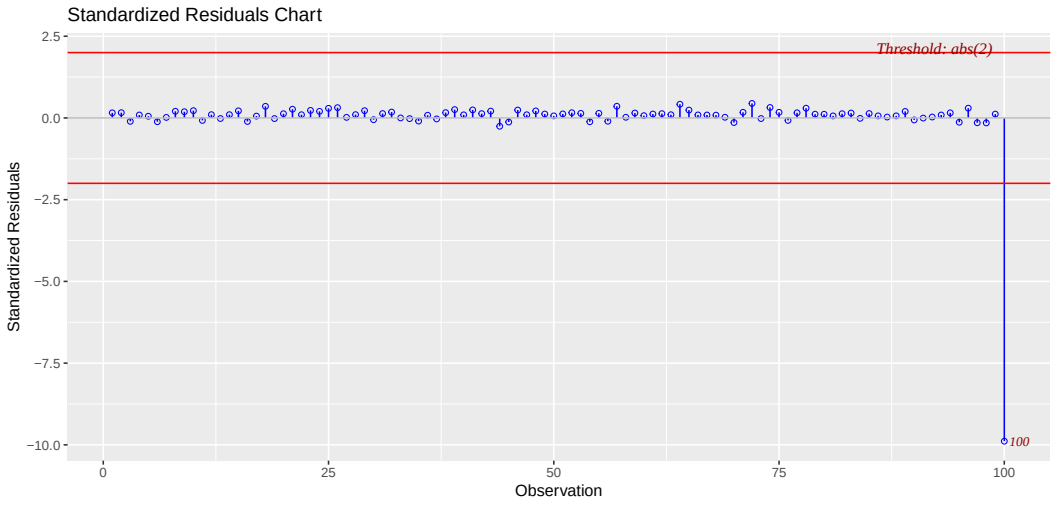
Observations with studentized residual values larger than 3 in *absolute* value could be considered outliers.

We can plot the studentized and standardized residuals:

```
olsrr::ols_plot_resid_stud(mdl_1_fit)
```

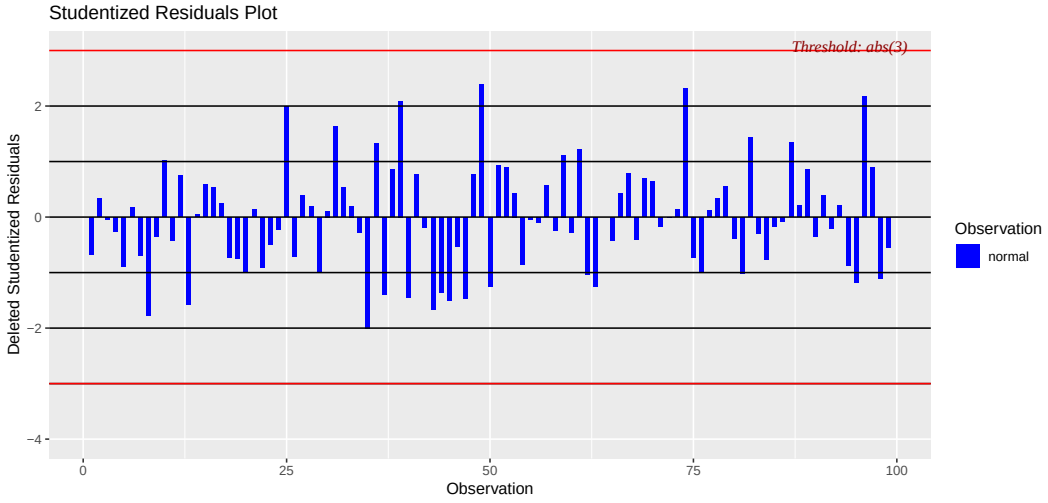


```
olsrr::ols_plot_resid_stand(mdl_1_fit)
```

