# Of small numbers with big influence – The Sum Of Squares

*Dr. Peter Paul Heym – Sum Of Squares*

Often, the small things make the biggest difference in life. Sometimes these things we do not recognise at first as big but as soon as we draw our attention to them, they become more important.

In analytical method validation, one of these small things is the so-called *sum of squares* or *residual sum of squares* (*RSS*). The *(residual) sum of squares* you will often find as a number in validation reports that, at first sight, might be of no interest at all. That is why, in this article, we will explain in more detail what this number actually means and why it is of importance. We will discuss its meaning and its importance, as well as the pros and cons and what can be done to avoid statistical pitfalls when using the *RSS*.

## Analytical Method Validation and ICH Q2(R1)

When validating an analytical method, the ICH guideline Q2(R1) is of importance. This guideline describes which characteristics of the method need to be evaluated in order to verify that a method is suitable for its intended purpose (*fit-for-purpose)*. How to do that is well-described in detail in "**Validation of Analytical Procedures: Text and Methodology Q2(R1)"**, where ICH stands for *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* [1]. For assays or quantitative impurities, not only characteristics as *specificity* and *range*, or *accuracy* and *precision* have to be evaluated, but also, the linearity of the method needs to be assessed. Linearity in this context is defined as the ability (within a given range) to obtain test results which are directly proportional to the concentration (amount) of analyte. The basis for evaluating linearity is the linear relation between the analyte concentration and the signal. It is easy to visualise the data and the linear relationship, as displayed in Figure 1.

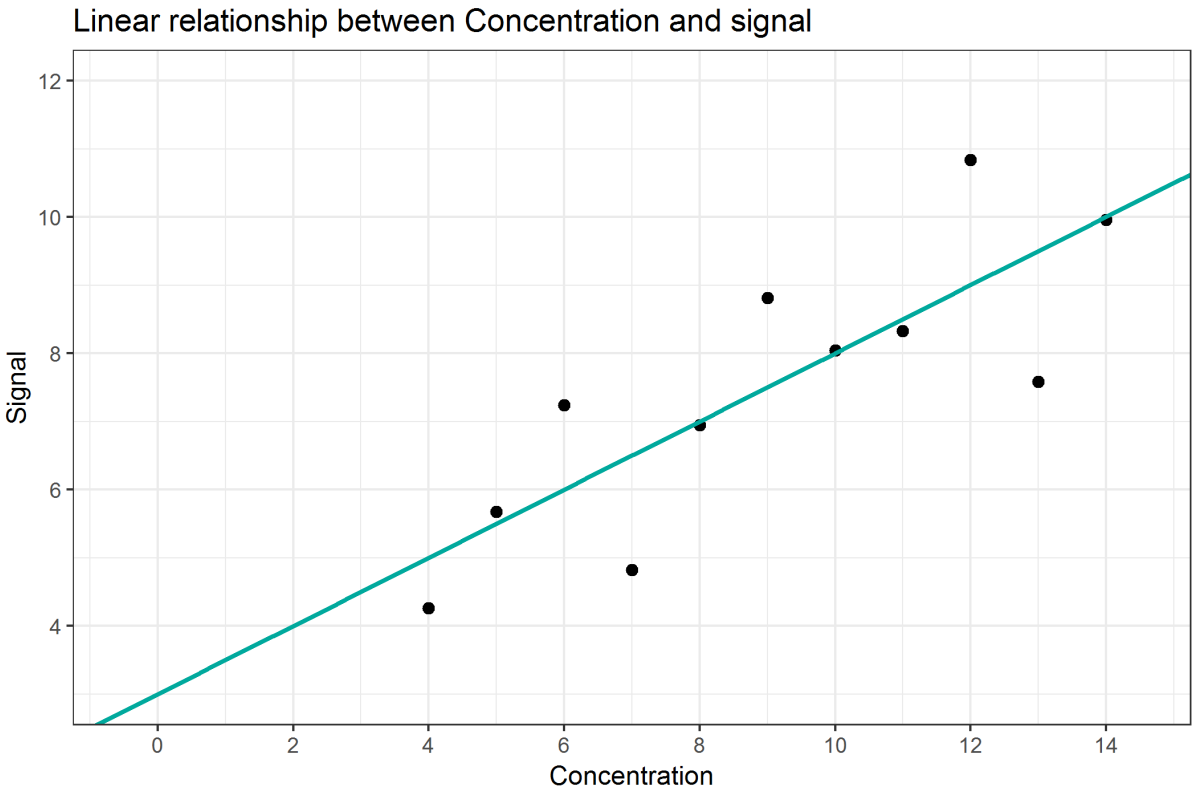


*Figure 1: Linear Relationship between analyte concentration and signal*

Using simple linear regression, the relation between analyte concentration and signal can be established. Here, with increasing concentration, a signal of increasing intensity is observed. The linear relationship can be described using the slope and the y-intercept of a straight line. In the example, at a concentration of 0, we would still get a signal intensity of 3. And for every increase in one concentration unit we would get an increase of 0.5 units of signal intensity. Therefore, the mathematical relationship between concentration and signal would be: $\text{signal} = 3 + 0.5 * \text{concentration}$ or $\text{y} = 3 + 0.5 * \text{x}$. It is not difficult to use programs like Excel or other statistical tools to identify this relationship, when applying simple linear regression and the method of least squares.

Considering the actual data points, we notice that, for example, we have actually observed a signal of 8.81 for a concentration of 9. Inserting concentration $x = 9$ into the equation would result in a concentration of $3 + (0.5 * 9) = 7.5$. The measured values and regression line prediction do actually differ. The regression line describes the best-possible linear relationship between signal and concentration, and still, it is not capable of exactly predicting the observed values due to the fact that every single measurement value is deviating from the regression line. These differences of observed vs. predicted values are the reason why the regression line is not able to describe this relation 100% exactly. In our example, the regression line can only describe this relationship with 66.7%. This percentage corresponds to the well-known characteristic “coefficient of determination”, or $R^2$ and it describes the percentage of variability in the data that can be explained by the applied regression line. The remaining percentage of 33.3% remains unexplained by the regression line.

