

# Data Analysis

www.crgraph.com

## Multiple Regression

### Contents

Keywords .....	2
Introduction .....	2
Purpose and benefit .....	2
Basics.....	2
Analyses of Variance (Model ANOVA) .....	5
Prediction Measure $Q^2$ .....	6
$R^2$ and $Q^2$ is small.....	6
$R^2$ high and $Q^2$ very small.....	7
Lack of Fit.....	7
Analyses of Variance overview.....	7
Reproducibility.....	7
Test of the coefficient of determination.....	8
Test of the regression coefficients, the p-value .....	8
Standard deviation of the model RMS.....	9
Confidence interval for the regression coefficient.....	9
Confidence interval for the response.....	9
Standardize to -1 ... +1 .....	10
Standardize to standard deviation .....	10
The correlation matrix.....	10
Response transformation (Box-Cox) .....	11
Application in Visual-XSel .....	13

### Requirement and related topics

Basics of statistics are beneficial for these descriptions. Further and related topics include:

[www.weibull.de/COM/Statistics.pdf](http://www.weibull.de/COM/Statistics.pdf)

[www.weibull.de/COM/Poisson\\_Regression.pdf](http://www.weibull.de/COM/Poisson_Regression.pdf)

[www.weibull.de/COM/Design\\_of\\_Experiment.pdf](http://www.weibull.de/COM/Design_of_Experiment.pdf)

## Keywords:

Data analysis, evaluation, Multiple Regression, stepwise regression, regression, linear regression model, DoE, least square method, ANOVA, Effect, Lack of fit, Box-Cox, categorical factors, prediction measure, reproducibility, confidence, regression coefficients, p-value

## Introduction

In a regression, a connection is made between the influencing parameter(s) and a target variable. For a parameter  $x$ , a regression line is used for this  $y = b_0 + b_1 x$ .

A multiple regression expands the relationships to several influencing parameters. The magnitude of the influences is determined using the least squares error method.

Multiple regression is the standard method for evaluating experimental designs (DoE), but also for general data evaluation.

## Purpose and benefit

The aim is to determine the coefficients (intensities) and possible interactions. The model equation then obtained can be used to make predictions and optimization runs.

## Basics

One uses a multiple regression if more than one independent factor  $x$  is available. The simple linear model is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

It is presupposed that the features are normal distributed and linear. E.g., not linear parameters can be realized in most cases by re-modelling or by using squared terms:

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + \dots$$

In case of tabular values this means that one adds the column to  $x$  with the values in a new column copied and squared. E.g., a combination two influences which represents an interaction also can be carried out:

$$y = b_0 + b_1 x_1 + b_2 x_1 x_2 + b_3 x_2 + \dots$$

The corresponding table columns for  $x$  then have to be inserted in a new column as a product  $x^2$ . Further conversions are possible to reach the linear model. In matrix form the model equation is:

$$\hat{y} = b^T X$$

with  $\hat{y}$  = vector of the results from the parameter set  
 $X$  = matrix of the actual parameter values  
 $b$  = vector of the coefficients

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_z \end{bmatrix}$$

Hint: 1st column represents in  $X$  the constant

The sought-after vector  $b$  with the coefficients determines about the matrix operation

$$b = (X^T X)^{-1} X^T y$$

Example: Interaction model is given:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2$$

The individual steps of the equation

$$b = (X^T X)^{-1} X^T y \quad \text{arise as follows}$$

$$X' = X^T X \quad \text{with} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix} \quad z+1$$

columns and  $n$  rows

The respective cells are calculated after each other:

$$x'_{j,i} = \sum_{k=1}^n x_{k,i}^{(T)} x_{j,k} \quad (1\text{st index} = \text{column}, 2\text{nd index} = \text{row})$$

The first column represents the constant  $b_0$ . The following columns are the factors  $x_1$  and  $x_2$  and the last column is the product of  $x_1$  and  $x_2$  (interaction).

experiment: results Y

$V_1$	-1	-1	3
$V_2$	1	-1	5
$V_3$	-1	1	7
$V_4$	1	1	11
$V_5$	0	0	6

$$X = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & 0 \\ -1 & -1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 \end{bmatrix}$$

etc. cells

$$j=1 \quad i=1 \\ x'_{1,1} = (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) + (1) \cdot (1) = 5$$

$$j=2 \quad i=2 \\ x'_{2,2} = (-1) \cdot (-1) + (1) \cdot (1) + (-1) \cdot (-1) + (1) \cdot (1) + (0) \cdot (0) = 4$$

as a result, yields:

$$X' = X^T X = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

and the revers matrix is:

$$(X^T X)^{-1} = \begin{bmatrix} 1/5 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$$

and via the intermediate step

$$X^T y = \begin{bmatrix} 32 \\ 6 \\ 10 \\ 2 \end{bmatrix}$$

one gets the result for the sought-after coefficients:

$$b = (X^T X)^{-1} X^T y = \begin{bmatrix} 6,4 \\ 1,5 \\ 2,5 \\ 0,5 \end{bmatrix}$$

So, the equation of the beginning is:

$$y = 6,4 + 1,5x_1 + 2,5x_2 + 0,5x_1x_2$$

## Categorical Factors

Categorical or qualitative factors whose variations are indicated in the form of textual names must be brought in suitable number form. One uses -1 and +1 for two attitudes in a column. If the categorical factor is e.g., a component of supplier A and supplier B, then A gets the value -1 and B the value 1. As of every broader feature (variation) an additional column is laid out:

	F [B]	F [C]	F [D]
A	-1	-1	-1
B	1	0	0
C	0	1	0
D	0	0	1

The attitude A of the generally mentioned factor F represents the basic level. The corresponding line therefore contains -1 everywhere. The other variations have one in their column 1.

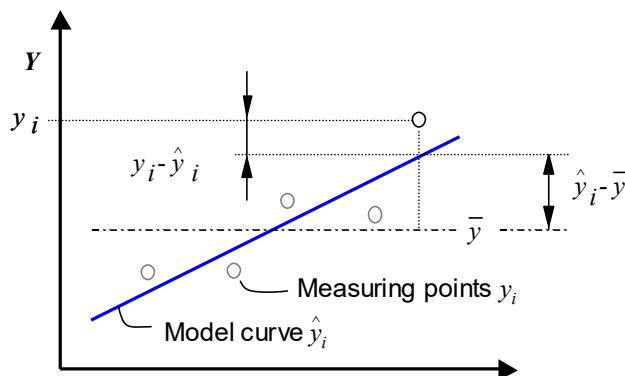
Partial correlations of  $r$  have construction caused test plans with categorical factors  $r = 0.5$  or more greatly.