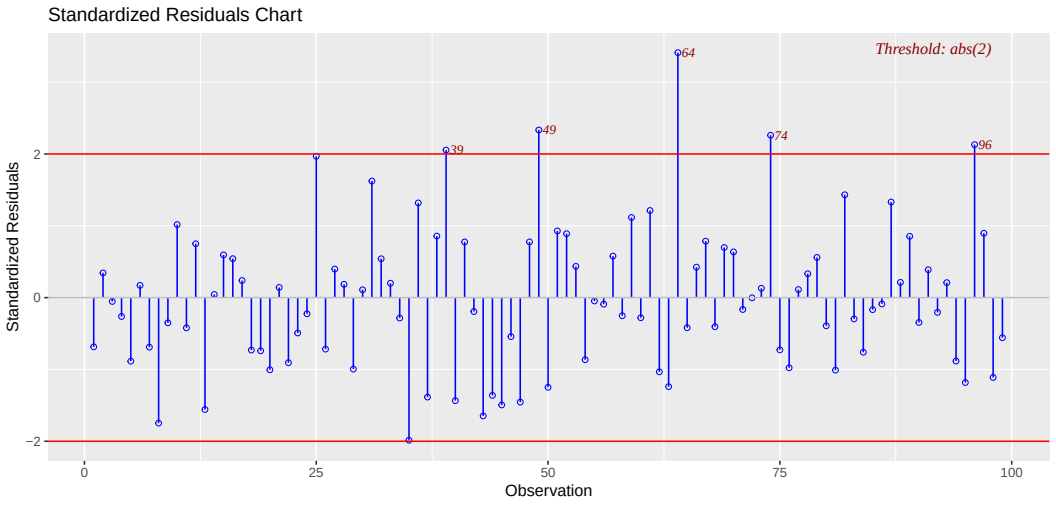


We can examine the same plots on the model, with the outlier observation removed from the data:

```
olsrr::ols_plot_resid_stud(lm(y[-N] ~ 1 + x[-N]))
```



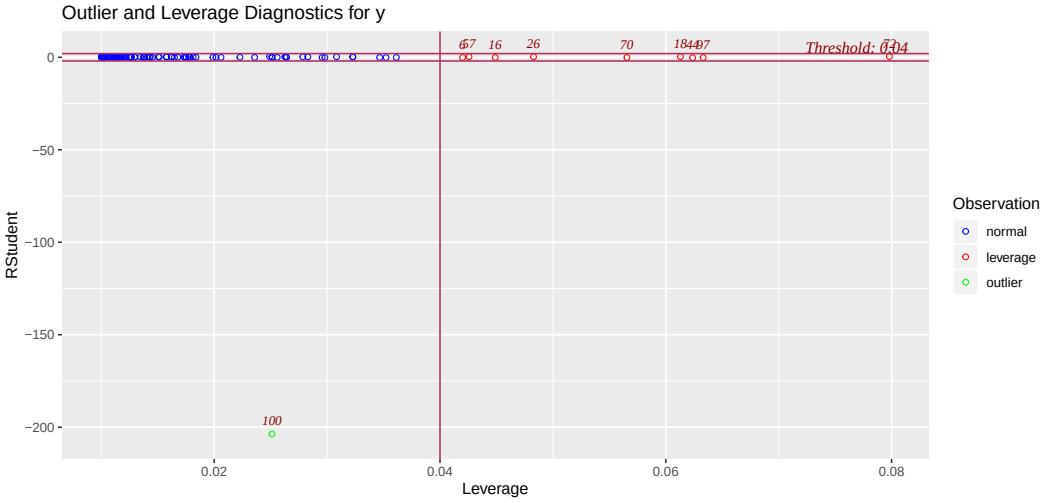
```
olsrr::ols_plot_resid_stand(lm(y[-N] ~ 1 + x[-N]))
```