## Will one graph be enough to match ICH requirements?

Let's assume we already have calculated the regression line describing the linear relationship between concentration and signal, hence we also know its slope and y-intercept. Using Excel or other tools for calculating the regression line we also know the coefficient of determination. Are these measures enough to define our method as being fit-for-purpose regarding linearity? Let's have a look into the guideline:

*"A **linear relationship** should be evaluated across the range … of the analytical procedure. … Linearity should be evaluated by visual inspection of a **plot of signals** as a function of **analyte concentration** or content. If there is a linear relationship, test results should be evaluated by appropriate statistical methods, for example, by **calculation of a regression line**… . The **correlation coefficient, y-intercept, slope** of the regression line and **residual sum of squares** should be submitted."*

Most of the characteristics were already calculated. A graphical representation of the data and the regression line (including the formula showing slope and y-intercept) can be set up easily in Excel, the $R^2$ value can be included in the figure with a single click in Excel as well. Rooting $R^2$ results in r, which is the correlation coefficient, also required by the ICH. According to ICH, we already have almost everything we need to address for evaluating linearity. The only thing left is hidden in the last sentence. So, the last thing we need to clarify is the *residual sum of squares*.

## From single values to *residual sum of squares*

What is the so-called *residual sum of squares*, which is oftentimes abbreviated with *RSS*? Is there any link from the data or regression line to the *RSS*? And why would we need this value to satisfy ICH requirements? As we have already noticed, the regression line predicts a signal of 7.5 for the concentration of 9, although the actual measurement was 8.81. We keep in mind that the $R^2 = 0.67$ means that the regression line is able to explain only 67% of variation in the data. Therefore, the remaining 33% of variation cannot be explained by the regression line. This corresponds to the fact, that the regression line predicts values for the signal which are different from the actual values. As in the example of concentration 9, the regression line predicts a signal of 7.5, which deviates by 1.31 units from the actual data point $(8.81 - 7.50) = y - \hat{y} = 1.31$. In this case, the regression line *under*estimates the actual value by 1.31 units. Similarly, the difference between the predicted and actual signal for concentration 8 is $(6.95 - 7.00) = -0.05$, meaning that here, the regression line *over*estimates the signal by 0.05 units. This difference is called *error*, *residuum* or *residual*, and is often denoted by $\varepsilon$ (epsilon). The residuum for concentration $x = 4$ we would write as follows:
$\varepsilon_1 = y_1 - \hat{y}_1 = 4.26 - 5.00 = -0.74$. Graphically, we can illustrate the residuals of each data point as grey lines in the scatter plot (Figure 2). The length of each line corresponds to the magnitude of prediction error, $\varepsilon$, for each concentration. Why don't we sum up all the errors to get one value for the total error of prediction? Wouldn't this correspond to the "goodness" of the model?

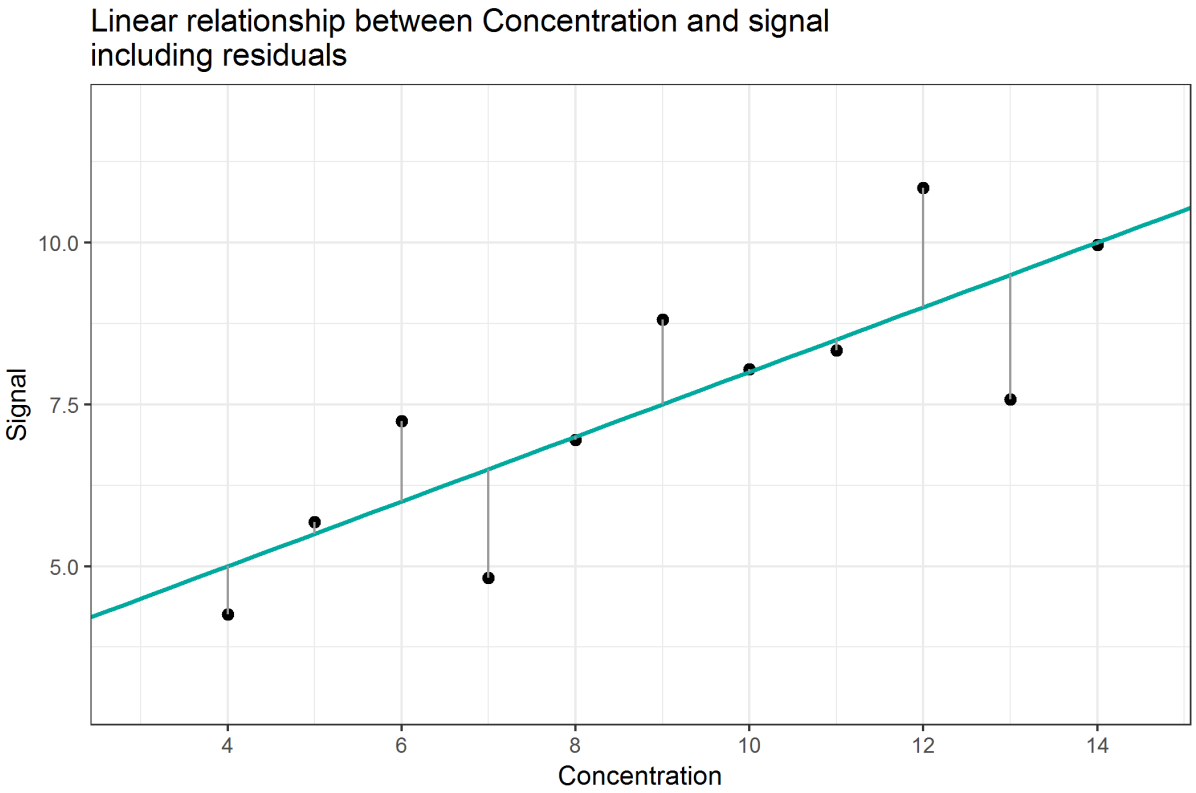


*Figure 2: Linear Relationship between analyte concentration and signal, including residuals*

As calculated in Table 1, the sum of all errors (the *sum of residuals*) is resulting in 0. This is because errors can positive or negative, as well – the model underestimates and overestimates. By summing up the errors, all errors compensate each other. This is a fundamental characteristic of the regression line and the method of least squares. Therefore, the sum of residuals can't be used as an indicator of how well the regression line fits the data.

*Table 1: Residuals*

| X-values | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Residuals** | -0.740 | 0.179 | 1.239 | -1.680 | -0.050 | 1.309 | 0.039 | -0.171 | 1.838 | -1.921 | -0.041 | 0 |

Since the sum of residuals is not an adequate measure to evaluate the validity of the regression model, we need to find another way to still use the residuals. To get rid of all negative values, we square the residuals (Table 2). This simple transformation has two advantages: First, the square of a number is always positive and as a consequence, each squared error cannot be negative anymore. Additionally, for small errors, e.g. errors less than 1 unit, the square of it will be even smaller and closer to 0 (e.g. $0.5^2 = 0.25$). And for residuals bigger than 1, the squared residuals will be even bigger (e.g. $2.5^2 = 6.25$). That means that small deviations from the regression line will be "rewarded", while bigger deviations will be "penalised".