## Analyses of Variance (Model ANOVA)

For assessment of the regression model the most important index is the coefficient of determination  $R^2$  and then adjusted coefficient of determination  $R^2_{adj}$ .

The closer  $R^2$  is to the value 1, the better the model  $y$  is described through  $x$ . The smaller  $R^2$  is the values scatter is higher and there is not the slightest connection to  $y$ .

The following picture shows the connection between measuring and the model for one factor



$$SS_{Total} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad SS_{Reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad SS_{Res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$SS_{Total} = SS_{Reg} + SS_{Res}$$

$$R^2 = \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Res}}{SS_{Total}} \quad 0 \leq R^2 \leq 1$$

One frequently also finds the adjusted coefficient of determination  $R^2_{adj}$ . The corresponding degrees of freedom are taken into account

$$R^2_{adj} = 1 - \frac{SS_{Res} / DF_{Res}}{SS_{Total} / DF_{Total}} = 1 - \frac{MS_{Res}}{MS_{Total}}$$

$MS$  : Variance

$DF_{Reg}$  : Degrees of Freedom of regression -> number of X-variables in model  $DF_{Reg} = z - 1$   
( $z$  = number of model-terms  $x_1, x_2, x_3, x_1 \cdot x_2, x_1^2, \dots$ )

$DF_{Res}$  : Degrees of Freedom of the residuals  $DF_{Res} = n - z - 1$   
( $n$  = Number of experiments)

$DF_{Total}$  : Degrees of Freedom total  $DF_{Total} = n$

For great data sizes are like A and B brought closer. The smaller the data size gets, the bigger the deviation is.  $R^2$  overestimates the declared amount of deviation considerably at a small number of degrees of freedom from time to time. Great differences between  $R^2$  and  $R^2_{adj}$  indicate unnecessary terms in the model.

## Prediction Measure $Q^2$

The Prediction measure is the fraction of variation of the response that can be predicted by the model.

In principle  $R^2$  rises with increasing the coefficients in the model because these then can adapt to the test points always better ( $SS_{res}$  decreases).  $R^2$  isn't suitable to recognize whether the model is over-determined. For this the  $Q^2$  measure has been defined:

$$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

with  $\hat{y}_i$  = model prediction for not measured points

$Q^2$  also can get negative if the point is bigger than the denominator.

Hints:

## $R^2$ and $Q^2$ is small

The customization of the model is bad. This can have several causes:

- Outliers
- Wrong test order
- Bad reproducibility

Corrective: Checking the measurements for plausibility. Perhaps carrying out the tests once again.

Bad test plan, possible carry out a new plan for one.

## R<sup>2</sup> high and Q<sup>2</sup> very small

The model offers a good description, is, however, unstable. Tendency toward the over-determination

There are too many terms or interactions taken into account. The model should be reduced. The terms with the smallest effects should be deleted from the model, but be careful with significant interactions.

- There are dominant outliers
- One response must be transformed
- The investigations should be going on

Note:

- In case of lean experiments (screening plans), often the  $Q^2$  is worse than the model is.
- In case of many repetitions, the  $Q^2$  is better than the model is. Therefore, it should be analyzed much more the lack of fit.

## Lack of Fit

Some further information can be analysed from the residual.  $SS_{res}$  is put together out:

$$SS_{res} = SS_{LoF} + SS_{p.e.}$$

$SS_{LoF}$  is the Lack of Fit, with the degrees of Freedom  $DF_{LoF} = n - z - DF_{p.e.} - 1$   
 $SS_{p.e.}$  is the pure error determined from repetitions.

$$SS_{p.e.} = \sum_{j=1}^r \sum_{k=1}^{r_j} (Y_{j,k} - \bar{Y}_j)^2 \quad \text{with the Degrees of Freedom} \quad DF_{p.e.} = \sum_{j=1}^r (r_j - 1)$$

Is  $SS_{res}$  and  $SS_{p.e.}$  known, the equation for the Lack of Fit is:

$$SS_{LoF} = SS_{res} - SS_{p.e.}$$

The quotient of the variances is then the Lack of Fit:

$$\frac{MS_{LoF}}{MS_{p.e.}} = \frac{SS_{LoF} / DF_{LoF}}{SS_{p.e.} / DF_{p.e.}} > F_{DF_{LoF}, DF_{p.e.}, \gamma}$$

The result is to compare to a critical F-worth ( $\gamma$ =confidence interval). Obviously if this is bigger than the model terms are contained too little.

## Analyses of Variance overview

The following picture shows an overview to the total Analyses of Variance:

## Reproducibility

The Reproducibility is described through the following equation:

$$\text{Reproducibility} = 1 - \frac{MS_{p.e.}}{MS_{total}}$$

This is a relative indicator which says as good we are able to reproduce the tests. This indicator can only be determined with repetitions of tests.

## Test of the coefficient of determination

As you described at the beginning is the regression result all the better the nearer the coefficient of determination is due to 1. The question is worth as of which value under 1 the deviation by chance or already is only significant. To this one builds the null hypothesis: All regression coefficients are 0., i.e., no connection between y and x etc. insists. A weighted F value is calculated as test quantity:

$$F_{pr} = \frac{R^2(n-z-1)}{(1-R^2)z}$$

with  $n$  number of series of experiments = and  $z$  = number of model terms  $x_1, x_2, x_3, x_1, x_2, x_1^2$  etc.. As the result is significantly the regulation becomes the F-distribution with the degrees of freedom to

$$f1 = z, \quad f2 = n - z - 1$$

used. According to the significance standard, e.g. 5% or 1%, the regression result is all the better with respect to the correlation coefficient, the nearer the value of the F-distribution is due to 0 and the null hypothesis must be rejected.

The corresponding statistical basics you find in the statistical-literature.