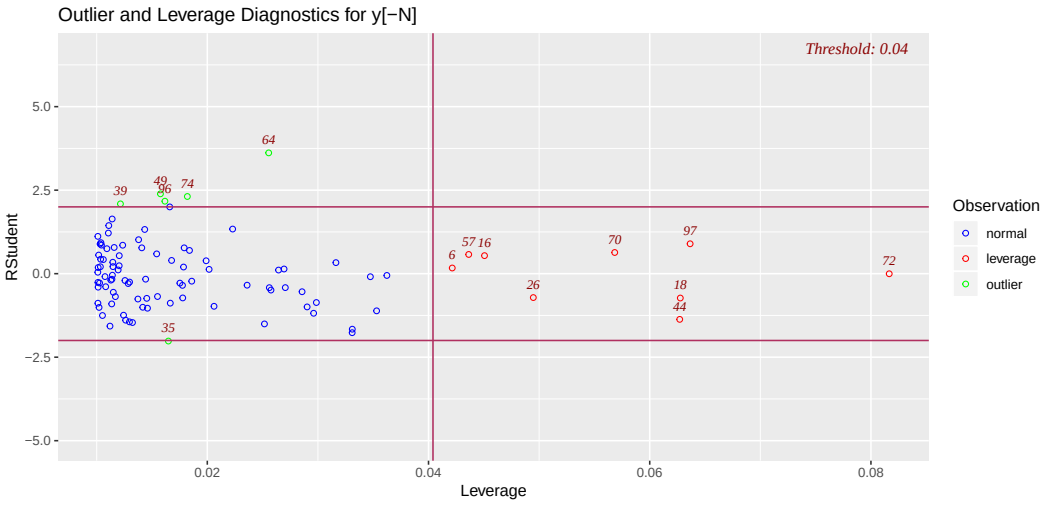
While the studentized residuals appear to have no outliers, the *standardized* residuals indicate that a few observations may be influential. Since we have simulated the data, we know that our data contained only one outlier. Consequently, we should not treat all observations outside the threshold as definite outliers.

We may also be interested in plotting the studentized residuals against the leverage points:

```
olsrr::ols_plot_resid_lev(mdl_1_fit)
```



```
olsrr::ols_plot_resid_lev(lm(y[-N] ~ 1 + x[-N]))
```



This plot combined the leverage score, which shows influential **explanatory variable** observations, and the studentized residual plot, which shows outlier residuals of the difference between the actual and fitted **dependent variables**.

Influential observations

Influential observations are defined as observations, which have a large effect on the results of a regression.

DFBETAS

The $DFBETA_i$ **vector** measures how much an observation i has effected the estimate of a regression coefficient **vector** . It measures the difference between the regression coefficients, calculated for all of the data, and the regression coefficients, calculated with the observation i deleted:

$$DFBETA_i = \frac{\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}}{\sqrt{\widehat{\sigma}^2_{(i)} \text{diag}\left((\mathbf{X}^\top \mathbf{X})^{-1}\right)}}$$

Observations with a $DFBETA$ value larger than $2/\sqrt{N}$ in absolute value should be carefully inspected.
The recommended general cutoff (absolute) value is 2.

We can calculate the appropriate *DFBETAS* for the last 5 observations as follows:

```
dfbetas_manual <- NULL
for(i in (N-4):N){
  mdl_2_fit  <- lm(y[-i] ~ 1 + x[-i])
  numerator  <- mdl_1_fit$coef - mdl_2_fit$coef
  denominator<- sqrt((summary(mdl_2_fit)$sigma^2) * diag(solve(t(cbind(1, x)) %*% cbind(1, x))))
  dfbetas_manual <- rbind(dfbetas_manual, numerator / denominator)
}
print(dfbetas_manual)
```

```
##        (Intercept)           x
## [1,]   0.028743821 -0.022789554
## [2,]   0.030744687 -0.034844559
## [3,]   0.020403791 -0.024298429
## [4,]   0.006702931 -0.004242548
## [5,] -29.230784828 25.362876769
```

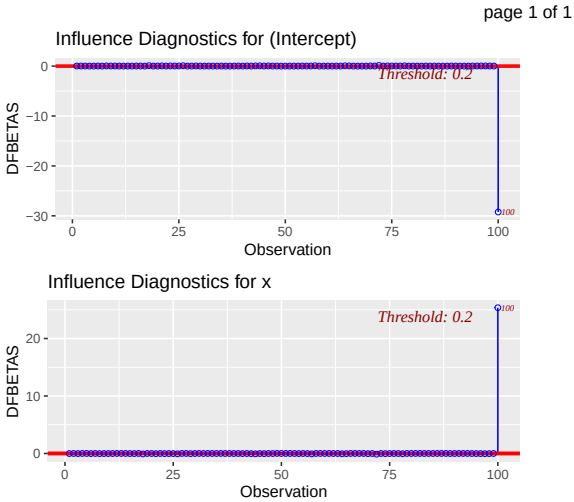
While these calculations are a bit more involved, we can use the built-in functions as well:

```
print(tail(dfbetas(mdl_1_fit), 5))
```

```
##       (Intercept)           x
## 96    0.028743821 -0.022789554
## 97    0.030744687 -0.034844559
## 98    0.020403791 -0.024298429
## 99    0.006702931 -0.004242548
## 100 -29.230784828 25.362876769
```