*Table 2: X-values, residuals and squared residuals*

| X-values | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Residuals | -0.740 | 0.179 | 1.239 | -1.680 | -0.050 | 1.309 | 0.039 | -0.171 | 1.838 | -1.921 | -0.041 | 0 |
| Squared residuals | 0.547 | 0.032 | 1.535 | 2.822 | 0.002 | 1.713 | 0.001 | 0.029 | 3.378 | 3.690 | 0.001 | 13.754 |

Consequently, in a model with "larger" deviations, the sum of the squared residuals will be "larger" than in a model with "smaller" deviations. And therefore, we can use the sum of the squared residuals as a measure to evaluate model quality. As can be seen in Table 2, the sum of the squared residuals results in 13.75. This is actually the so-called *residual sum of squares*, or *RSS*. This is the value that the ICH requires in method validation. It describes how much difference exists between the measured values and the ideal values of the regression line. Thus, small RSS values should always represent good mathematical models respectively calibration lines.

Now we know what the residual sum of squares is, but we still have not discussed the actual meaning of this measure. We know that smaller *RSS* might indicate a "better" model, or a "good" relationship between signal and concentration. But then, does this indicate that the method is *fit-for-purpose*? What is the threshold for the *RSS*? Or is there a threshold at all? In the ICH Q2(R1) guideline, there is no reference or comment regarding this aspect. It says that we need to state this number, but no interpretation of this number is needed. This is strange because the *RSS* is an indicator of model quality.

Since there is no *RSS* limit based on which we could declare the method as being *fit-for-purpose*, can we set a limit by ourselves? And if we are aiming for small *RSS*, could we just transform the data to decrease the *RSS* and make it look better? We could try by changing the units of the signal (Table 3). Let's assume we would have a *RSS* value of 13.75, and the signal is given in Litres. What would happen if we use the exact same data, but transforming them to Gallons (1 Gallon = 4.54609 L)? What would happen to the regression line characteristics, and especially to the *residual sum of squares*?

*Table 3: Differences in slope, y-intercept, RSS and R-squares, depending on signal units*

| | slope | y-intercept | RSS | $R^2$ |
|---|---|---|---|---|
| **Litres** | 0.500 | 3.000 | 13.754 | 0.667 |
| **Gallons** | **0.109** | **0.660** | **0.666** | **0.667** |
| **Transformation** | $0.5 = 0.109 * 4.54609$ | $3.0 = 0.66 * 4.54609$ | $13.754 = 0.666 * 4.54609^2$ | |

The change of signal units would result in a change of regression characteristics, especially the slope, y-intercept and also in the *residual sum of squares*. Only, the $R^2$ value stays the same, which makes sense because there is still the same relationship between concentration and signal, it is independent of units. Astonishingly, the transformation results in a *RSS* of 0.666, a reduction of about 95% (!). That would sound much better in the validation report, wouldn't it? ... having a *RSS* of less than 1 instead of more than 13.75...

And still, you cannot fool statistics because you can always back-transform the data using the factor of 4.54609, which will reveal the original values. Therefore, it doesn't make sense to define any upper

limit for the *RSS* because this characteristic depends on the method itself. It depends on the number of data points, e.g. on the number of replicates and concentrations, but also on the chosen values of concentrations. Thus, the *RSS* must always be considered together with the method itself, the $R^2$ and the characteristics of the regression line. Providing *RSS* without the other aspects is without any value. Therefore, it makes sense that the ICH is requiring the *RSS* **without** any limit for the model. There cannot be a single limit for the *RSS*. As the *RSS* should always be regarded with all other data of the regression model, the following sentence of the guideline is absolutely warrantable:

*"The … **residual sum of squares** should be submitted. A **plot of the data** should be included. In addition, an **analysis of the deviation of the actual data points from the regression line may also be helpful** for evaluating linearity."*

This sentence is of importance, even if the word *may* does not suggest that. Although the *RSS* gives us no hint whether the method satisfies any *RSS* condition, it could still happen that most of the data points contribute little to the overall *RSS* but one data point is far away from the regression line, such that its residual has a very big contribution to the overall *RSS*. Since the *RSS* equals the squared deviations from the regression line for **all** data points, we still do not know if there are some data points influencing the *RSS* more than others are doing. Therefore, it is suggested to further analyse the deviations and not only rely on the *RSS* itself.

If the method is *fit-for-purpose*, then the regression line should represent a physical or chemical relationship that exists between concentration and signal which, again, is dependent on the method itself. Therefore, the observed data points should also satisfy this relationship and we would not expect too large deviations from the physico-chemical relationship in our data. Each single data point should be close to this relationship and therefore close to the regression line. The *RSS* as a single measure cannot answer the question if one or more data points deviate too much from that relationship. To investigate the contribution of each data point to the *RSS*, we need to have a closer look into the data points and their deviation from the regression line. This might be the actual intention of the guideline: *„**analysis of the deviation of the actual data points from the regression line**"*.

## From *residual sum of squares* back to single values

One idea to investigate the contribution of each data point is to measure its contribution by dividing its squared residual by the *RSS*. By doing so, we get a percentage of contribution and recognise that only 3 out of 14 data points (in bold) contribute about 70% to the total *RSS*. The data is summarised in Table 4:

*Table 4: Contributions to the RSS by each data point*

| X-values | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Residual** | -0.740 | 0.179 | 1.239 | -1.680 | -0.050 | 1.309 | 0.039 | -0.171 | 1.838 | -1.921 | -0.041 | 0 |
| **Squared residual** | 0.547 | 0.032 | 1.535 | 2.822 | 0.002 | 1.713 | 0.001 | 0.029 | 3.378 | 3.690 | 0.001 | 13.754 |
| **Contribution (%)** | 3.98 | 0.23 | 11.16 | **20.52** | 0.01 | 12.45 | 0.01 | 0.21 | **24.56** | **26.82** | 0.01 | 100 |