## Test of the regression coefficients, the p-value

To determine the significance of a factor, frequently the so-called p-value is used. At first the hypothesis is defined that a coefficient of a factor  $b=0$ . Then the p-value is the probability to reject the hypothesis mistakenly. This probability is determined via the t-distribution:

$$t = \frac{b}{s_b}$$

$b$  = coefficient from the multiple regression

$s_b$  = deviation of the coefficient

With using the double value of  $t$  because of the two-way test and the degrees of freedom  $f = n - z - 1$  ( $n$  = count of experiments,  $z$  = count of model terms  $x_1, x_2, x_3, x_1 \cdot x_2, x_1^2$  etc.). With the index  $j$  for each factor  $t$  is defined with:

$$t_j = \frac{b_j}{s_{b_j}}$$

The spread of the regression coefficient is determined through:



$$s_{b_j} = \sqrt{s^2 X''_{j,j}}$$

in which  $s$  is the standard deviation of the complete model.  $s$  is calculated through the sum of squares between the model and the measured values

$$s^2 = \frac{1}{n - z - 1} \sum_{i=1}^n \left( Y_i - b_o - \sum_{j=1}^z x_{j,i} b_j \right)^2$$

with  $b_o$  = constant term of the model.

$X''$  is calculated through:

$$X'' = (X^T X)^{-1} \quad \text{with} \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{z1} \\ 1 & x_{12} & \dots & x_{z2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{zn} \end{bmatrix}$$

The greater the  $t$ -worth is the smaller the p-value becomes. Usually the significance level is 5%, that means if there is a p-value smaller than 0.05 the coefficient is significant.

## Standard deviation of the model RMS

The so-called RMS-Error (Root mean squared error) represents the standard deviation of the complete model. It is calculated through:

$$RMS = \sqrt{\frac{SS_{Res}}{n - z - 1}} \quad \text{with} \quad SS_{Res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The relative standard deviation is related to the middle data area

$$RMS / Y_m$$

and is a further control criterion. This value can also analogously be seen by Taguchi to the reciprocal of the not squared signal-to-noise ratio (without the pre-factor 10 lied)

## Confidence interval for the regression coefficient

The confidence interval for the regression coefficient is determined with the spread already introduced above:

$$b_j \pm \sqrt{s^2 X''_{j,j}} t_{n-z-1; 1-\gamma/2}$$

## Confidence interval for the response

For certain values of the factors (adjusting) the response value can be calculated to  $Y$  about the model equation (forecast). The corresponding value has a confidence interval because of the spread of the tests and because of the simplification of the model to the reality. This can be decided on the following relation:

$$\hat{Y} \pm \sqrt{s^2 x^T X^{-1} x} \quad t_{n-z-1; 1-\gamma/2}$$

with  $X^{-1} = (X^T X)^{-1}$  (see above) and  $x$  for the corresponding factor adjustments

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_z \end{pmatrix}$$

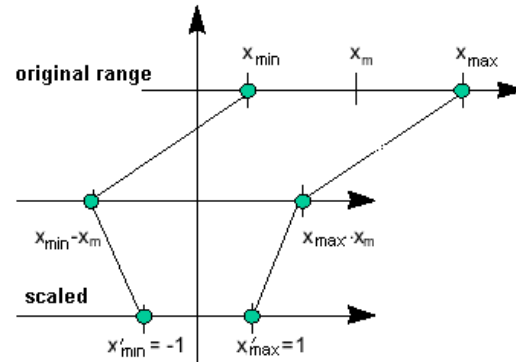
and  $\gamma$  for the confidence level, normally 5%.

This form is valid under this one assumption that one parameter each are changed, the others however are fixed values (principle as in the case of the effect chart -> non simultaneous confidence interval).

## Standardize to -1 ... +1

All data are transformed that the range is between -1 and 1.

$$x_n = \frac{(x - \bar{x})}{(x_{\max} - x_{\min})}$$



Through this one get a better comparable and relative influence sizes under each other. In addition, the multiple regression is circumstances permitting only hereby possible when the data areas lie far from each other.

## Standardize to standard deviation

At the standardized form the data values are related and put centrally to her standard deviation:

$$x_s = \frac{(x - \bar{x})}{s}$$

The standardization should be used at historical data or tests not planned since the data values can happen uneven regarding her size (not orthogonal).

## The correlation matrix

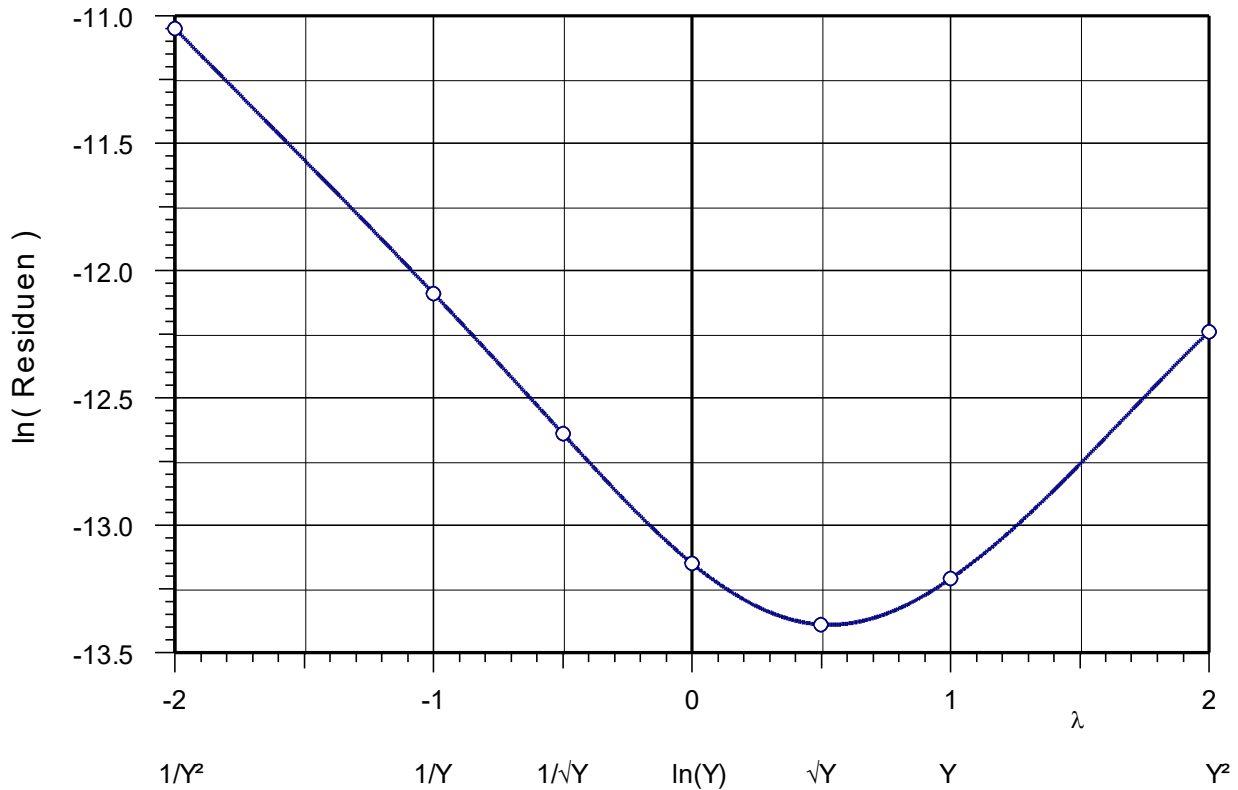
One understands by a correlation a more or less high linear dependence between two variables. The correlation between two factors or between  $x$  and  $y$  is defined through:

