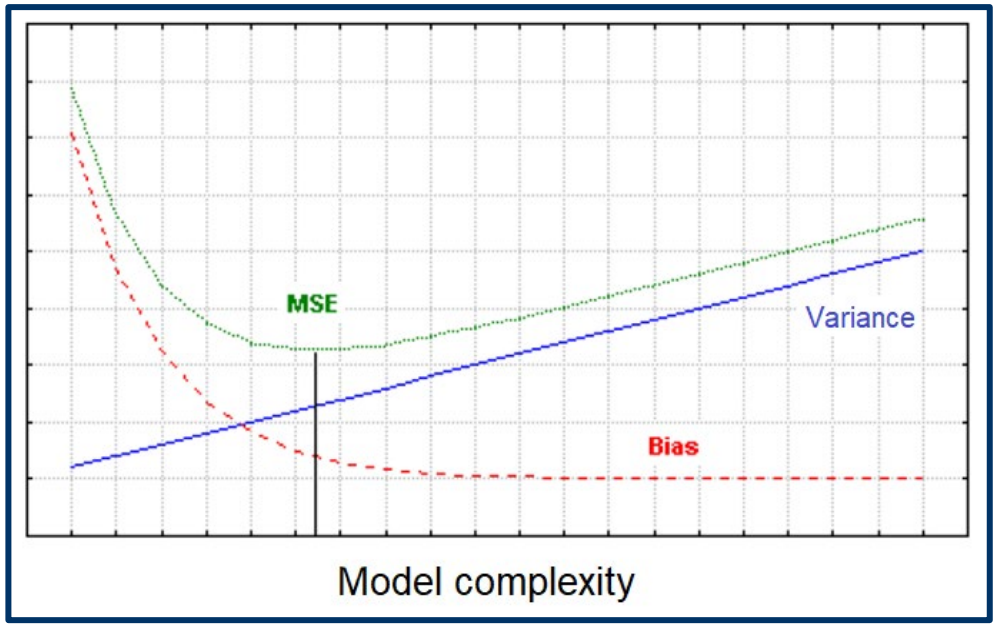
## Bias of a model

As already explained, the estimator **b** of a parameter $\beta$ is unbiased if $E(\mathbf{b}) = \beta$.

If the estimator is biased, its bias can be defined as $B(\mathbf{b}) = \beta - E(\mathbf{b})$, thus it corresponds to the systematic error related to the estimator **b**.

The quality of an estimator **b** is related to the minimum value of its Mean Square Error (MSE), which, in turn, can be described as the combination of two effects:




As a general rule, variance is increased with the model complexity (*i.e.*, with the number of variables), whereas bias is decreased.

The goal of biased methods is searching for a compromise between the model complexity and its variability, in order to minimize MSE.

The MSE for a biased estimator is defined as follows:

$$\mathbf{MSE}(\mathbf{b}) = E\,(\mathbf{b}-\beta)^2 = E\left[\left(\mathbf{b}-E(\mathbf{b})\right)-\left(\beta-E(\mathbf{b})\right)\right)^2\right]$$

The expectation for the square of the binomial shown in the second member of the equation can be expressed as the sum of the expectations of the squared terms of the binomial:

$$E\left[\left(\mathbf{b}-E(\mathbf{b})\right)^2\right]+E\left[\left(\beta-E(\mathbf{b})\right)^2\right]+E\left[-2\left(\mathbf{b}-E(\mathbf{b})\right)\left(\beta-E(\mathbf{b})\right]\right.$$

However, if the double product is made explicit, its expectation is easily found to be equal to 0:

$$E\left[-2\,\mathbf{b}\,\beta+2\,\mathbf{b}\,E(\mathbf{b})+2\beta E(\mathbf{b})-2E(\mathbf{b})E(\mathbf{b})\right]=$$

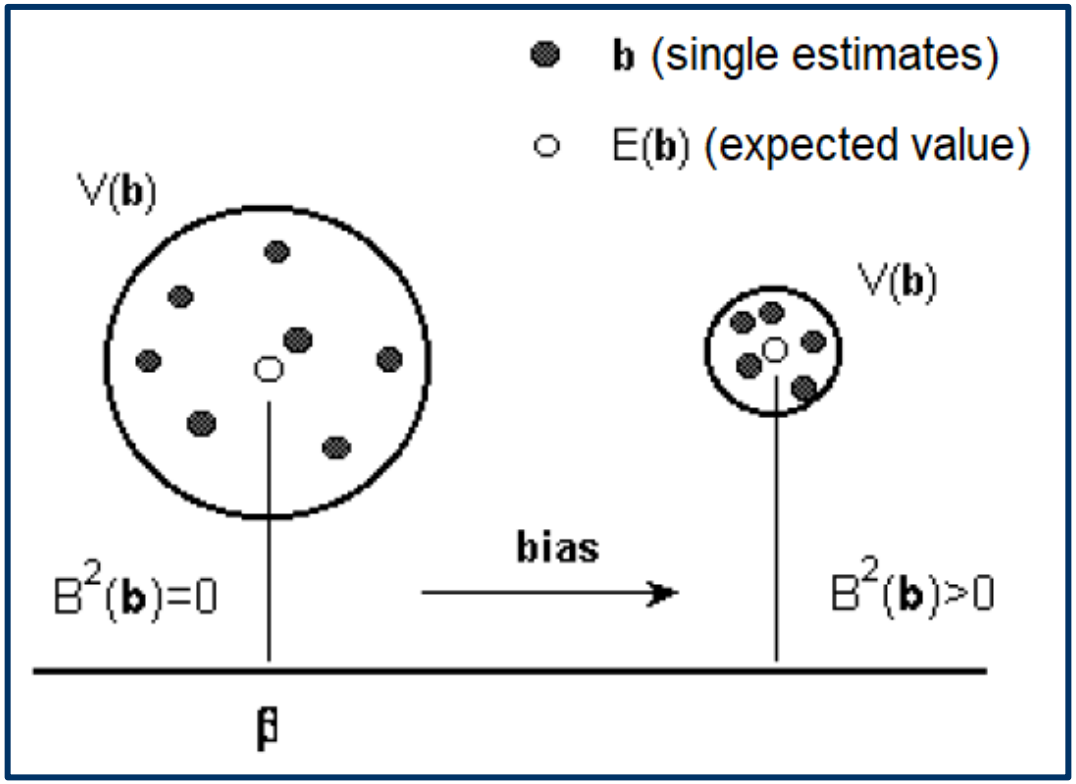
$$=-2\beta E(\mathbf{b})+2E(\mathbf{b})E(\mathbf{b})+2\beta E(\mathbf{b})-2E(\mathbf{b})E(\mathbf{b})=0$$

Consequently:

$$\mathbf{MSE}(\mathbf{b}) = E\left[\left(\mathbf{b}-E(\mathbf{b})\right)^2\right]+E\left[\left(\beta-E(\mathbf{b})\right)^2\right]$$

$$=E\left[\left(\mathbf{b}-E(\mathbf{b})\right)^2\right]+\left(\beta-E(\mathbf{b})\right)^2=V(\mathbf{b})+B^2(\mathbf{b})$$

The concept of bias can be represented graphically using the following figure:




On the left side of the figure a model characterized by a higher variance (represented by the circle area) but unbiased ($B^2(\mathbf{b}) = 0$) is represented.

On the right side, a model with a lower variance but characterized by a slight bias ($B^2(\mathbf{b}) > 0$) is shown.