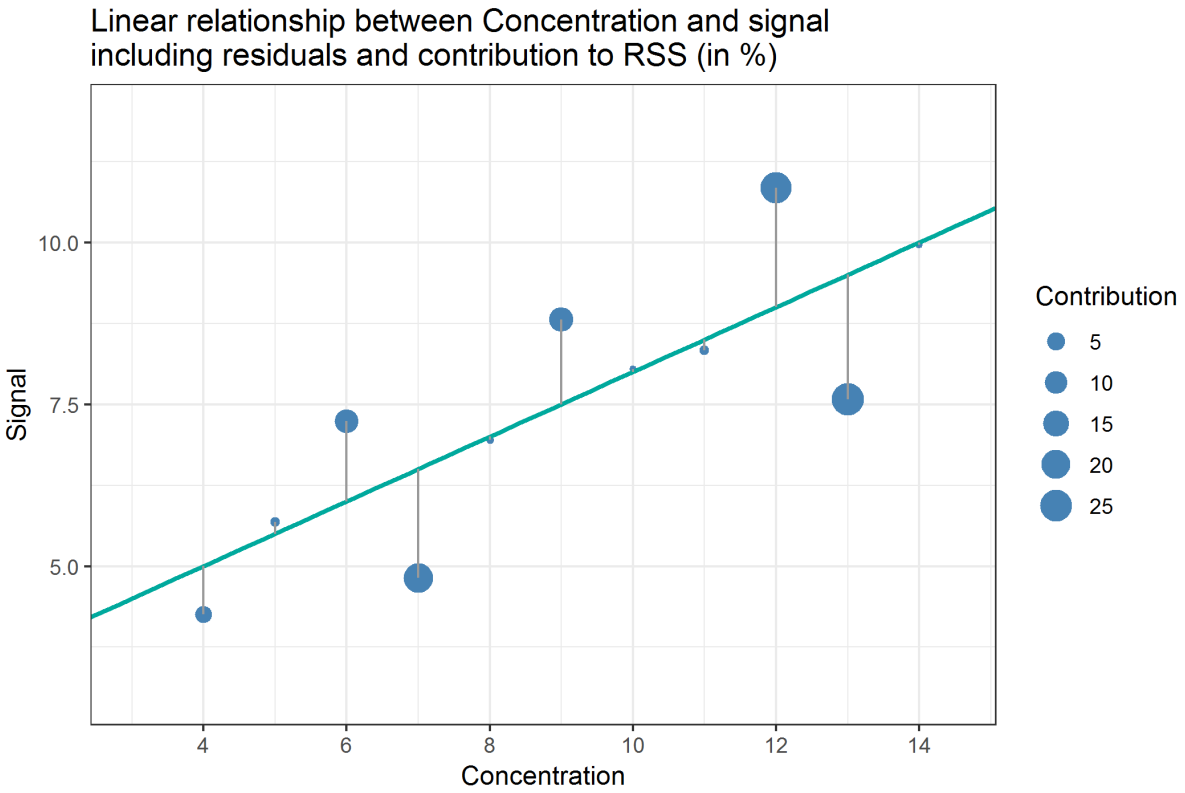


*Figure 3: Linear Relationship between concentration and signal; size of data point corresponds to its contribution to RSS*

We can also display the contribution of each data point by scaling the size of each data point according to its contribution (Figure 3). By doing so, we can at least visualise which data points contribute more than others. Now, which pattern of sizes would we expect as *"normal"* contributions? Again, using the percentages does not give us an exact answer to that. So, is there anything else we can do to find answers about that?

## *Hat Values* and *Cook's Distance* – what is really influencing the regression line?

So far, we were thinking about the influence of data points, but have actually not clarified what *influence* actually means. One intuitive way to think about that is to consider what would happen to the regression line if a single data point would be removed from the data set. If one data point has a big influence on the regression line, then, removing that data point should change the regression line a lot, which can be measured by a difference in the slope and / or y-intercept. This can be done and is shown in the following Figure 4:

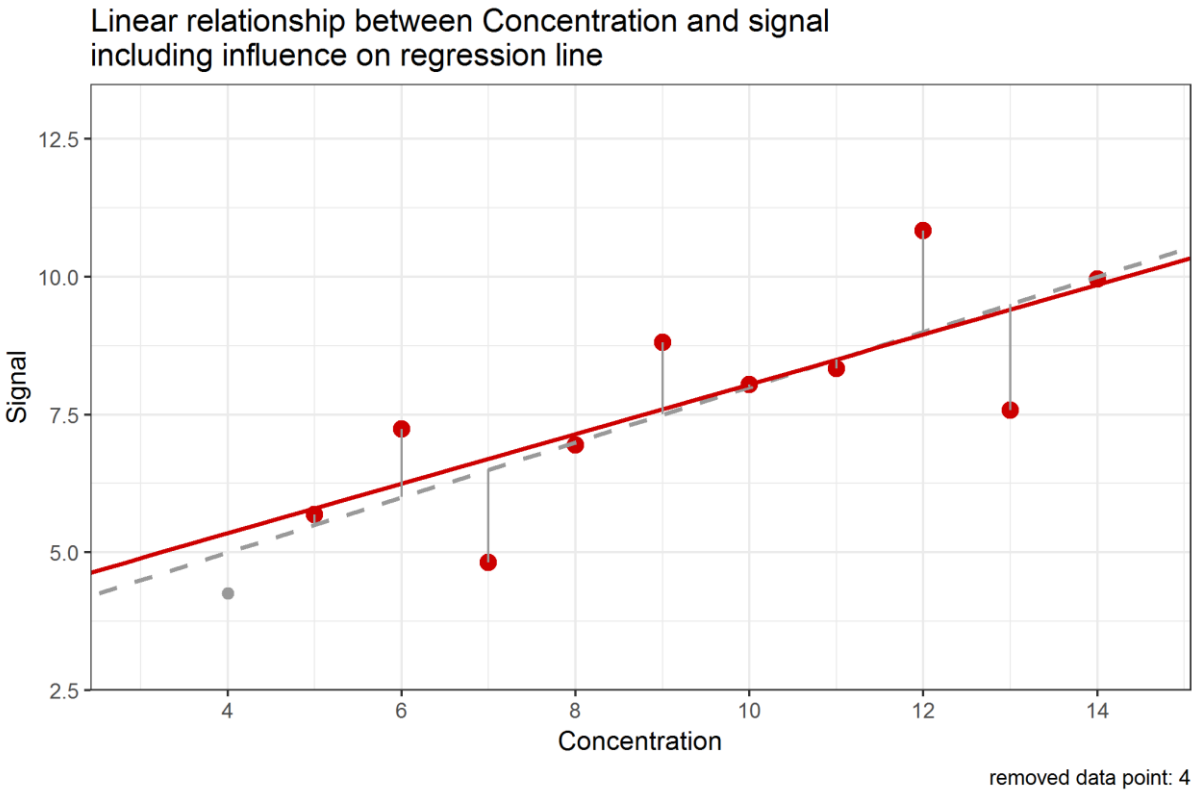


*Figure 4: Change of regression line by removing single data points when calculating*

We notice (from Table 4) that the data point with concentration 9 has about 12% contribution to the RSS. But leaving this data point out of the data set would merely change the slope or intercept of the regression line. In contrast to that, leaving out data point with concentration 4 would influence the regression line in a similar manner, although this data point only has a 4% contribution to the RSS. This is an example showing that a data point being "far away" from all other data points has a higher influence to the regression line than data points being close to the centre of the data set. This type of influence is called *leverage* and data points with high leverage are called *high-leverage points*. In method validation, it is important to notice that *high-leverage points* are most likely data points with very low (or high) concentrations, e.g. those close to the limit of quantification or detection (LOQ / LOD). Because of that, we have to carefully investigate the influence of these data points on the regression line.