$$r_{xy} = \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

If there is a strong correlation between two  $x$  factors, in most cases one of both can be left out.

## Response transformation (Box-Cox)

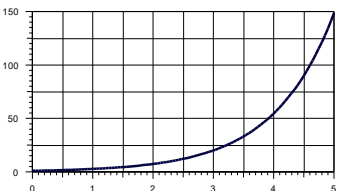
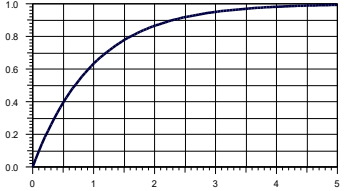
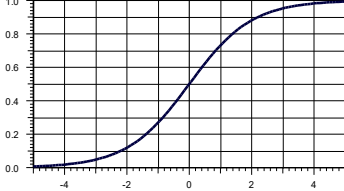
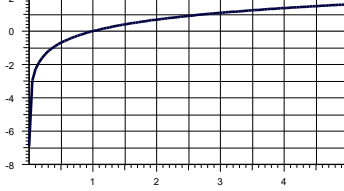
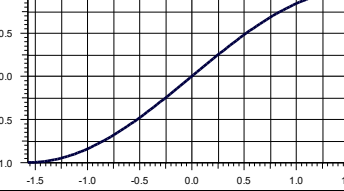
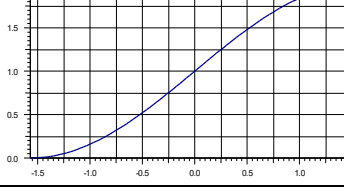
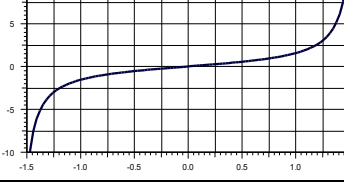
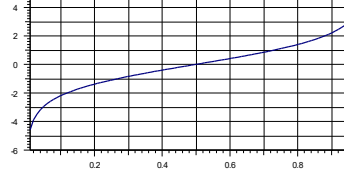
For checking a possibly necessary response transformation the so-called **Box-Cox**-transformation is used.



One after another the response is transformed according to the functions displayed below and the residues (SSr) are determined.

$$Y^{(\lambda)} = \begin{cases} \lambda^{-1} \bar{Y}^{1-\lambda} (Y^\lambda - 1) & \text{if } \lambda \neq 0 \\ \bar{Y} \ln(Y) & \text{if } \lambda = 0 \end{cases}$$

The smaller the residues and therefore the deviations from the model to the measured data, the better is the transformation to be chosen. This has to be adjusted under the category data, as mentioned in the beginning. It must be pointed out that after the transformation single significances can be changed. Therefore, on the side coefficients it has to be checked, if the model has to be corrected. The Box-Cox-transformation can just be executed, if a target factor-transformation has not yet been chosen.