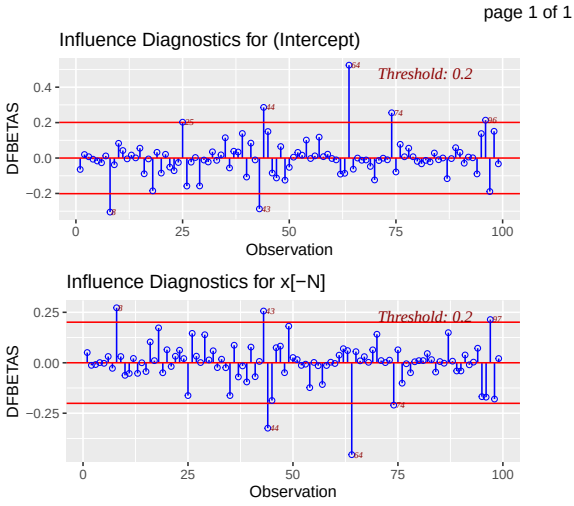
If we wanted, we could also plot these values:

```
olsrr::ols_plot_dfbetas(mdl_1_fit)
```



If we were to remove the last observation and examine the *DFBETAS* plot:

```
olsrr::ols_plot_dfbetas(lm(y[-N] ~ 1 + x[-N]))
```



We see that there are some observations, which may be worth examining. In this case, we know that there are no more outliers because we have simulated the data ourselves. So this is a good example that you should not blindly trust the above charts, as the **influential observations are not necessarily outliers**.

DFFITS

DFFITS measures how much an observation i has effected the fitted value of a regression. It is defined as a Studentized difference between the fitted values from a regression, estimated on all of the data, and the fitted values from a regression, estimated on the data with observation i deleted:

$$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\widehat{\sigma}^2_{(i)}\sqrt{h_{ii}}} = t_{i(i)}\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

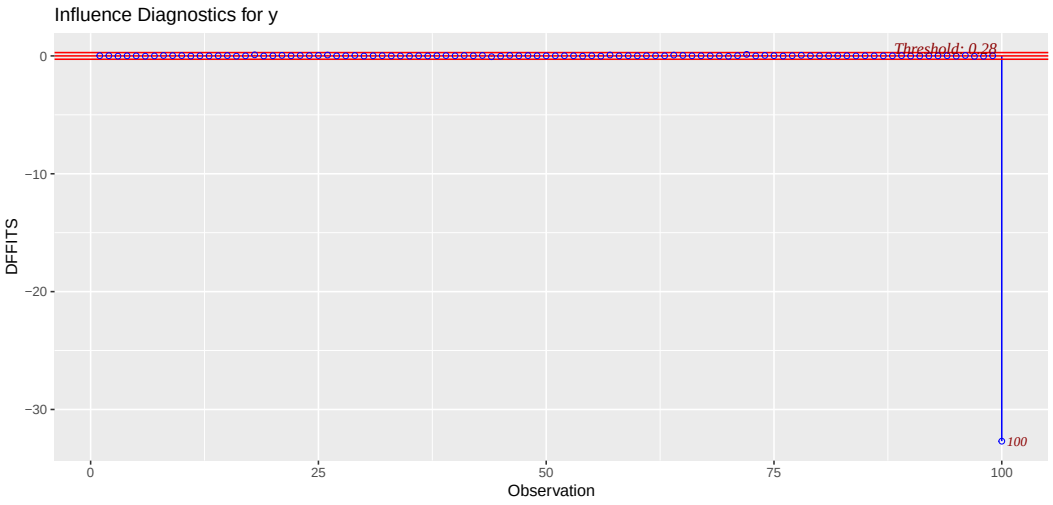
where $t_{i(i)}$ is the **externally** studentized residual.

```
tmp_val <- dffits(mdl_1_fit)
print(format(tail(cbind(tmp_val), 10), scientific = FALSE))
```