Depending of the number of data points $n$, there is a limit for which a data point will likely be a *high-leverage data point*. In our example, this limit will be $2 * (2/n) = 0.363$, where $n = 11$. A synonym for *leverage* is *Hat Value* or *H Value*. A visual representation is given in the following Figure 5, indicating that in our example, there are no *high-leverage points*. The data points for the concentrations 4, 5, 13 and 14 are possessing high *leverage* values as they are "far away" from the other data points. But as their *leverage* values are below the calculated limit of 0.363, they aren't *high-leverage points*.

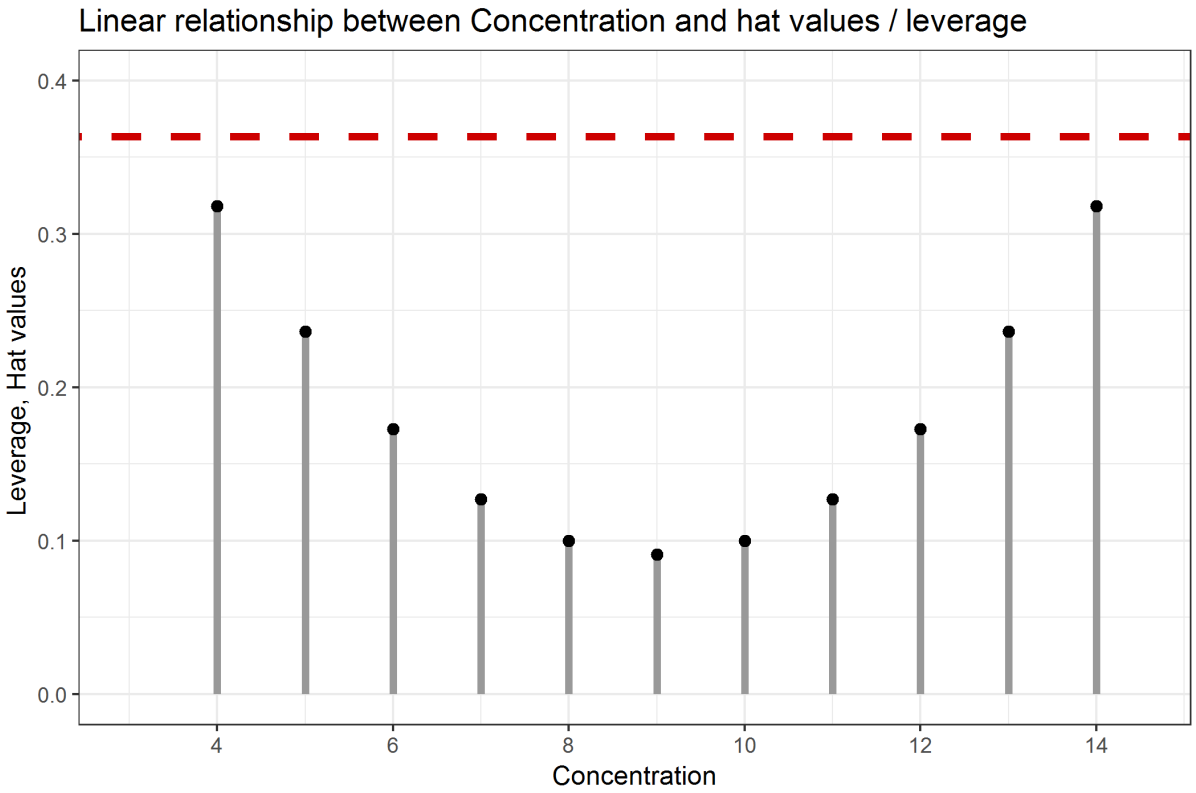


*Figure 5: Leverage of each data point, and relation to hat value limit of 0.363*

It is also important to clarify that data points with higher *leverage* don't imply that they have a big influence on the regression line because they are only dependent on the X-coordinate (or analyte concentration in this case). Y values (or signal values in our case) are of no importance for the calculation of the *leverage*. Thus it is possible that data points with low *leverage* values do possess a high influence on the regression line and hence the quality of the method. Therefore, to fully investigate the influence of a data point, we need to have an additional measure that also takes the Y values into account.

A measure that takes the X and Y coordinate for each data point into account is the so-called *Cook's Distance (D)* [2]. It relates each data point to all other data points, depending on the X and Y coordinates and is therefore giving a better measure for identifying so-called *influential observations*. These *D* values should be as small as possible and less than a limit which, for *D*, is calculated as $4/(n-2)$, with $n$ representing the number of data points. In Figure 6, all data points are represented as *Cook's Distance* and the limit here is $4/9 = 0.445$. Consequently, data point $X = 13$ is identified as an *influential observation*, meaning, it would change the slope and / or y-intercept too much when removed. But, when looking only at its *leverage* this data point would not be identified as unusual.