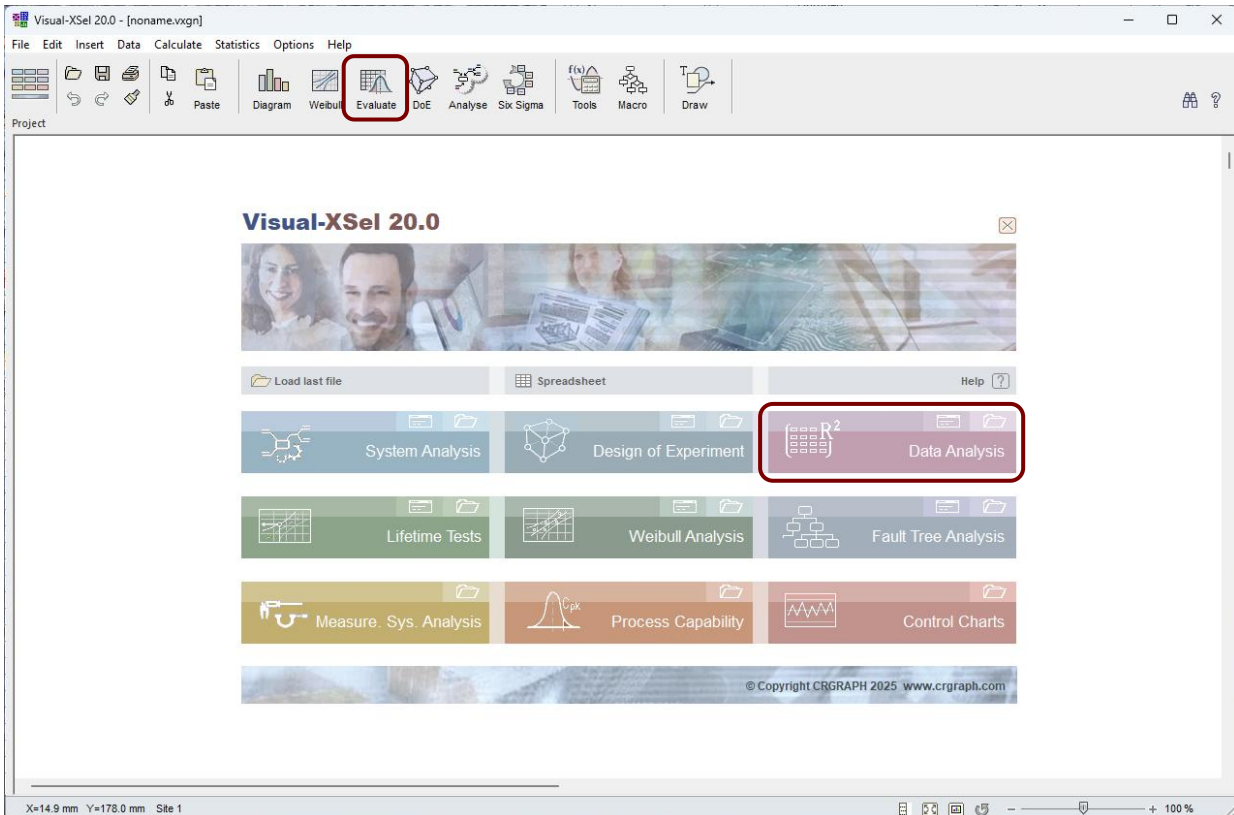
	Transformation	Inverse function	Example for a'=1, b'=1 c'=0
1	$Y' = a' e^{b'Y} + c'$	$Y = \frac{1}{b'} \ln \left( \frac{Y' - c'}{a'} \right)$	 A graph showing the exponential function Y' = e^Y. The x-axis (Y) ranges from 0 to 5, and the y-axis (Y') ranges from 0 to 150. The curve starts at (0, 1) and increases rapidly.
2	$Y' = a'(1 - e^{-b'Y}) + c'$	$Y = \frac{1}{b'} \ln \left( \frac{1}{1 - (Y' - c')/a'} \right)$	 A graph showing the function Y' = 1 - e^{-Y}. The x-axis (Y) ranges from 0 to 5, and the y-axis (Y') ranges from 0.0 to 1.0. The curve starts at (0, 0) and increases, approaching 1.0 as Y increases.
3	$Y' = a' \left( 1 - \frac{b'}{e^{c'Y} + 1} \right)$	$Y = \frac{1}{c'} \ln \left( \frac{b'}{1 - Y'/a'} - 1 \right)$	 A graph showing the function Y' = 1 - 1/(e^Y + 1). The x-axis (Y) ranges from -4 to 4, and the y-axis (Y') ranges from 0.0 to 1.0. The curve is an S-shape, passing through (0, 0.5).
4	$Y' = a' \ln(b'Y + c')$	$Y = \frac{1}{b'} \left( e^{\left( \frac{Y'}{a'} \right)} - c' \right)$	 A graph showing the function Y' = ln(Y). The x-axis (Y) ranges from 1 to 5, and the y-axis (Y') ranges from -8 to 2. The curve starts at (1, 0) and increases slowly.
5	$Y' = a' \sin(b'Y + c')$	$Y = \frac{1}{b'} \left( \text{ArcSin} \left( \frac{Y'}{a'} \right) - c' \right)$	 A graph showing the function Y' = sin(Y). The x-axis (Y) ranges from -1.5 to 1.5, and the y-axis (Y') ranges from -1.0 to 1.0. The curve is a sine wave passing through the origin.
6	$Y' = a'(1 + \sin(b'Y + c'))$	$Y = \frac{1}{b'} \left( \text{ArcSin} \left( \frac{Y'}{a'} - 1 \right) - c' \right)$	 A graph showing the function Y' = 1 + sin(Y). The x-axis (Y) ranges from -1.5 to 1.5, and the y-axis (Y') ranges from 0.0 to 2.0. The curve is a sine wave shifted upwards, passing through (0, 1).
7	$Y' = a' \tan(b'Y + c')$	$Y = \frac{1}{b'} \left( \text{ArcTan} \left( \frac{Y'}{a'} \right) - c' \right)$	 A graph showing the function Y' = tan(Y). The x-axis (Y) ranges from -1.5 to 1.5, and the y-axis (Y') ranges from -10 to 10. The curve has vertical asymptotes at Y = +/- pi/2.
8	$Y' = \ln \left( \frac{Y}{1 - Y} \right)$	$Y = \frac{1}{1 + e^{-Y'}}$	 A graph showing the function Y = 1/(1 + e^{-Y'}). The x-axis (Y') ranges from -4 to 4, and the y-axis (Y) ranges from 0 to 1. The curve is an S-shape, passing through (0, 0.5).



## Application in Visual-XSel

www.crgraph.com

Our software Visual-XSel is a powerful tool for all important statistical quality and reliability methods. To get started, use the topic areas in the guide (see also [crgraph.de/en/themen-index](http://crgraph.de/en/themen-index)), or the icon *Evaluation*.



Here you can find an introduction and a short video:

[crgraph.de/visual-xsel-software/](http://crgraph.de/visual-xsel-software/)

Here you can also find some introductory videos:

[crgraph.de/downloads/software/Visual-XSel Basis Functions.mp4](http://crgraph.de/downloads/software/Visual-XSel_Basis_Functions.mp4)

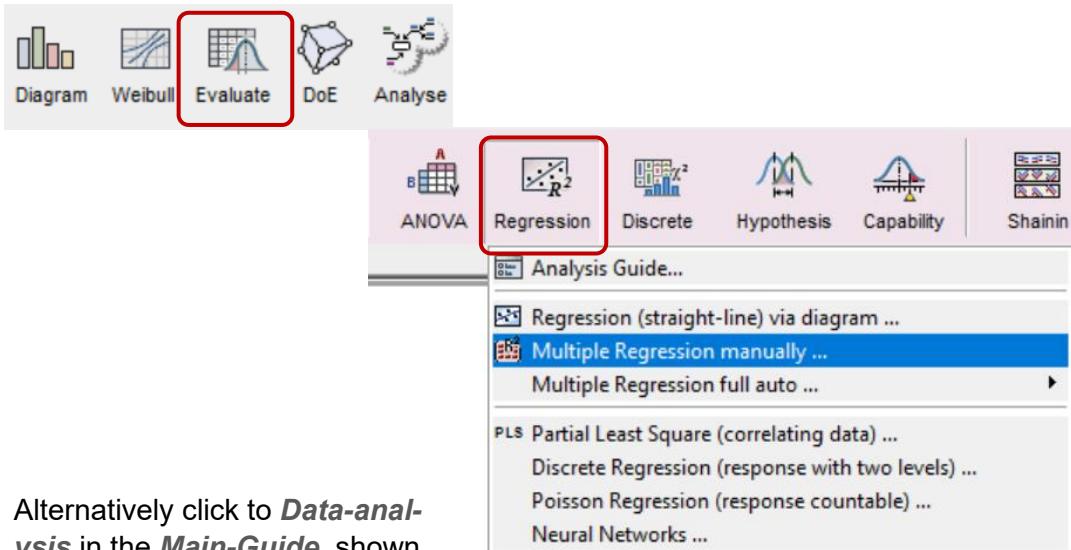
[crgraph.de/downloads/software/Visual-XSel Methods.mp4](http://crgraph.de/downloads/software/Visual-XSel_Methods.mp4)