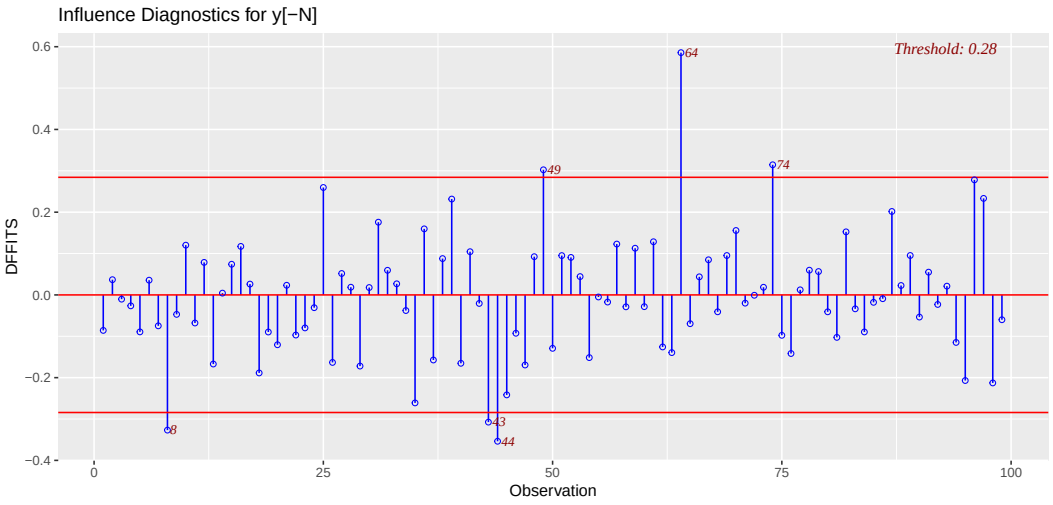
```
##      tmp_val
## 91  " -0.0005235787"
## 92  "  0.0031760359"
## 93  "  0.0091761236"
## 94  "  0.0199891169"
## 95  " -0.0226748095"
## 96  "  0.0376495768"
## 97  " -0.0379725909"
## 98  " -0.0287153104"
## 99  "  0.0125545090"
## 100 "-32.6910440413"
```

Observations with a *DFFITS* value larger than $2\sqrt{(k+1)/N}$ in absolute value

```
olsrr::ols_plot_dffits(mdl_1_fit)
```



```
olsrr::ols_plot_dffits(lm(y[-N] ~ 1 + x[-N]))
```



Similarly to what we have observed with *DFBETAS* - we should not blindly trust that each value outside the cutoff region is an outlier. Instead, we should treat them as influential observations, which need additional analysis to determine whether they are acceptable.

Cook's distance

Cook's D measures the aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. It can be used to:

- indicate influential data points (i.e. potential outliers);
- indicate regions, where more observations would be needed;

Cook's distance for observation i is defined as:

$$D_i = \frac{\sum_{j=1}^{N}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{(k+1)\widehat{\sigma}^2} = \frac{\widehat{\epsilon}_i^2}{(k+1)\widehat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

where:

- $\widehat{Y}_{j(i)}$ is the fitted value of Y_j, obtained by excluding the i-th observation and re-estimating the same model via OLS.
- $\widehat{\sigma}^2 = \dfrac{\widehat{\varepsilon}^\top \widehat{\varepsilon}}{N-(k+1)}$ is the mean squared error of the error term.

Note: in practical terms, it may be easier to use the leverage score expression of D_i instead of re-estimating the model for each observation case.

```r
tmp_val <- cooks.distance(mdl_1_fit)
print(format(tail(cbind(tmp_val), 10), scientific = FALSE))
```

```
##     tmp_val
## 91  "0.0000001384804"
## 92  "0.0000050955563"
## 93  "0.0000425310902"
## 94  "0.0002017917033"
## 95  "0.0002596785491"
## 96  "0.0007154000322"
## 97  "0.0007282312180"
## 98  "0.0004164378945"
## 99  "0.0000796089714"
## 100 "1.2596831219103"
```