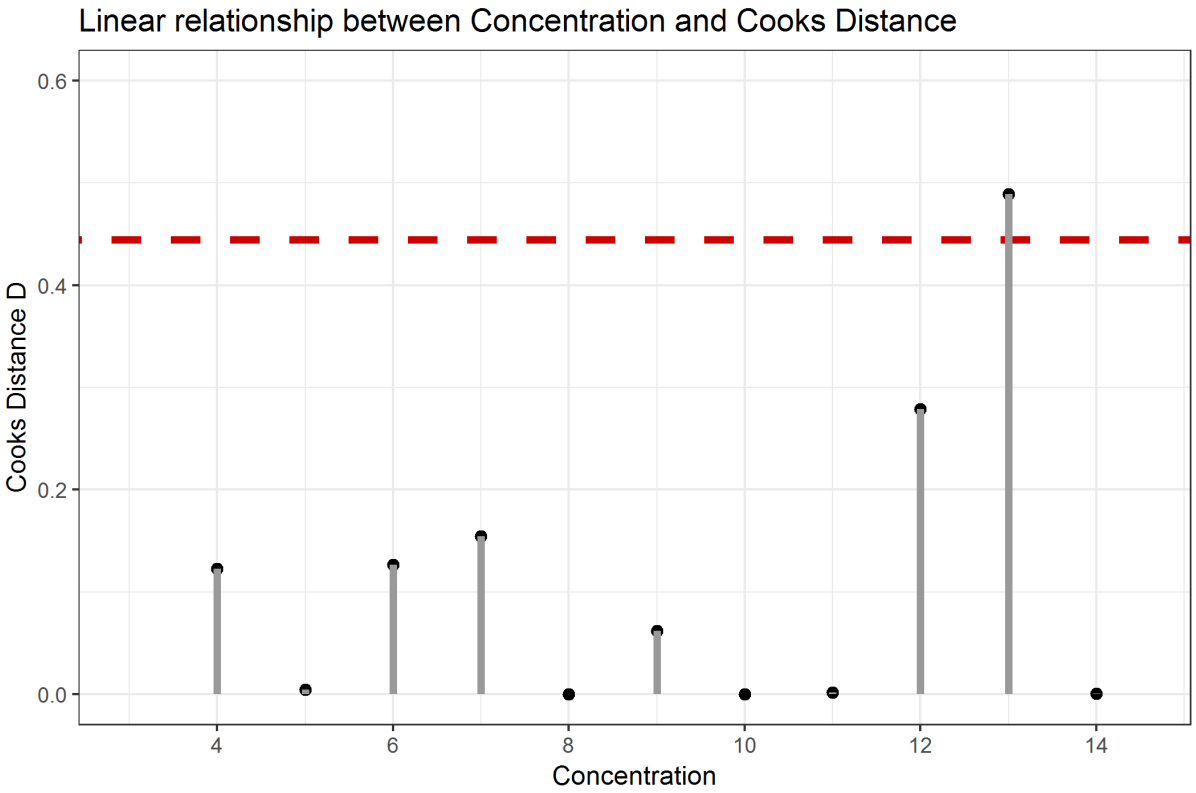


*Figure 6: Cook's Distance used to identify influential observations*

The values of *Cook's Distance* are a mixture of the *RSS* and the *leverage* values. They combine the characteristics of all single data points by investigating the Y and X coordinates in relation to all other data points. Only with this combined measure the actual influence of a single data point can be revealed. To include this knowledge into the standard scatter plot and regression line, one might color-code data points having a statistical influence onto the whole data set. A figure containing all information might look like Figure 7. Here, the data, the regression line, and the information about the influence of each data point, is shown. The threshold is 0.445 and it can be seen that data point 13 is indicated an influential data point (blue colour). Also, it should to be remembered, that this information couldn't be made visible by only focussing on the residual sum of squares. Even for a "low" *RSS* value, there still might be influential data points present in the data set.

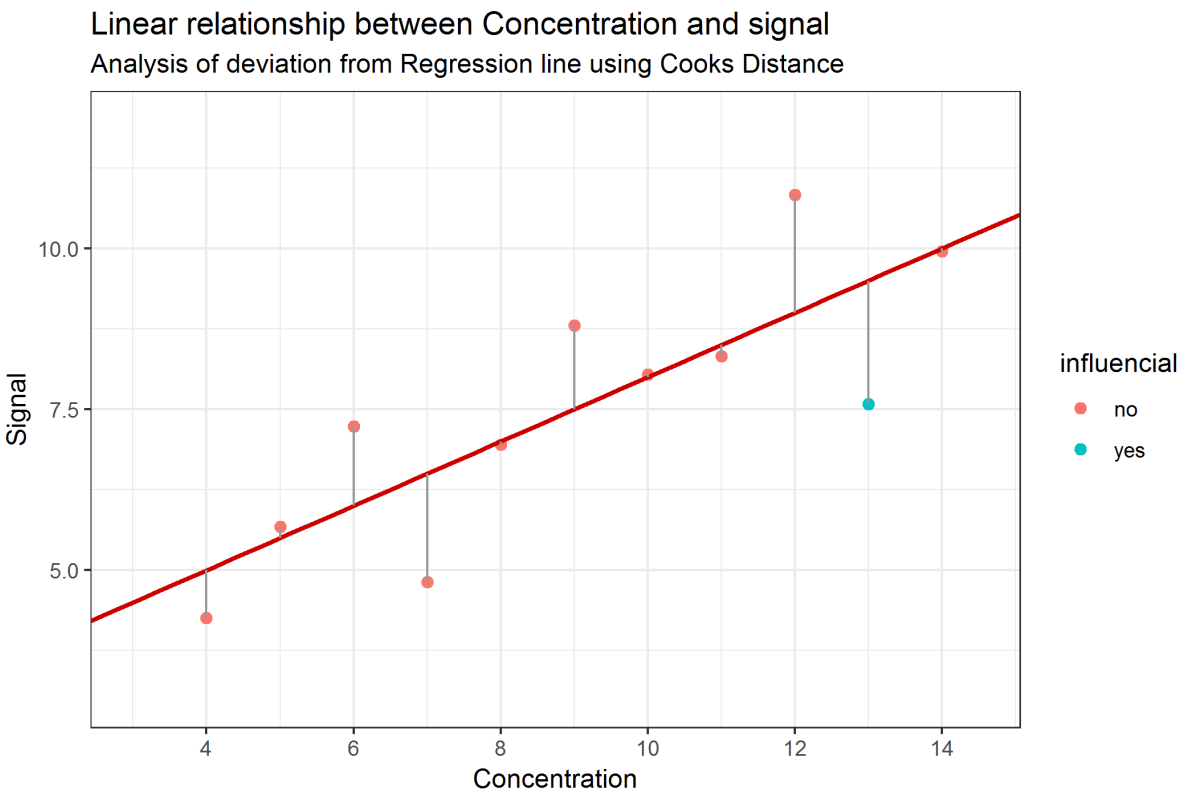


*Figure 7: Linear regression and influential data point identified by Cook's Distance*

## So, what to do when investigating linearity?

The ICH requirement of *„**analysis of the deviation of the actual data points from the regression line**"* can easily be assessed using programs like Excel or other software solutions. When it comes to presenting the residual sum of squares, one needs to be sure what the *RSS* means. Good scientific interpretation can only be possible when investigating each data point because each data point has its unique contribution to the overall *RSS*. Visual inspection of the data is required, but cannot provide scientific justification, therefore one should always perform further "residual" analysis – not only to satisfy ICH requirements, but also to gain knowledge about its own data.

The regression line, defined by slope and y-intercept reflects the general physico-chemical relation between concentration and signal. This relationship, based on the data, can be inaccurate and lead to wrong decisions, if influential data points are present and ignored. In that case, one should always check, why the influential data points behave differently than all the others. Also, influential data points are not necessarily outliers. Oftentimes, they reflect the natural behaviour and higher variability in certain concentration ranges.