# Validation of a model

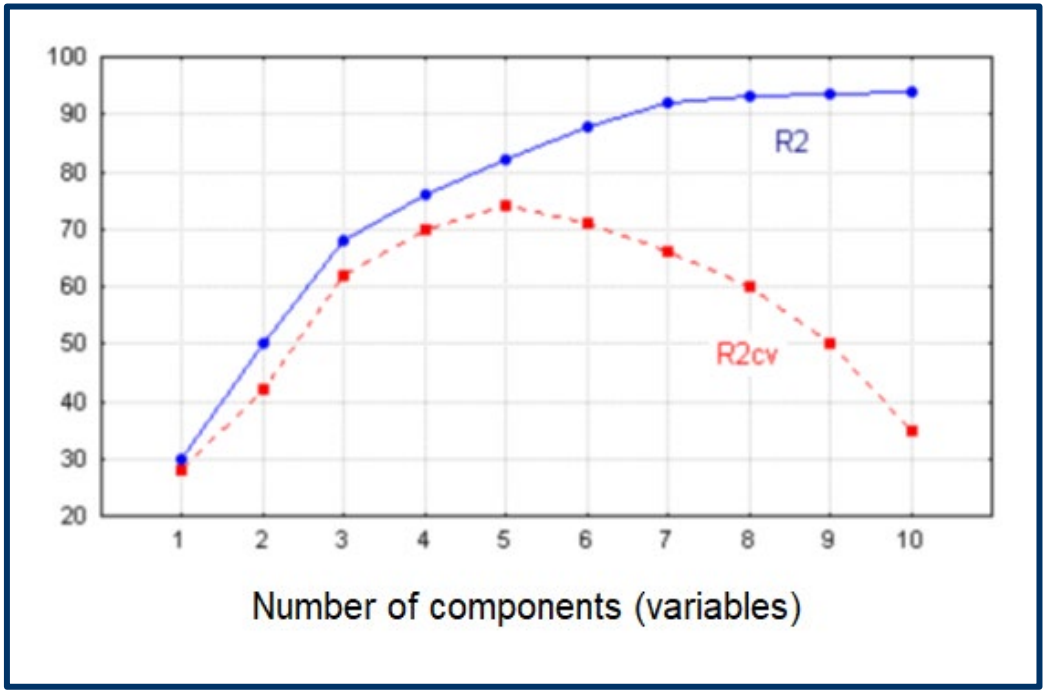
Validation of a model consists in searching the structure of model that makes its predictive capacity maximum.

At the same time the model must have stability characteristics that make it sufficiently independent by specific data exploited to build it.

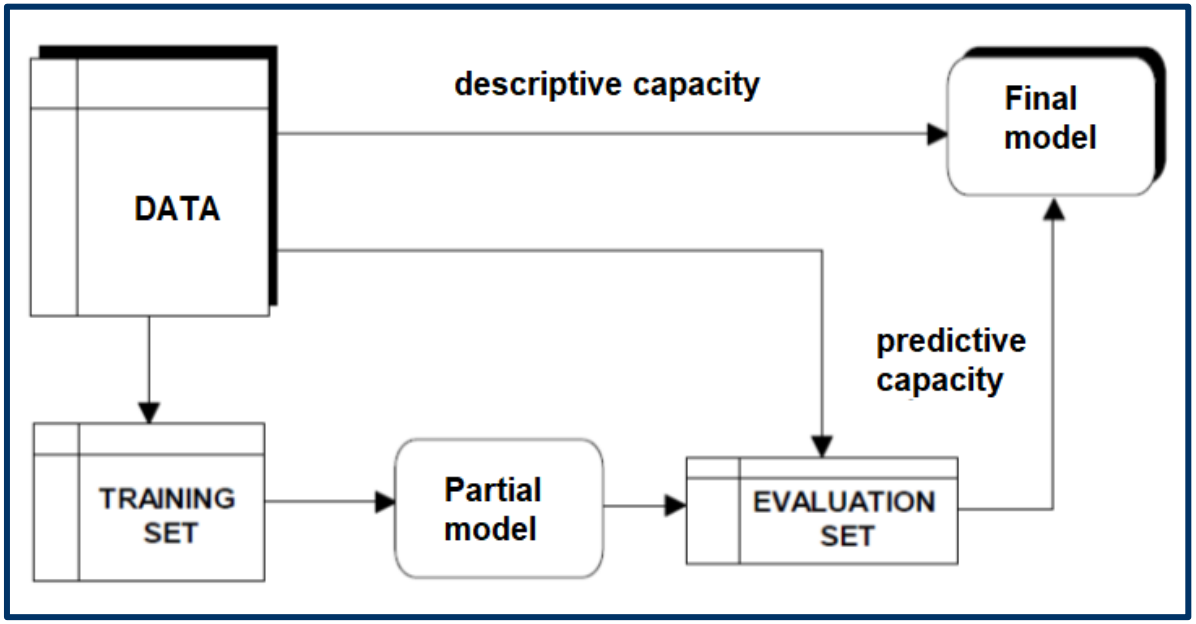
An increase in the model complexity (number of variables) leads to an increase in the descriptive quality of the model (fitting); however, an uncontrolled increase in the model complexity makes its predictive performance worse (overfitting).

As shown in the figure on the right, an increase in the number of components considered as significant in a regression model leads to an increase in the percentage of variance explained by fitting (R2).

On the other hand, when the number of components is greater than 5 variance explained in prediction (R2cv) is decreased.



Validation of a model usually occurs according to the following scheme:




As shown in the figure, part of available data is usually employed to build the training set, whereas the remaining part is used to build the evaluation set.

The training set is used to build a partial model, that is subsequently applied to predict the response of objects (samples) belonging to the evaluation set.

The ensemble of predictions made using data included in the evaluation set is exploited to calculate one or more parameters that enable the evaluation of the predictive capacity of the final model, that is usually build using all available data.

## Validation techniques

The four most common techniques adopted for model Cross Validation (CV) are:

1) leave-one-out
2) leave-more-out (also called k-fold CV)
3) training/evaluation splitting
4) bootstrap

### Leave-one-out

Given n objects (observations), the calculation of n models is made, with each model obtained using the remaining n-1 objects.
The square of the difference between experimental and predicted response is obtained for all the n objects that, in turn, are excluded from the model. Each of them corresponds to a MSE:

$$MSE_i = (y_i - \hat{y}_{\bar{i}})^2$$

Finally, the average of available MSE values is calculated:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

## Leave-more-out or k-fold CV

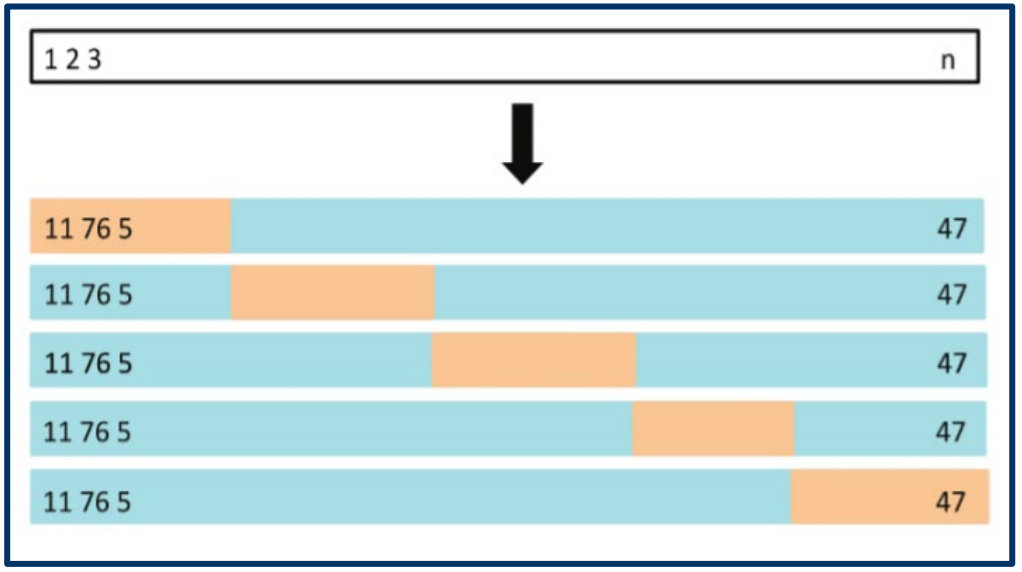
Since the Leave-one-out approach can lead to too optimistic predictive values, especially when the number of objects is high, a different approach, known as Leave-more-out, can be adopted:

In this case data are randomly divided into G cancellation groups, so that k objects, with k = n/G, form a single group and represent the evaluation set each time. The training set, consisting in n-k objects, is used to predict the response of the k excluded objects.

k estimates of $MSE_i$ are obtained using the Leave-more-out approach, thus a k-fold CV can be calculated as follows:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

The Leave-more-out approach is less intensive, from a computational point of view, of the Leave-one-out approach.