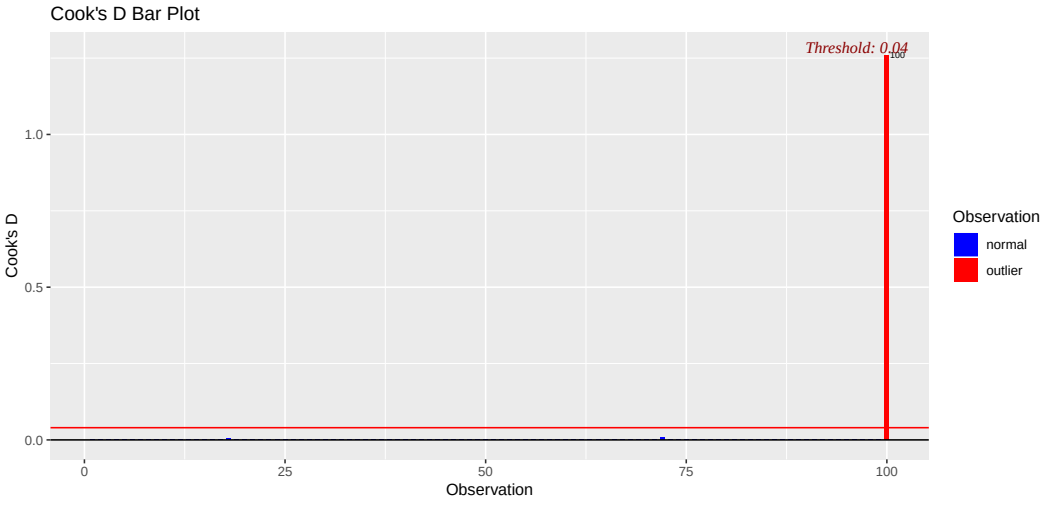
Cook's distance values, which are:

- larger than $4/N$ (the traditional cut-off);
- larger than $3 \times \frac{1}{N} \sum_{i=1}^{N} D_i$

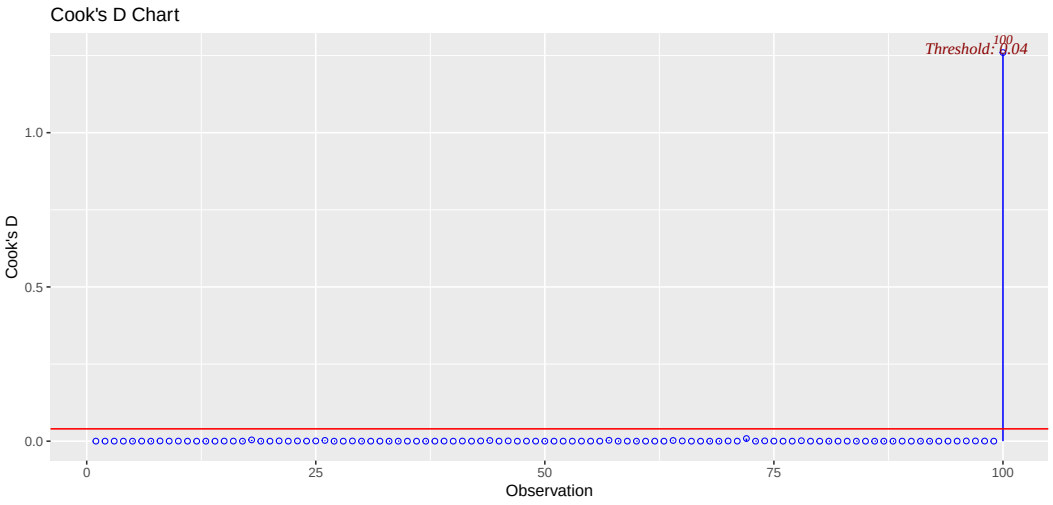
could be considered highly influential.