## Conclusion

The *residual sum of squares* is a statistic value which is applied e.g. in linear regression. Its importance is often neglected, but when falsely interpreted it can lead to misunderstandings. In the same way that the average summarizes the properties of many data points in one value, the information about the individual values is lost. Regarding *RSS*, the individual contributions of the error squares are lost. Therefore, they must be separately analyzed to interpret the *RSS* correctly. Statistical key figures such as *hat values* / *leverage* or *Cook's Distance* are useful to classify the contributions of the individual values correctly and thus help in the proper interpretation of the *residual sum of squares*.

In analytical method validation, the *RSS* must be specified in addition to general key performance indicators. In the guideline, there is no information about an upper limit for the *RSS*, and the performance of the residual analysis doesn't seem to be mandatory as the word "*may*" is used. In obvious cases, a visual analysis of the residuals may be sufficient, but it is unable to identify potential influential data points. However, it is precisely these *influential observations* that may jeopardize the *fitness-for-purpose* of the method's linearity, as they might distort the established linear relationship between analyte concentration and signal. Visual analyzes are always subjective - "scientific justification", on the other hand, can only be obtained by objective criteria, such as *Cook's Distance* or similar methods for residual analysis.

When evaluating (e.g. linear) regression methods, the *residual sum of squares* constitutes the starting point of an interesting journey, ending in the knowledge about what the (linear) relationship is actually worth. For this reason, the residual sum of squares is not only of great importance in analytical method validation.

# RSS, leverage and Cook's Distance in Excel 2016

## Prerequisites

It is possible to calculate the discussed measures in Excel 2016: The example discussed above is taken from a publication of Francis Anscombe [3]. The publication contains four data sets, the well-known *Anscombe Quartet*. These data sets are identical in many linear regression characteristics, although consisting of completely different data points. Here, we work with the first of these data sets (Table 5).

*Table 5: Data set used in the example*

| Data point | X value | Y value |
|---|---|---|
| 1 | 4 | 4.26 |
| 2 | 5 | 5.68 |
| 3 | 6 | 7.24 |
| *4* | 7 | 4.82 |
| 5 | 8 | 6.95 |
| 6 | 9 | 8.81 |
| 7 | 10 | 8.04 |
| 8 | 11 | 8.33 |
| 9 | 12 | 10.84 |
| 10 | 13 | 7.58 |
| 11 | 14 | 9.96 |

When transferring these data into Excel, we here assume that the X values (e.g. analyte concentration) and the Y values (signal) will be stored in columns A and B. For analysis, first, we click onto "**Data**" and "**Data Analysis**". There will be a new window from which we select "**Regression**". Then, we enter the X and Y values cell ranges from the Excel sheet, e.g. columns A and B. It is important to activate the **Residuals** button, then we confirm by clicking "**OK**". A new sheet is opening, where we will find all the information that we need for further analysis. These are stored in the following cells (Table 6):

*Table 6: Characteristics in Excel for linear regression analysis*

| | Cell | Value | Formula / Symbol |
|---|---|---|---|
| **Multiple Correlation coefficient** | B4 | 0.816 | $r$ |
| **Coefficient of Determination** | B5 | 0.666 | $R^2$ $(R^2 = r * r)$ |
| **Number of data points** | B8 | 11 | $n$ |
| **RSS (residual sum of squares)** | B14 | 13.762 | $RSS = \sum_i \varepsilon_i^2 = (y_i - \hat{y}_i)^2$ |
| **y-intercept** | B17 | 3.000 | $Y$ value for $X = 0$, $\beta_0$ |
| **Slope** | B18 | 0.500 | $\beta_1$ |
| **Estimates for Y** | B25:B35 | | $\hat{y}$ |
| **Residuals** | C25:C35 | | $\varepsilon = y - \hat{y}$ |

Since we will need the raw data and the values for the residuals for future calculations, we will merge them into one sheet. I merged the raw data into the new sheet, resulting in having the data point number in column A, the Y estimates in column B, the residuals in column C, and then analyte concentration and signal in columns D and E, they are shown from row 25 to 35 (Figure 8):

| 22 | OUTPUT: | | | | |
|---|---|---|---|---|---|
| 23 | | | | | |
| 24 | *data point number* | **Y Estimate** | **Residual** | **Concentration** | **Signal** |
| 25 | 1 | 5,000 | -0,740 | 4 | 4,26 |
| 26 | 2 | 5,501 | 0,179 | 5 | 5,68 |
| 27 | 3 | 6,001 | 1,239 | 6 | 7,24 |
| 28 | 4 | 6,501 | -1,681 | 7 | 4,82 |
| 29 | 5 | 7,001 | -0,051 | 8 | 6,95 |
| 30 | 6 | 7,501 | 1,309 | 9 | 8,81 |
| 31 | 7 | 8,001 | 0,039 | 10 | 8,04 |
| 32 | 8 | 8,501 | -0,171 | 11 | 8,33 |
| 33 | 9 | 9,001 | 1,839 | 12 | 10,84 |
| 34 | 10 | 9,501 | -1,921 | 13 | 7,58 |
| 35 | 11 | 10,001 | -0,041 | 14 | 9,96 |

*Figure 8: Data preparation: data for linear regression and residual analysis*

To calculate the values for C*ook's Distance D* for all data points *i,* $D_i$, we need the formula: $D_i = \frac{isr_i^2}{p} * \frac{h_i}{1-h_i}$. In order to do so, we need to calculate two more things, namely the *leverage* $h_i$ and squares of the $isr$ values. These can be calculated from the r*esiduals*, which we already have.

## Calculation of *Residual Sum of Squares*

The *residuals* for each data point can be calculated in the sheet be subtracting the Y estimates from the signal values, or E25 - B25 (for the first data point). The *RSS* then is the sum of all the squared residuals (E25 – B25)^2. Here, I put the residuals into column F (cells F25:F35) and the squared residuals into column G (cells G25:G35), see Figure 9. The value of the sum, or *RSS*, I will save in cell F37, its value will be the same as in cell B14, which is part of the data analysis output.

## Calculation of *leverage* or $H$ values

For further calculations, we need the average values (the mean; $\bar{x}$), as well as the standard deviation ($s_x$) of all x values (concentration). The average of x we will get using AVERAGE(D25:D35), and we save it in cell D37. The standard deviation of x we will get using STDEV.S(D25:D35) and we save it in cell D38. We will need those for the following steps, the calculation of *leverage* respectively *hat values*:

The formula for *hat value* calculation is $h_i = \frac{1}{n} + \frac{1}{n-1} * \left(\frac{x_i - \bar{x}}{s_x}\right)^2$. Using the currently calculated values for $\bar{x}$ and $s_x$, we use the following Excel formula to calculate the *hat value* for the first data point (D25): (1/$B$8) + 1/($B$8-1) * (($D25 - $D$37) / $D$38)^2.