It is also more severe than the other approach in the evaluation of the model predictive performance.

## Training/evaluation splitting

This approach can be applied in two different ways.

a) Single Evaluation Set, SES: the n objects are randomly divided into a training set and an evaluation set so that 10 to 50% of data are included in the latter.

This procedure usually leads to quite unstable models, since, obviously, a strong dependence of results on the dimension of the evaluation set and on the objects randomly included in the latter exists.

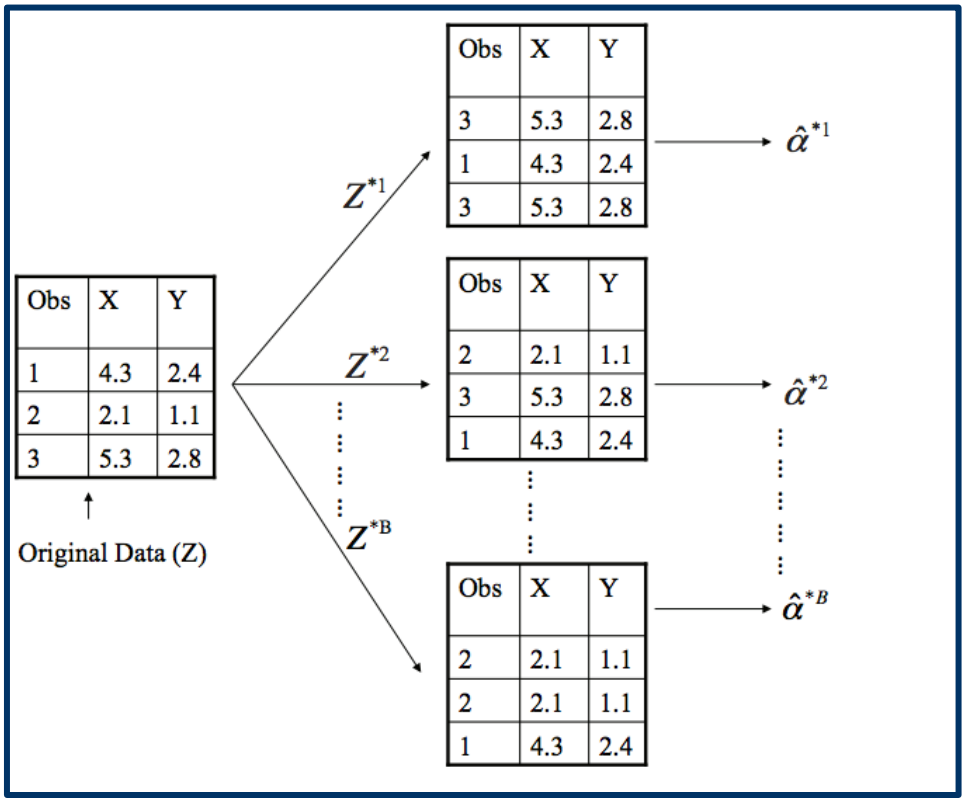
b) Repeated Evaluation Set, RES: the SES procedure is repeated many times, so that a reliable average value of a parameter can be obtained.

The main disadvantage of this approach is the remarkable computational time required, thus it cannot be considered a common strategy for model validation.

## Bootstrap

In this approach, B sets are constructed by extracting randomly, and also with repetitions, n objects from the original set of n data.

This means that some objects will compare more times in the same extracted set, whereas other ones may not compare at all:




The model obtained with extracted objects is used to predict the response of excluded objects.

This approach can be adopted only if powerful computers are available, since thousands of extractions of training sets are required to obtain a reliable estimate of a parameter average value when the number of observations is high.

## Validation of regression methods

As shown before, three fundamental quantities related to a regression method are the total (TSS), residual (RSS) and regression/model (MSS) sum of squares:

$$SS_{Tot} \equiv TSS = \sum_i \left(Y_i - \overline{Y}\right)^2 \qquad SS_{\mathrm{Res}} \equiv RSS = \sum_i (Y_i - \hat{Y}_i)^2 \qquad SS_{\mathrm{Reg}} \equiv MSS = \sum_i (\hat{Y}_i - \overline{Y})^2$$

The coefficient of determination is calculated as:

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$$

and corresponds to the square of the multiple correlation coefficient, indicated as r or R.

Given a response Y and m predictors (variables) and indicating as $\boldsymbol{R_{Y(1,\ldots,m)}}$ the correlation coefficient between the response and all predictors, the following properties can be cited:

1. $0 \leq \boldsymbol{R_{Y(1,\ldots,m)}} \leq 1$;

2. If $\boldsymbol{R_{Y(1,\ldots,m)}} = 0$, also the correlation between the response and each of the predictors, $\mathbf{R_{Y(j)}}$, is equal to 0 ;

3. If a single predictor is involved, $\boldsymbol{R_{Y(1)}} = |\, r_{YX} \,|$ , *i.e.*, the correlation of Y with the single predictor coincides with the absolute value of correlation coefficient between Y and X;

4. $R_{Y(1)} \leq R_{Y(1,2)} \leq R_{Y(1,2,3)} \leq \ldots \leq R_{Y(1,\ldots,m)}$ , *i.e.*, at the increase of the number of predictors the coefficient of correlation is decreased.

An adjusted-$R^2$ can also be defined for a regression model, to account for the degrees of freedom:

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \frac{RSS}{TSS}\left(\frac{n-1}{n-p}\right) = 1 - (1-R^2)\left(\frac{n-1}{n-p}\right)$$

This parameter has been introduced to evaluate the convenience in adding a further variable to the model.

Indeed, $R^2_{adj}$ exhibits a maximum when the model reaches the optimal complexity and is decreased when the addition of a further variable is not adequately compensated by a significant increase in the R value.

It is worth noting that both $R^2$ **e** $R^2_{adj}$ measure exclusively the model capacity to fit contemporarily all the n objects used in the construction of the model.

Different parameters need to be considered to evaluate the predictive capacity of the model.

## Parameters measuring the predictive capacity of the model

When the sum of residual squares is calculated by considering the values of predicted responses, instead of calculated responses, a new parameter is obtained, the Predictive Error Sum of Squares, formally identical to RSS:

$$PRESS = \sum_i (Y_i - \hat{Y}_{i/i})^2$$

where $\hat{Y}_{i/i}$ is the value predicted for the i-th sample using a model in which the sample has not been taken into account.

When PRESS is used instead of RSS, the percentage of variance explained by the model in prediction, $Q^2$, corresponding to the coefficient of determination for cross validation, $R^2_{CV}$, can be calculated:

$$Q^2 \equiv R^2_{CV} = 1 - \frac{PRESS}{TSS}$$

Similarly to $R^2_{adj}$, $R^2_{cv}$ exhibits its maximum when the model reaches its optimal complexity; however, it is evaluated with respect to the predictive power, not to the fitting level, of the model.