It is not for nothing that this software is used in many well-known companies:

[References](#)

The following description is a guide and introduction to the topic shown

By using the icon **Evaluate** an icon bar appears with the most important data analysis tool **Regression**.

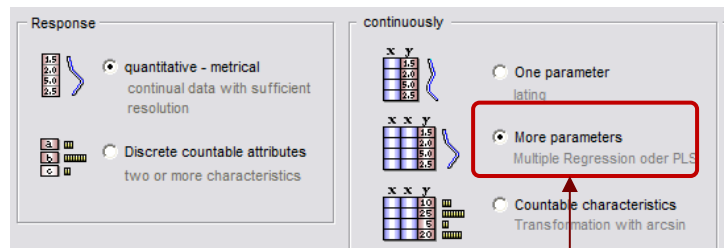


Alternatively click to **Data-analysis** in the **Main-Guide**, shown at the first page and follow the speech bubbles.



The further steps are explained using the sample data *Example\_MulReg.vxt*, which is available via menu **File/Examples**.

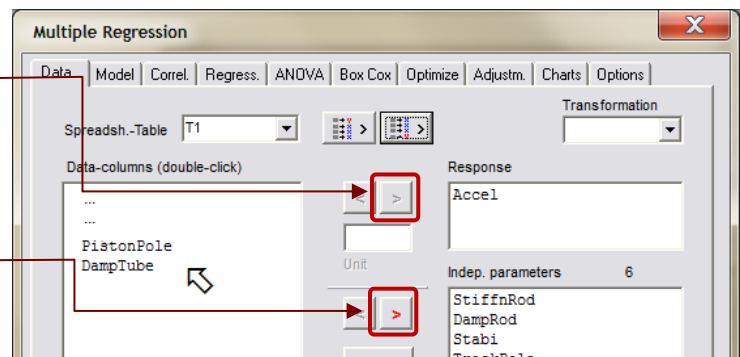
Depending from the entrance, normally the **Data Analysis Guide** will help you find out the right method for the used data. In this case, a normal Regression without transformation is suitable (continuously measurements).



Use more parameters for **Regression/ANOVA**

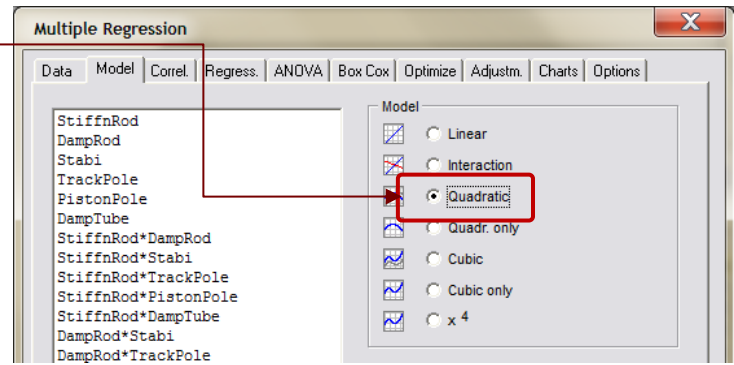
In the dialog **Multiple Regression**, the response and the factors (here independent parameters) must be selected with the respective buttons.

Note: If in the list of Data-columns a double click is used, the names will be moved in this field, where the button is red marked.



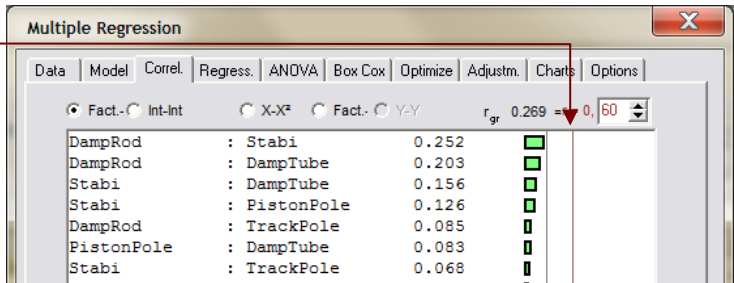
In the rubric **Model** the **Quadratic** Model has to be used analog to the experiment-definition.

As of version 16.0 also triple interactions can be selected for some defined combinations in the same way like by defining the Experiment.



The next step is to check the correlation. Because of the DoE the data is not critical here. There is a limit with the a red line, to decide if the MR is suitable. This limit comes more from experience and is not a statistical factor.

If this limit will be overstepped automatically a dialog-box appears to give some alternatives, how to proceed.

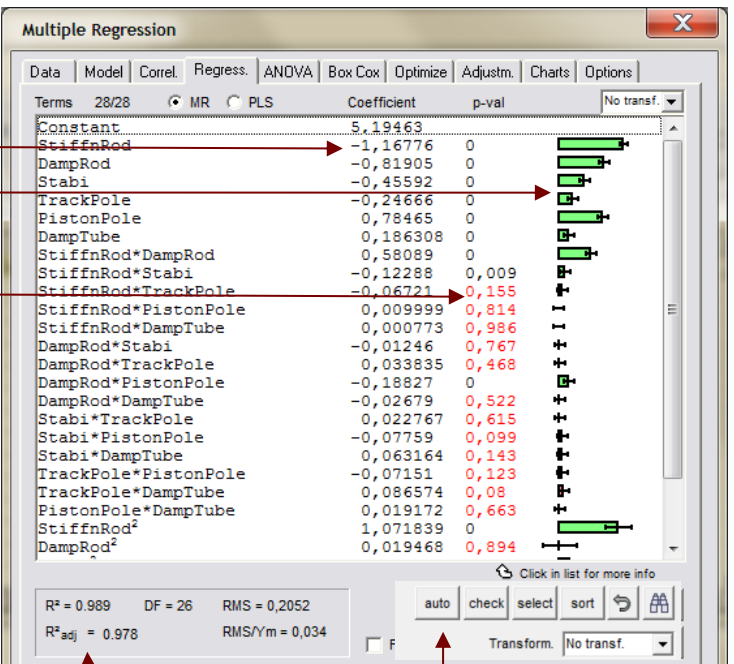


The result of the regression is shown on the next rubric. The **Coefficients** are the weight of the influence of each term.