We can plot the D_i points:

```
olsrr::ols_plot_cooksd_bar(mdl_1_fit)
```

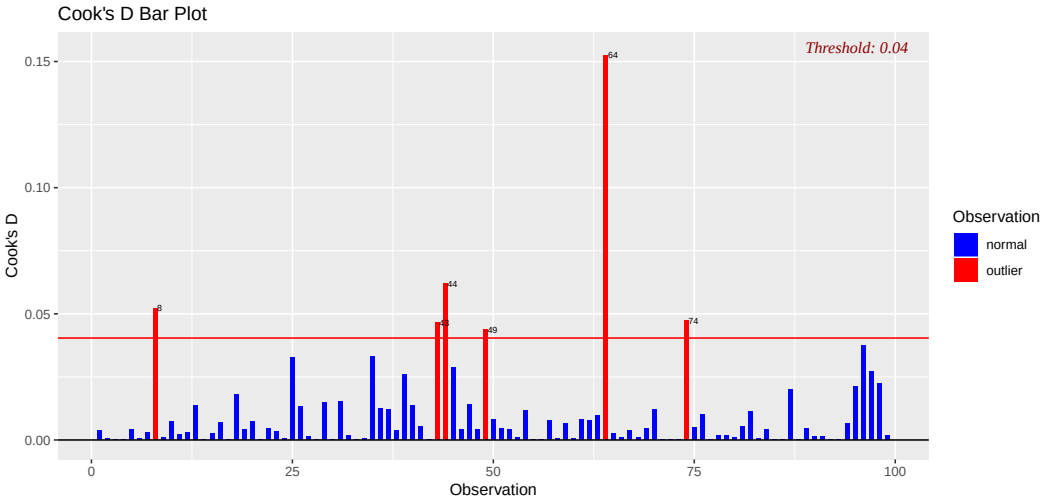


```
olsrr::ols_plot_cooksd_chart(mdl_1_fit)
```

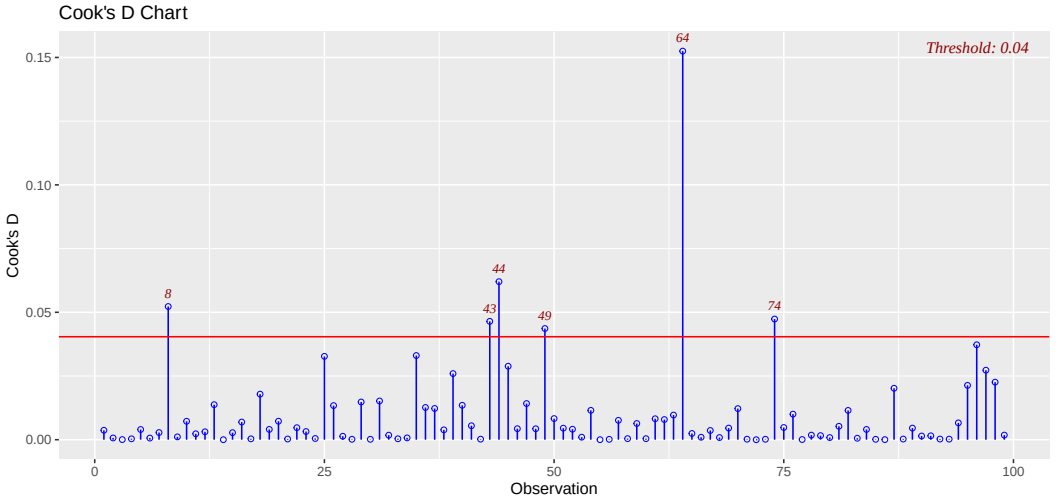


As well as plot the D_i on the data without the outlier observation:

```
olsrr::ols_plot_cooksd_bar(lm(y[-N] ~ 1 + x[-N]))
```



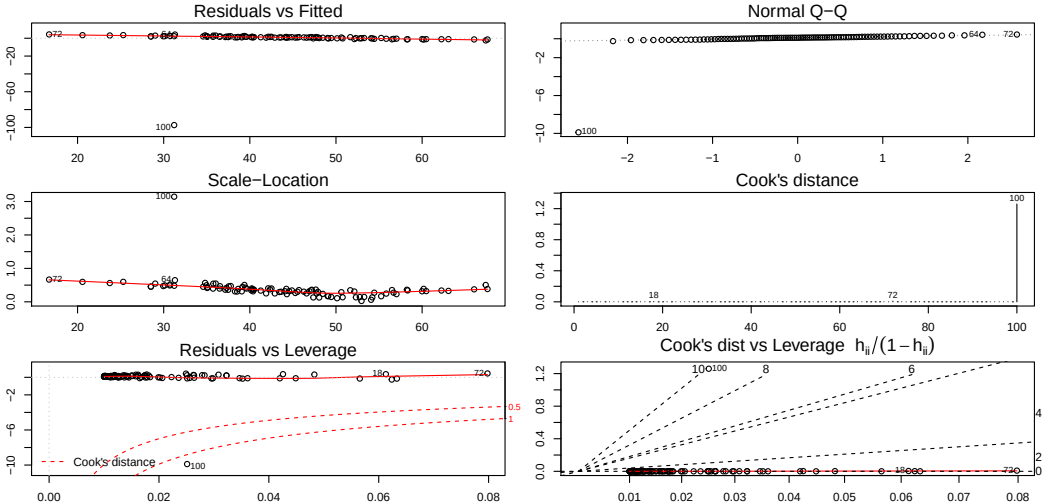
```
olsrr::ols_plot_cooksd_chart(lm(y[-N] ~ 1 + x[-N]))
```