Further useful quantities related to RSS and PRESS are:

1. Standard Deviation Error in Calculation, SDEC (or SEC)

$$SDEC = \sqrt{\frac{RSS}{n}}$$

2. Standard Deviation Error in Prediction, SDEP (or SEP)

$$SDEP = \sqrt{\frac{PRESS}{n}}$$

3. Standard Error of Estimate, $S_y$ (or $S_e$)

$$s_y = \sqrt{\frac{RSS}{n-p}}$$

# A numerical example

Let us consider the following 17 responses (y), related to 5 independent variables ($x_1$, …, $x_5$):

| *ID* | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $y$ |
|---|---|---|---|---|---|---|
| 1 | 15.57 | 2463 | 472.92 | 18.0 | 4.45 | 566.52 |
| 2 | 44.02 | 2048 | 1339.75 | 9.5 | 6.92 | 696.82 |
| 3 | 20.42 | 3940 | 620.25 | 12.8 | 4.28 | 1033.15 |
| 4 | 18.74 | 6505 | 568.33 | 36.7 | 3.90 | 1603.62 |
| 5 | 49.20 | 5723 | 1497.60 | 35.7 | 5.50 | 1611.37 |
| 6 | 44.92 | 11520 | 1365.83 | 24.0 | 4.60 | 1613.27 |
| 7 | 55.48 | 5779 | 1687.00 | 43.3 | 5.62 | 1584.17 |
| 8 | 59.28 | 5969 | 1639.92 | 46.7 | 5.15 | 2160.55 |
| 9 | 94.39 | 8461 | 2872.33 | 78.7 | 6.18 | 2305.58 |
| 10 | 128.02 | 20106 | 3655.08 | 180.5 | 6.15 | 3503.93 |
| 11 | 96.00 | 13313 | 2912.00 | 60.9 | 5.88 | 3571.89 |
| 12 | 131.42 | 10771 | 3921.00 | 103.7 | 4.88 | 3741.40 |
| 13 | 127.21 | 15543 | 3865.67 | 126.8 | 5.50 | 4026.52 |
| 14 | 252.90 | 36194 | 7684.10 | 157.7 | 7.00 | 10343.81 |
| 15 | 409.20 | 34783 | 12446.33 | 169.4 | 10.78 | 11732.17 |
| 16 | 463.70 | 39204 | 14098.4 | 331.4 | 7.05 | 15414.94 |
| 17 | 510.22 | 86533 | 15524.00 | 371.6 | 6.35 | 18854.45 |
| mean | 148.27 | 18163.2 | 4480.62 | 106.32 | 5.89 | 4978.48 |
| st. dev. | 156.23 | 20642.8 | 4760.14 | 104.73 | 1.54 | 5394.51 |

The correlation matrix, $R_{xx}$, for the five predictors is:

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $x_1$ | 1 | 0.9074 | 0.9999 | 0.9357 | 0.6712 |
| $x_2$ | 0.9074 | 1 | 0.9071 | 0.9105 | 0.4466 |
| $x_3$ | 0.9999 | 0.9071 | 1 | 0.9332 | 0.6711 |
| $x_4$ | 0.9357 | 0.9105 | 0.9332 | 1 | 0.4629 |
| $x_5$ | 0.6712 | 0.4466 | 0.6711 | 0.4629 | 1 |

As apparent, a strong correlation exists between many of the considered variables.

OLS regression based on the following model (p = 6) is performed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

The main results in terms of coefficients of determination and related quantities are:

| *n.param.* | $R^2$ | $R^2_{adj}$ | $R^2_{cv}$ | $F$ | *SDEC* | $s_y$ | *SDEP* |
|---|---|---|---|---|---|---|---|
| 5 + 1 | 99.08 | 98.67 | 93.49 | 31.61 | 516.5 | 642.1 | 1376.2 |

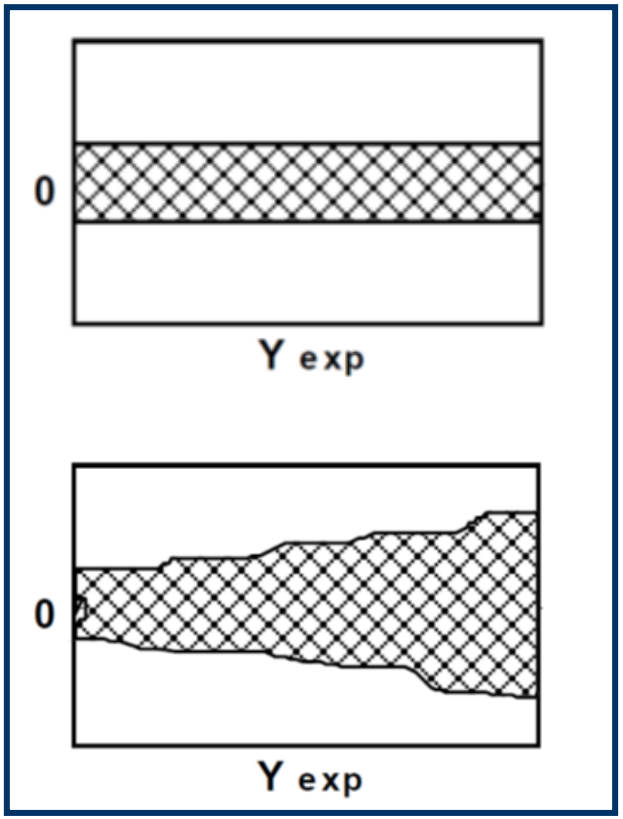
where: $F = \dfrac{MSS/m}{RSS/(n-p)} \sim F_{m,n\text{-}p}$ with m = 5 and n-p = 17-6 = 11

Since 31.61 > $F_{5,11(0.99)}$ = 5.32, the regression is significant at a 1% significance level.

# Diagnostic methods for regression: residuals and leverage values

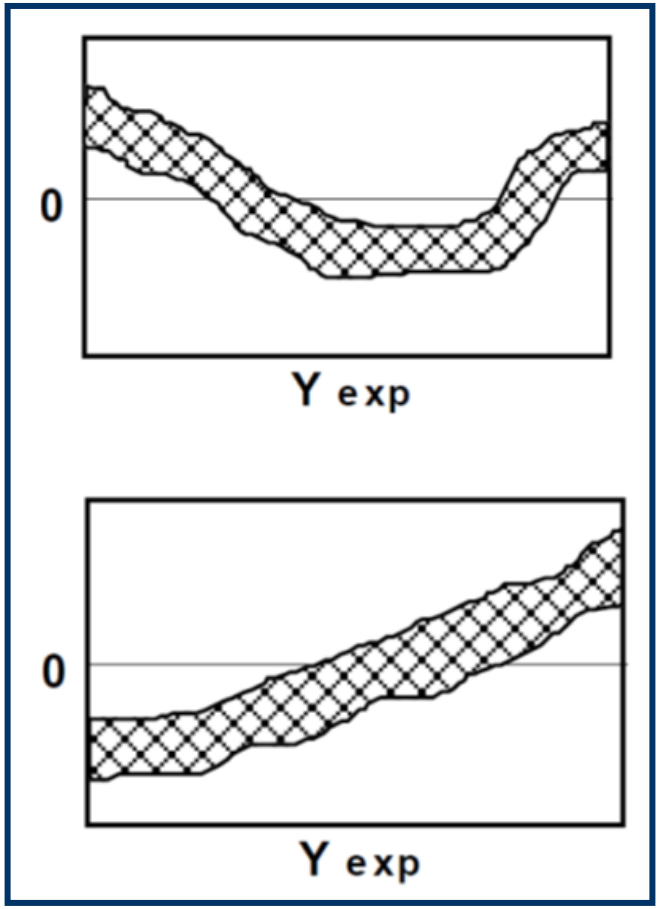
Residuals are exploited as a powerful quantity to assess the reliability of a regression model, since any deviation from the basic assumptions made for them (expected value for residual mean = 0 and random distribution around mean value) may indicate the presence of problems in the reliability of the model.

The graphical representation of residuals against the experimental response can rapidly provide interesting information:




Ideal case: residuals are distributed with mean = 0 and constant variance (homoscedasticity).

Residuals are increased at the increase of response, thus variance is not distributed homogeneously (heteroscedasticity)



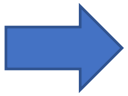

The choice of model is not adequate: non random trends are observed for residuals (they are systematically higher than 0 only for low and high values of the response, and systematically lower than 0 for intermediate value of the response).

A steady trend is observed for residuals: the presence of a systematic error can be suspected.

As shown for multiple linear regression, the vector of residuals, **e**, can be expressed as:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y}$$

where **I** is the identity matrix and **H** is the so-called hat matrix.