Again, as stated above, we recognise that *leverage* is just depending on the x values (concentration, column D) of the data and on its relation to the other values (indicated with cells D37 and D38), but not on the y values (signal).

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 22 | OUTPUT: | | | | | (1/$B$8)+1/($B$8-1)*(($D25-$D$37)/$D$38)^2 | | |
| 23 | | | | | | | | |
| 24 | *data point number* | **Y Estimate** | **Residual** | **Concentration** | **Signal** | **residuals** | **squared residuals** | **leverage** |
| 25 | 1 | 5,000 | -0,740 | 4 | 4,26 | -0,740 | 0,548 | 0,318 |
| 26 | 2 | 5,501 | 0,179 | 5 | 5,68 | 0,179 | 0,032 | 0,236 |
| 27 | 3 | 6,001 | 1,239 | 6 | 7,24 | 1,239 | 1,536 | 0,173 |
| 28 | 4 | 6,501 | -1,681 | 7 | 4,82 | -1,681 | 2,825 | 0,127 |
| 29 | 5 | 7,001 | -0,051 | 8 | 6,95 | -0,051 | 0,003 | 0,100 |
| 30 | 6 | 7,501 | 1,309 | 9 | 8,81 | 1,309 | 1,714 | 0,091 |
| 31 | 7 | 8,001 | 0,039 | 10 | 8,04 | 0,039 | 0,002 | 0,100 |
| 32 | 8 | 8,501 | -0,171 | 11 | 8,33 | -0,171 | 0,029 | 0,127 |
| 33 | 9 | 9,001 | 1,839 | 12 | 10,84 | 1,839 | 3,381 | 0,173 |
| 34 | 10 | 9,501 | -1,921 | 13 | 7,58 | -1,921 | 3,691 | 0,236 |
| 35 | 11 | 10,001 | -0,041 | 14 | 9,96 | -0,041 | 0,002 | 0,318 |
| 36 | | | | | | | | |
| 37 | AVERAGE(D25:D35) | | **Mean X** | 9 | | **RSS** | 13,763 | |
| 38 | STDEV.S(D25:D35) | | **SD X** | 3,317 | | | | |

*Figure 9: Calculation of leverage*

## Calculation of *Cook's Distance* / *D* values

Now, we use the *leverage* to calculate the *Cook's Distance*. For that, we first calculate the $isr$ value for each data point: $isr_i = \frac{\varepsilon_i}{s_E\sqrt{1-h_i}}$. $\varepsilon_i$ are the *residuals*, which we already have calculated in column C and F. We saved the *leverage* values $h_i$ in column H. Value $s_E$ can be calculated with SQRT(($B$8-1)/($B$8-2)*STDEV.S(F25:F35)^2), we save it in cell D39. Therefore, the $isr$ for the first data point (row 25) is: F25/$D$39/SQRT(1-$H25), we save all $isr$ values in column I.

With that, we calculate *Cook's Distance* $D$. For data point $i$, we calculate $D_i$ as follows: $D_i = \frac{isr_i^2}{p} * \frac{h_i}{1-h_i}$ and for data point 1 (row 25), we have: $I25^2 / 2 * $H25 / (1-$H25). We save all values in column J. That is everything we need to calculate the *Cook's Distance* in Excel, thus being able to provide all information; not only regression analysis but also residual analysis.

The following Figure 10 shows a summary of all the steps taken:

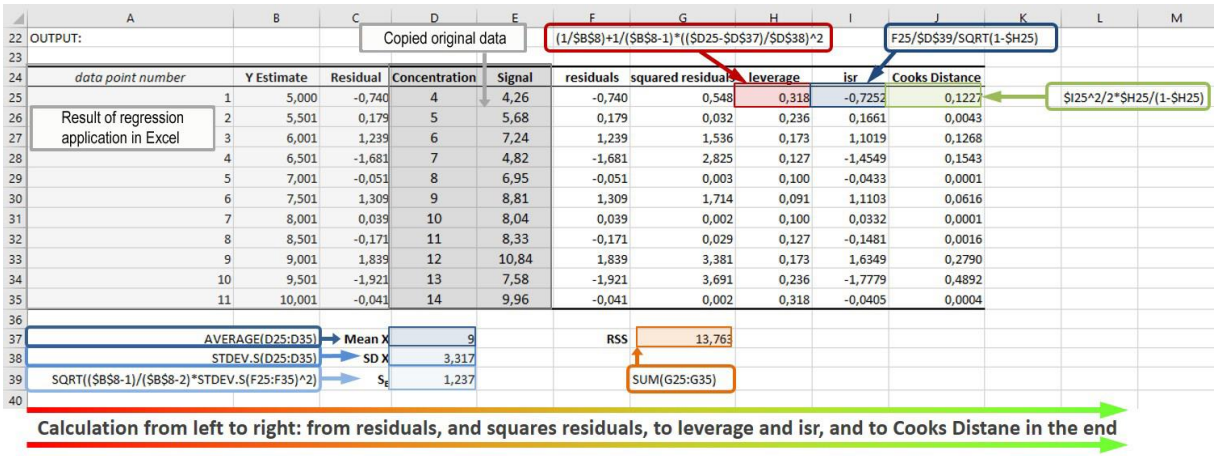


*Figure 10: Overview over all steps needed to calculate leverage and Cook's Distance for a given data set based on linear regression in Excel 2016.*

# Sources


[1] **Q2(R1) Validation of Analytical Procedures: Text and Methodology:** http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q2_R1/Step4/Q2_R1__Guideline.pdf

[2] **Cook's Distance:** Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". Technometrics. American Statistical Association. 19 (1): 15–18.

[3] **Anscombes Quartett:** F. J. Anscombe: *Graphs in Statistical Analysis*. In: *American Statistician*. 27, No. 1, 1973, page 17–21.


# About the author


**Dr. Peter P. Heym** wrote this article as guest author for Loesungsfabrik. He studied bioinformatics and obtained his PhD at the Leibnitz Institute for Plant Biochemistry Halle with the topic "In silico characterization of AtPARP1 and virtual screening for AtPARP inhibitors to increase resistance to abiotic stress". He is the CEO of Sum Of Squares - Statistical Consulting (www.sumofsquares.com), a service company specializing in statistical consulting for students, individuals, and companies. In addition to statistical advice he offers support for university theses, evaluation of surveys, workshops (e.g. in programming language R), seminars, and trainings, also in GMP topics.