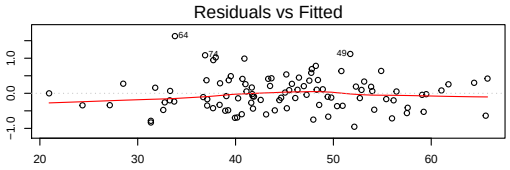
We again see a similar result, as with *DFBETAS* and *DFFITS*.

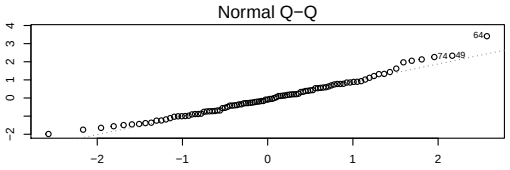
Also note that R has a lot of different plots for the default `lm` model output:

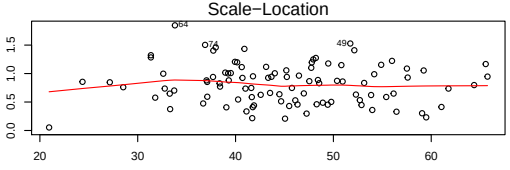
```
par(mfrow = c(3, 2), mar = c(2, 2, 2, 2))
for(i in 1:6){
  plot(mdl_1_fit, which = i)
}
```

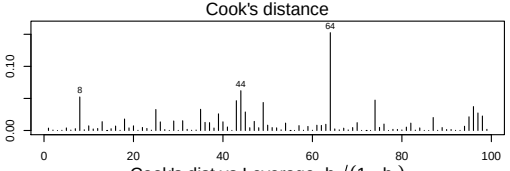


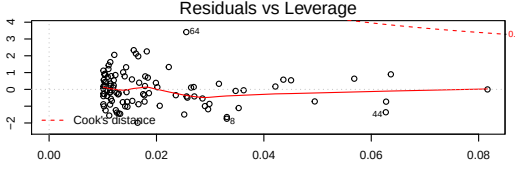
```r
par(mfrow = c(3, 2), mar = c(2, 2, 2, 2))
for(i in 1:6){
  plot(lm(y[-N] ~ 1 + x[-N]), which = i)
}
```

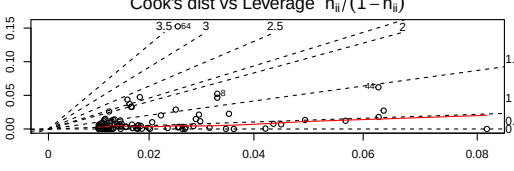












Addressing Outliers

After determining that a specific observation is indeed an outlier, we want to address them in some way.

Capping the Outliers

If we find that the explanatory variables $X_{1,i}, ..., X_{k,i}$ of an outlier variable Y_i are similar to other observations, with non-outlier values of Y_i, we may cap the value of the outlier, to match the values.

Replacing Outliers with Imputed Values

If we are certain that the outlier is due to some error in the data itself - we could try to impute the observations by treating them as missing values and substituting them for some average value of Y.
The Expectation-maximization (EM) algorithm could be utilized for missing data imputation.

Deleting Outliers

In some cases, if we are absolutely sure that the observation is an outlier, which is either completely unlikely, or impossible to encounter again, we could drop it.

Robust Regression

In addition to the methods mentioned before, we could also run a Robust regression.

In our example, we know that the last observation was differently generated, and is thus an outlier, which we can delete.

We can compare how would our model look like with the whole dataset, and if we were to drop the outlier observation:

```
plot(x, y)
lines(x, mdl_1_fit$fitted.values, col = "red")
lines(x[-N], lm(y[-N] ~ 1 + x[-N])$fitted.values, col = "blue")
points(x[N], y[N], pch = 19, col = "red")
legend("topleft", lty = 1, col = c("red", "blue"), legend = c("with outlier", "deleted outlier"))
```