The variance of residuals is: $V(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$  $\mathrm{V}(e_i) = \sigma^2 (1 - h_{ii})$

where $h_{ii}$, the i-th element of the principal diagonal of matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$,

is known as leverage value.

Notably, if simple linear regression is considered, $h_{ii}$ is given by: $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

thus the leverage value is a measure of the distance between the abscissa of the i-th value and that of the regression centroid.

At the centroid the minimum leverage value is obtained: $h_{\min} = \frac{1}{n}$

The sum of leverage values is equal to 2, *i.e.*, the number of parameters for simple linear regression, thus the average leverage value is 2/n.

In the case of multiple linear regression the average leverage value is thus p/n.

Equation $V(e_i) = \sigma^2 (1 - h_{ii})$ indicates that residuals referred to points with a high leverage have a smaller variance.

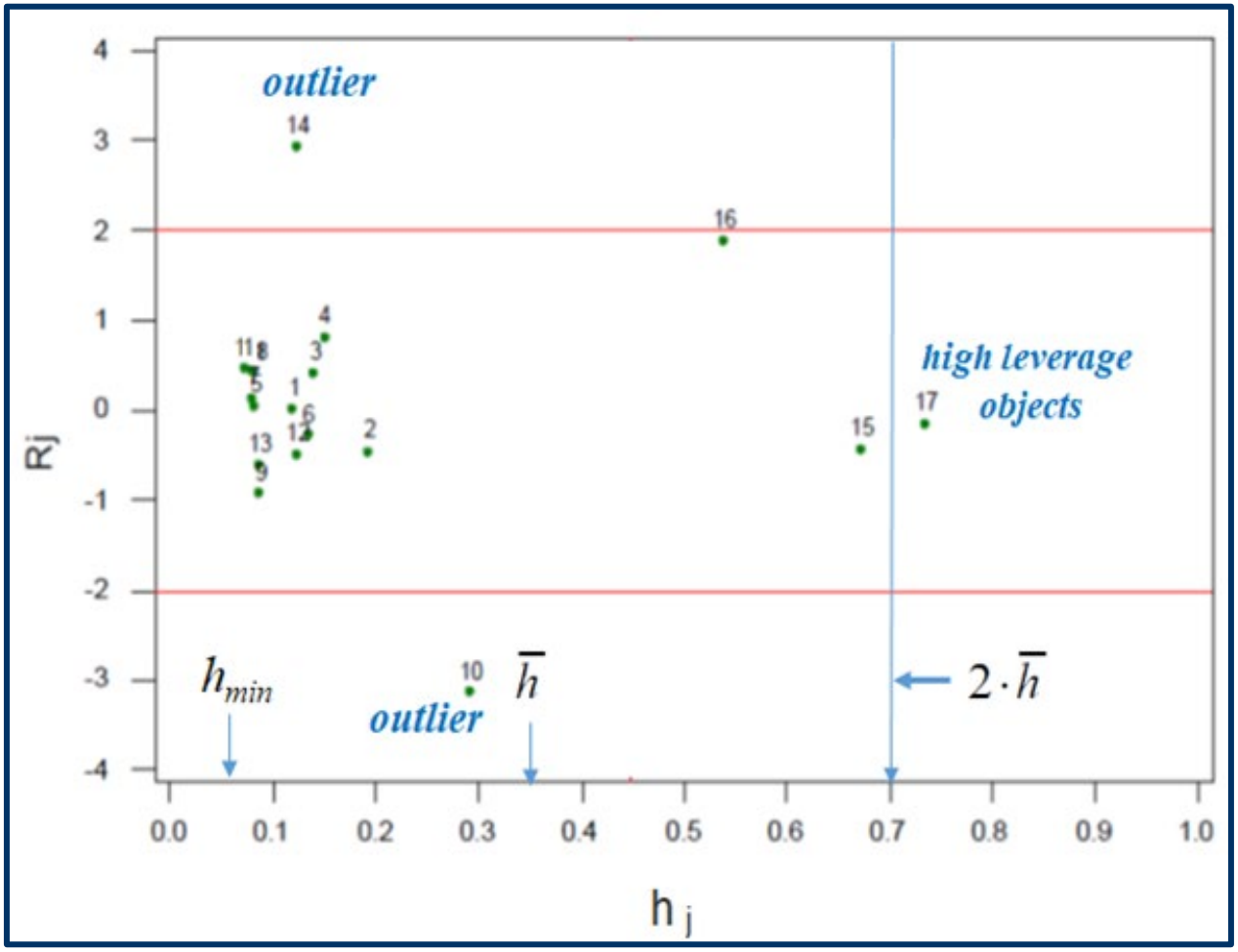
In order to consider jointly residuals and their variability, standardized residuals (also called internally studentized residuals) can be calculated:

$$r_i = \frac{e_i}{s_e \sqrt{(1 - h_{ii})}} \qquad i = 1,2,\ldots,n \quad \text{where:} \quad s_e^2 \equiv s_y^2 = \frac{RSS}{n - p}$$

A useful graphical representation is the Williams Plot, in which standardized residuals are plotted versus leverages:

In this case eventual outliers can be recognized from their position above or below horizontal lines corresponding to a standardized residual greater than 2, in absolute value.

High leverage observations can be recognized from their position on the right with respect to the vertical line corresponding to a $h_j$ value greater then twice the mean value of all $h_j$ values.



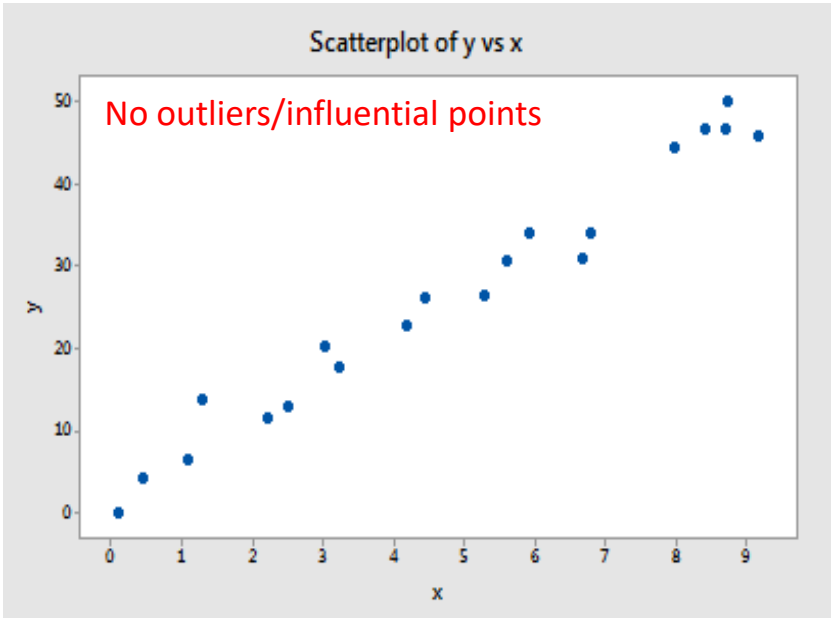

In the figure a set of n = 17 data was considered and a multiple linear regression based on 5 variables was performed, thus p = 6 and: $h_{min} = 1/17 = 0.059$ and $\bar{h} = 6/17 = 0.35$.