The visualisation of this coef. are the green horizontal bars on the right with additional confidence ranges.

The **p-val** (value) is the significance for the coefficients. If the defined limit of 0.05 is exceeded the recommendation is to exclude this term from the model. This will be done for all terms by stepwise regression (see button below).

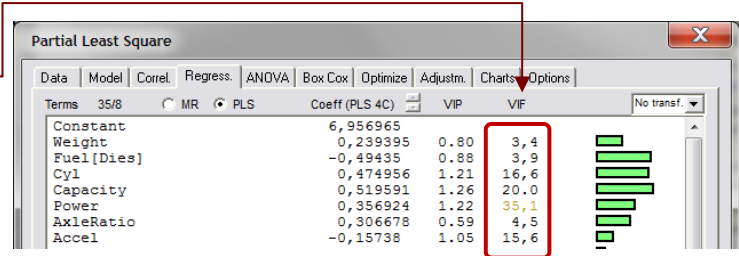
After excluding the non significant term those will be grayed, but can brought back manually (sometimes it is better to decide by technical understanding than by statistical issues).



the coefficient of determination shows how much the model can explain the data

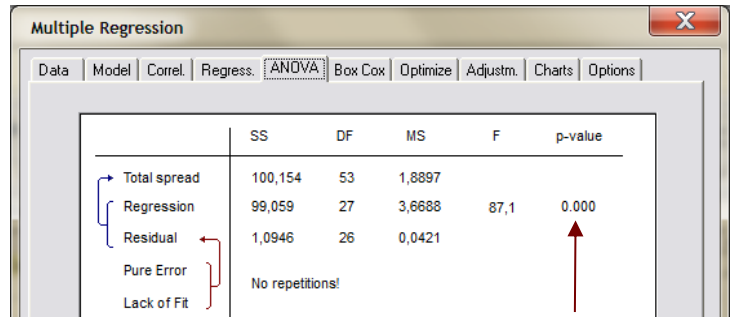
use the auto button to start the stepwise regression

Optional the so-called VIF (Variance Inflation Factor) can be shown. This is a measure of how far the model terms correlate with the others. The higher this is, the more critical the evaluation is. For more information, see the speech bubble, for each term by clicking on the respective term in the VIF column.



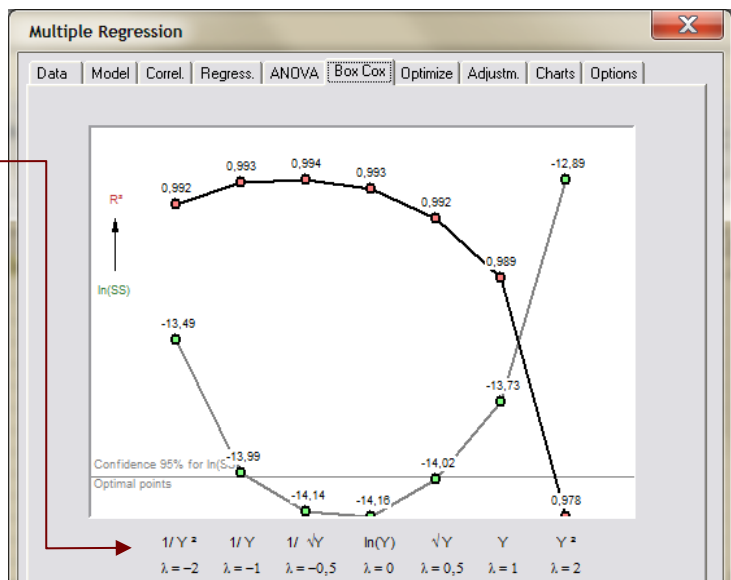
The (Model)-ANOVA gives enhanced information of how much trust one can have to the model.

For more information about the statistical values see the statistics-doc at the beginning.



this p-value shows the significance of the whole model

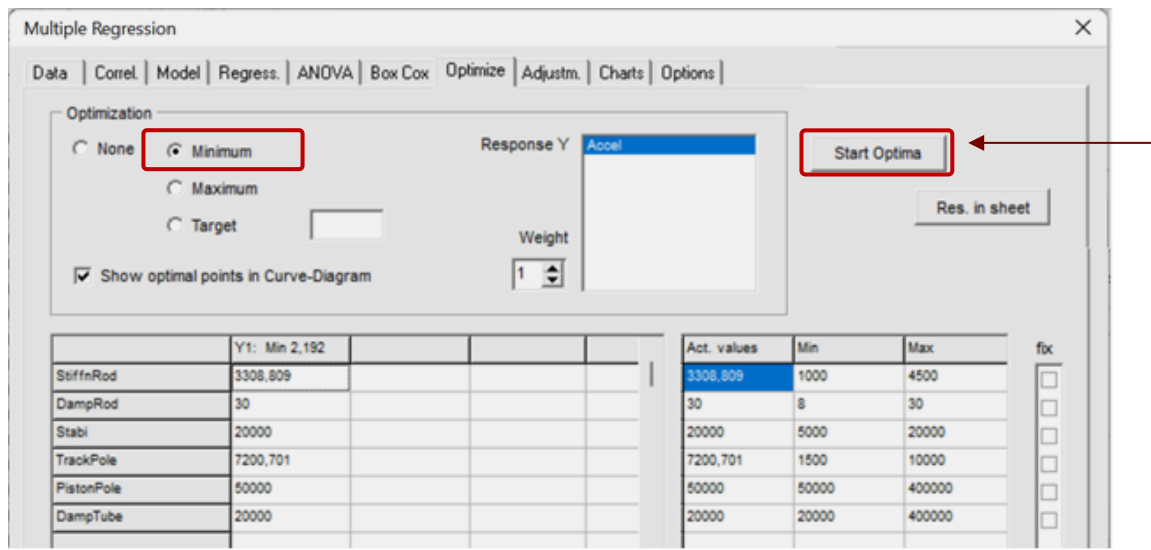
The so called Box-Cox-Transformation checks, whether the Y-data (response) should be used better by converting with mathematically standard formulas. The curve with the green points shows the special Box-Cox transformation with the goal to have the best normality of the data (must be as small as possible). The curve with the red points shows the best coefficient of determination  $R^2$  (must be as large as possible). Note: Sometimes the best transformation between the two arguments is not the same.