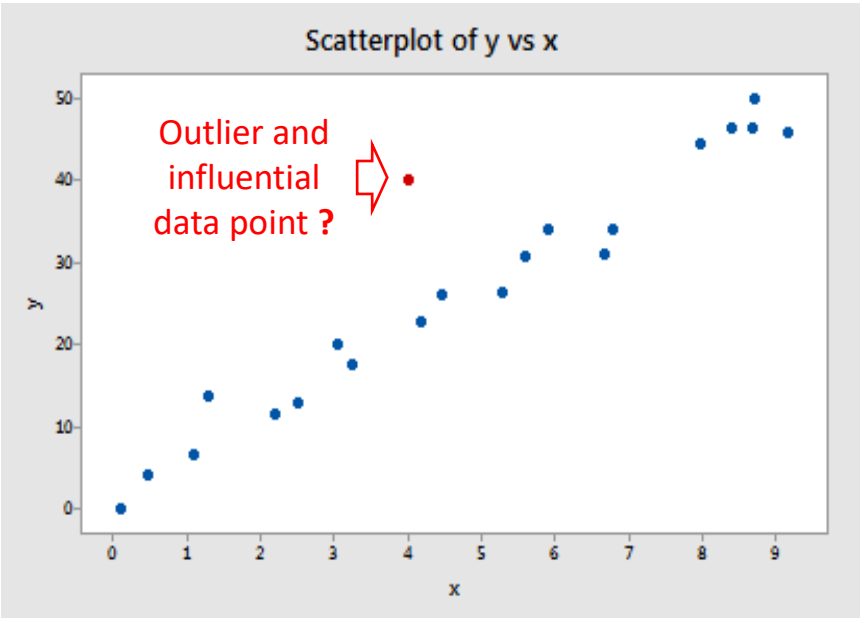
# Outliers and influential observations in simple linear regression

In the context of regression models, an outlier is a data point whose response y does not follow the general trend of data.

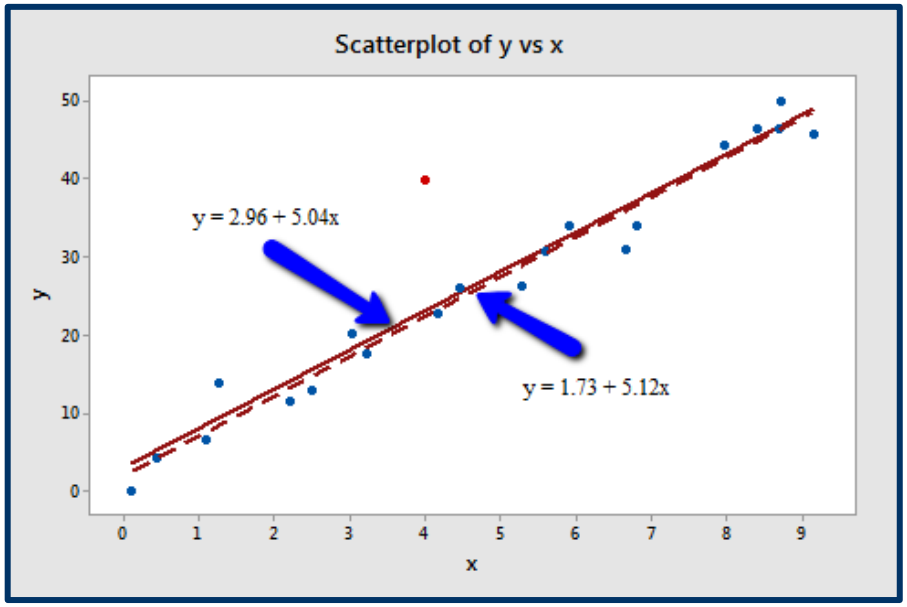
A data point is influential if it is able to influence remarkably any aspect of a regression analysis, such as estimated slope, predicted responses, etc.

A comparison between datasets including and not including a potential outlier/influential point, respectively, is shown in the following figure, referred to the case of a single regressor:



Best fitting lines obtained in the two cases are shown in the figure on the right:




Apparently, the presence of the presumed outlier in the dataset seems to have only a slight influence on regression parameters (especially on the slope).

Summaries of statistical information are the following:

Model Summary ***(red data point included)***

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 4.71075 | 91.01% | 90.53% | 89.61% |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 2.96 | 2.01 | 1.47 | 0.157 | |
| x | 5.037 | 0.363 | 13.86 | 0.000 | 1.00 |

Regression Equation

y = 2.96 + 5.037 x

Model Summary ***(red data point excluded)***

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 2.59199 | 97.32% | 97.17% | 96.63% |

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 1.73 | 1.12 | 1.55 | 0.140 | |
| x | 5.117 | 0.200 | 25.55 | 0.000 | 1.00 |

Regression Equation

y = 1.73 + 5.117 x

Minor side effects are observed when including the red data point:

1. The $R^2$ value is slightly lower but the relationship between y and x still appears strong;

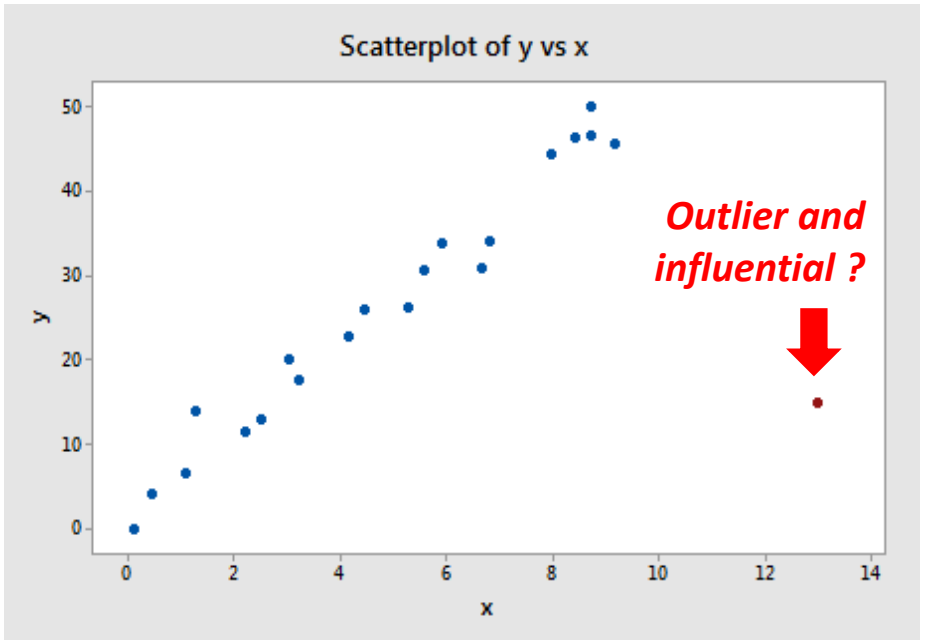
2. The standard error (SE) for parameter $b_1$, *i.e.*, the line slope, is larger when the red data point is included;

3. In each case the P-value referred to the hypothesis $H_0$: $\beta_1 = 0$ is less than 0.001. Consequently, there is sufficient evidence to conclude that, in the population, *x* is related to *y*.

Based on these results, the red data point cannot be considered influential but is still candidate as an outlier.

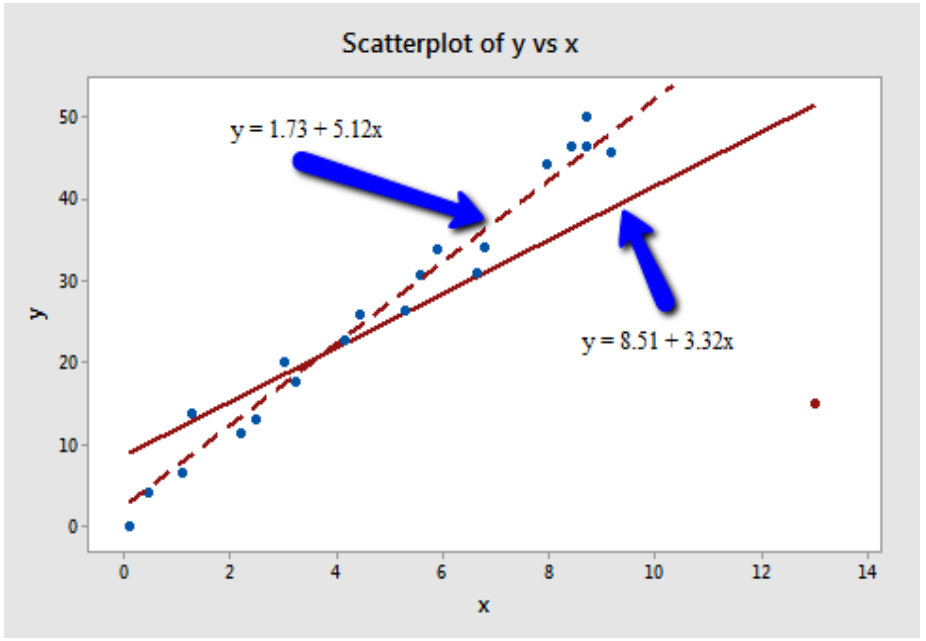
Notably, the standardized residual for that point is equal to 3.68. Based on the considerations made before about the Williams Plot, the point can actually be considered an outlier.

Let us consider a different data set:




Not surprisingly, the two best fitting lines (with or without the suspect point) are substantially different in this case.

Indeed, the presence of the red data point lowers the regression line slope from 5.117 to 3.320.



Further relevant consequences can be inferred from the summary of statistical information:

**Model Summary** ***(red data point included)***

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 10.4459 | 55.19% | 52.84% | 19.11% |

**Coefficients**

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 8.50 | 4.22 | 2.01 | 0.058 | |
| x | 3.320 | 0.686 | 4.84 | 0.000 | 1.00 |

**Regression Equation**

$y = 8.50 + 3.320\ x$

**Model Summary** ***(red data point excluded)***

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 2.59199 | 97.32% | 97.17% | 96.63% |

**Coefficients**

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 1.73 | 1.12 | 1.55 | 0.140 | |
| x | 5.117 | 0.200 | 25.55 | 0.000 | 1.00 |

**Regression Equation**

$y = 1.73 + 5.117\ x$

1. The $R^2$ value (R-sq) is significantly decreased (from 97.32 to 55.19%);
2. The SE on $b_1$ is almost 3.5 times larger (from 0.200 to 0.686);
3. The predicted responses are clearly affected (with $R^2$ in prediction, R-sq(pred), falling from 96.63 to 19.11%)
4. In any case, as suggested by P-Values, there is still sufficient evidence, at 0.05% significance level, that a relationship exists between x and y.

The red data point can thus be considered both an outlier and an influential point.

Influential points can also be recognized through a quantity known as the Cook Distance:

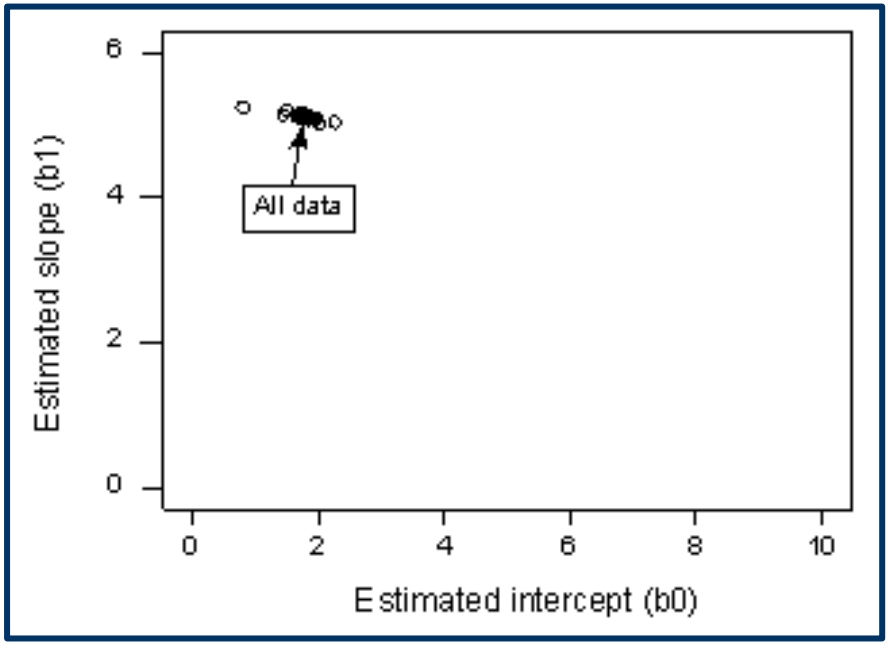
$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})} = \frac{(y_i-\hat{y}_i)^2}{p \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \qquad i=1,2,....,n$$

As apparent, the distance depends on both the residual and the leverage.

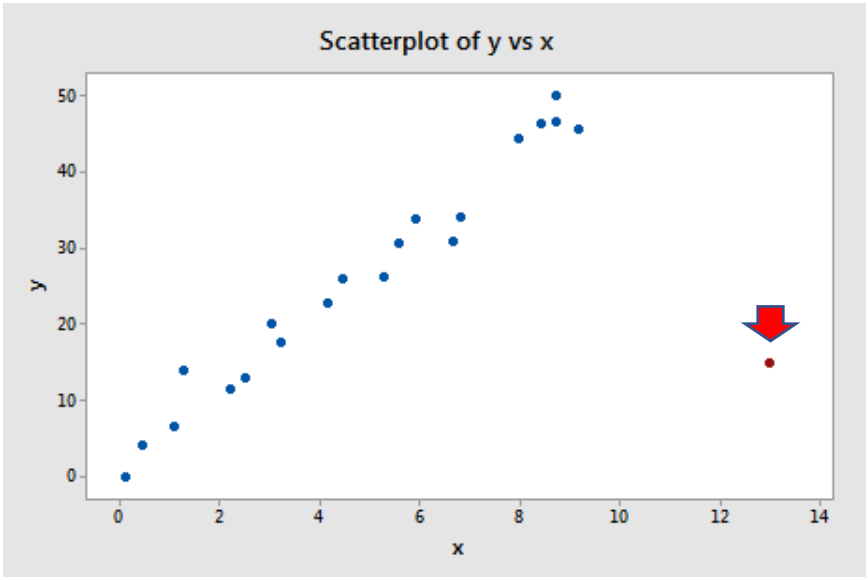
In order to appreciate the potential of Cook Distance in terms of influential points recognition, the first dataset shown before, the one without potential outliers or influential points, can be considered.

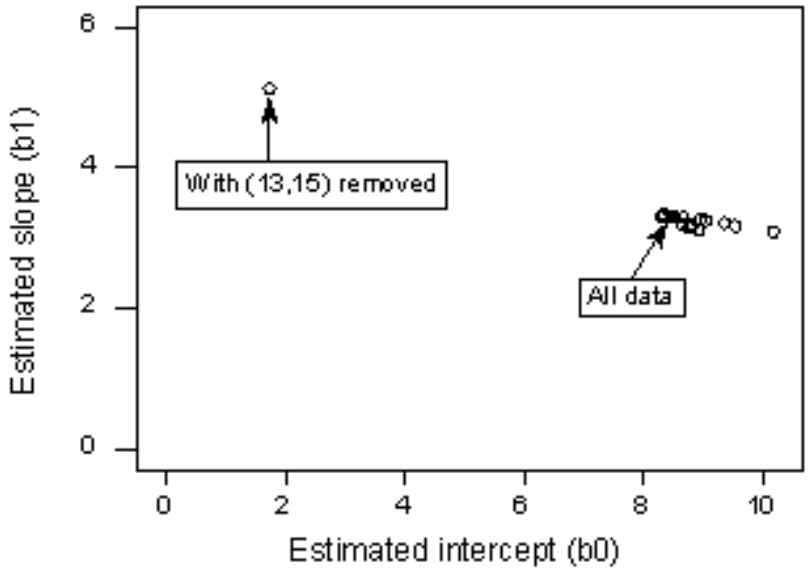
If slopes and intercepts obtained for regression lines when considering all the datapoints (black circle) or when deleting one of them at a time (open circles), are plotted, the plot shown on the right is obtained:

All estimated coefficients are grouped together, thus all Cook Distances are expected to be small.



This is clearly not the case for the dataset including also a datapoint with x = 13 and y =15:




The Cook Distance for point with co-ordinates (13, 15) is then expected to be much larger than those referred to other datapoints. This is confirmed by calculations. In fact, the Cook Distance for datapoint (13, 15) is equal to 4.048, whereas all the others are comprised between 0.00002 and 0.09180.

General guidelines for the use of Cook Distance are:

1) If $D_i > 0.5$, the i-th point is worthy of further investigation
2) If $D_i > 1$ the i-th point is likely influential
3) If $D_i >> 1$ the i-th point is most certainly influential

## Using leverages to identify extreme x values

Given an observation i, the predicted response can be written as a linear combination of the n observed responses:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \qquad \text{for } i = 1, \ldots, n$$

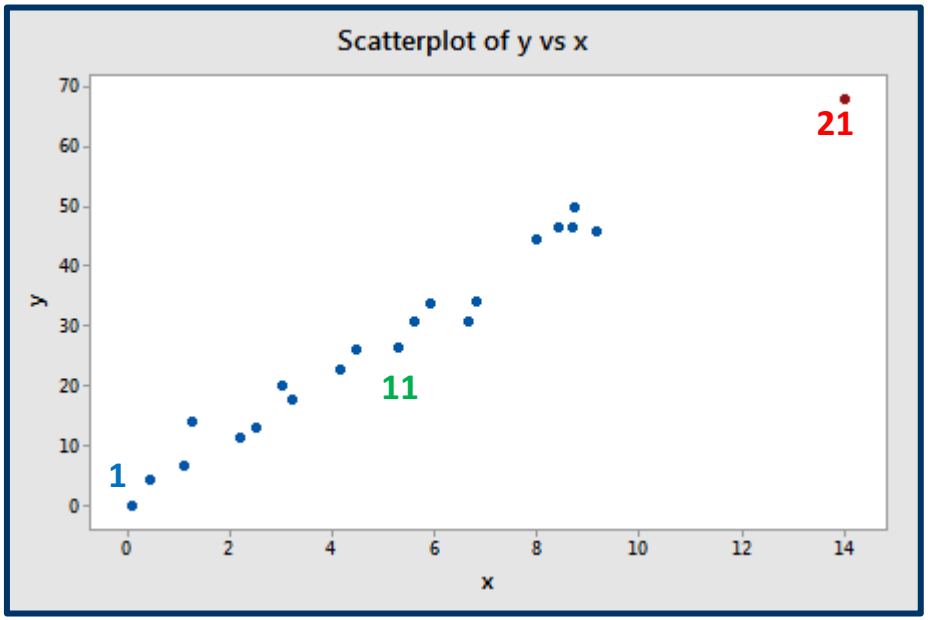
where weights $\mathbf{h}_{i1}$, $\mathbf{h}_{i2}$, ..., $\mathbf{h}_{ii}$, ..., $\mathbf{h}_{in}$ , *i.e.*, leverages, depend only on the predictor values.

All predicted values can thus be expressed as:

$$\begin{aligned}
\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n \\
\hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \cdots + h_{2n}y_n \\
&\vdots \\
\hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \cdots + h_{nn}y_n
\end{aligned}$$

It is then clear that a leverage quantifies the influence that the observed response $y_i$ has on its predicted value $\hat{y}_i$.

Let us consider the following scatter plot:




Data point #21 clearly appears to be quite far from all the others.

In this case leverage values for datapoints #11 and #1 are 0.048 and 0.153, respectively, whereas the one for datapoint #21 is 0.358.

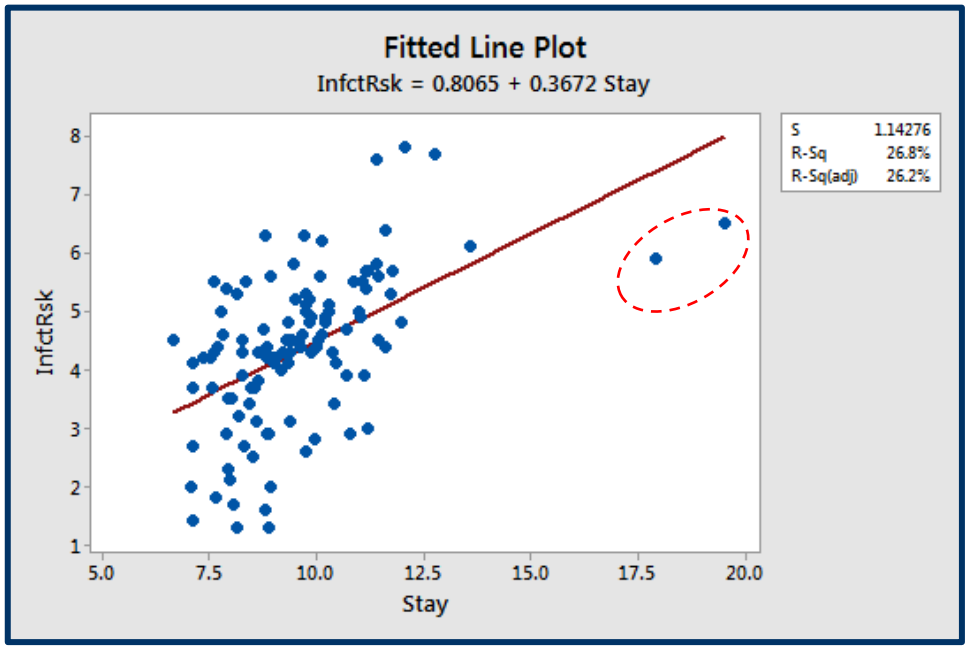
In the present case the average value of leverages, *i.e.*, 2/n, is equal to 2/21 = 0.095; leverage value for datapoint #21 is thus much larger than 2/n; specifically, it is even larger than 3(2/n), *i.e.*, 0.286.

Data point #21 can thus be considered a high leverage point but it is a non influential point, since, due to its ordinate, it appears located near the regression line obtained for all the other values.

# A peculiar real case

The regression plot shown in the figure on the right was obtained by considering data on hospital infection risk as a function of the average length of stay in 112 hospitals in the United States.

In two cases, the length of stay was quite large but the infection risk did not appear to be correspondingly large.




The following table was obtained from these data when using the Minitab's option «Fits and Diagnostics» in the Regression: Results window:

| Obs | InfctRsk | Fit | SE Fit | 95% CI | Resid | Std Resid | Del Resid | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1.600 | 4.045 | 0.117 | (3.812, 4.277) | -2.445 | -2.15 | -2.19 | R | |
| 40 | 1.300 | 3.802 | 0.137 | (3.532, 4.073) | -2.502 | -2.21 | -2.25 | R | |
| 47 | 6.500 | 7.988 | 0.585 | (6.828, 9.148) | -1.488 | -1.52 | -1.53 | | X |
| 53 | 7.600 | 4.996 | 0.150 | (4.699, 5.293) | 2.604 | 2.30 | 2.35 | R | |
| 54 | 7.800 | 5.238 | 0.179 | (4.884, 5.592) | 2.562 | 2.27 | 2.31 | R | |
| 93 | 1.300 | 4.082 | 0.115 | (3.853, 4.310) | -2.782 | -2.45 | -2.50 | R | |
| 111 | 5.900 | 7.393 | 0.494 | (6.415, 8.372) | -1.493 | -1.45 | -1.46 | | X |

In the table Standardized Residuals correspond to internally studentized residuals, whereas Deleted Residuals correspond to externally studentized residuals:

Std Resid

$$r_i = \frac{e_i}{s_e\sqrt{(1-h_{ii})}}$$

Del Resid

$$t_i = \frac{e_i}{s_{e(i)}\sqrt{(1-h_{ii})}}$$

Note that externally studentized residuals are calculated by considering a special estimate of $s_e$, indicated as $s_{e(i)}$, obtained by ignoring the i-th value in its calculation.

Two observations, those numbered as #47 and #111, were marked in the table by an «X» sign. They correspond to observations corresponding to longest average stays.

On the other hand, neither standardized nor deleted residuals were particularly large, in absolute value, for the two observations.
This outcome was clearly influenced by the presence of large residuals for several observations corresponding to shorter average stays.