Recommendation: Use this transformation only, if there is a great advantage by  $R^2$ , for example by lifetime data.

The optimizer calculate on the basis of the model the best-point, what you have defined (here for example the minimum of the response). This calculation finds mostly better parameter adjustments, than the best observation of the data in the table.





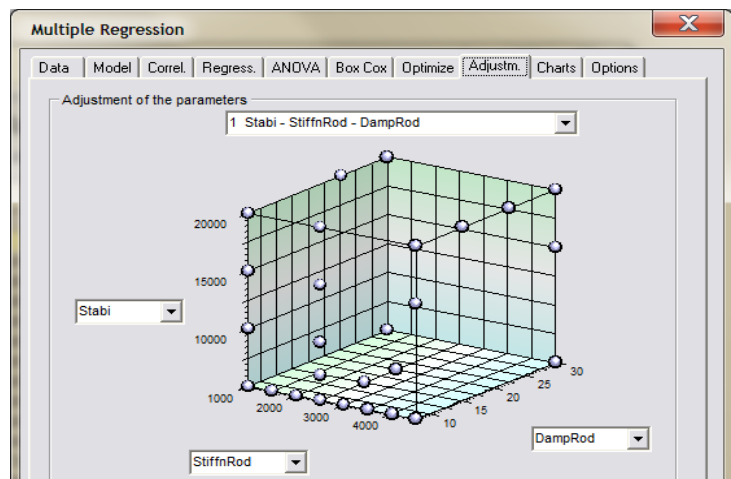
Click to start after selecting the options

If there more responses the optimizer try to find the best compromise.

If there are restrictions of some parameters, one can fix this. So only the non fixed parameter will be adjusted.

Select the "Optimal points..." to get a mark in the "Curve-diagram" later.

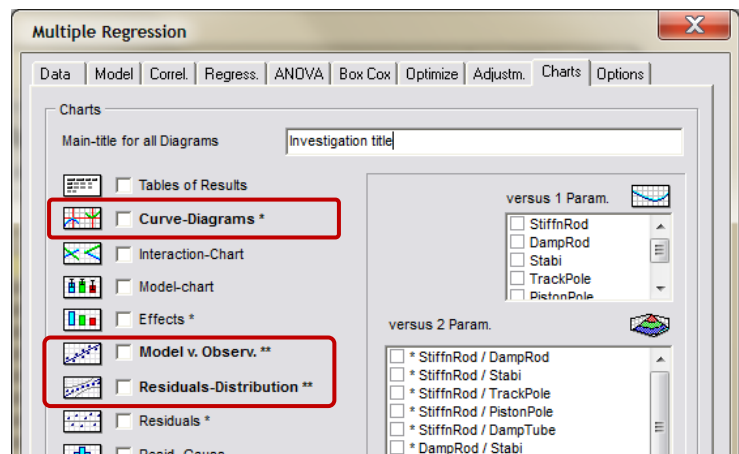
In the view of Adjustments one can judge, whether it is useful to complete experiments, especially if there are doubt about interactions. In this case it is recommended to add experiments of missing edges-points.



At the end one can select charts for the representation on the main window. The most important charts are bold marked.

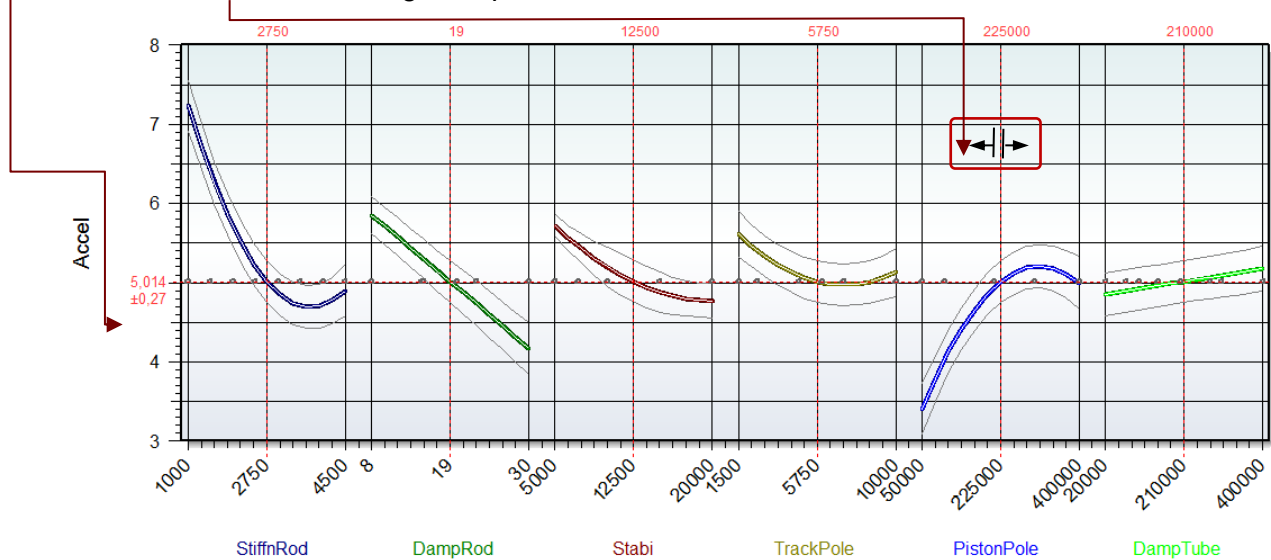
Note: do not forget to describe the project by a meaningful title in the top.

Click to the OK-button to create the charts in the main window. You can go back to the regression dialog at any time with the Data-analysis button.



The curve-diagramm shows the function of the model grafically. The steeper the slope of the curve is, the higher is the influence of the parameters.

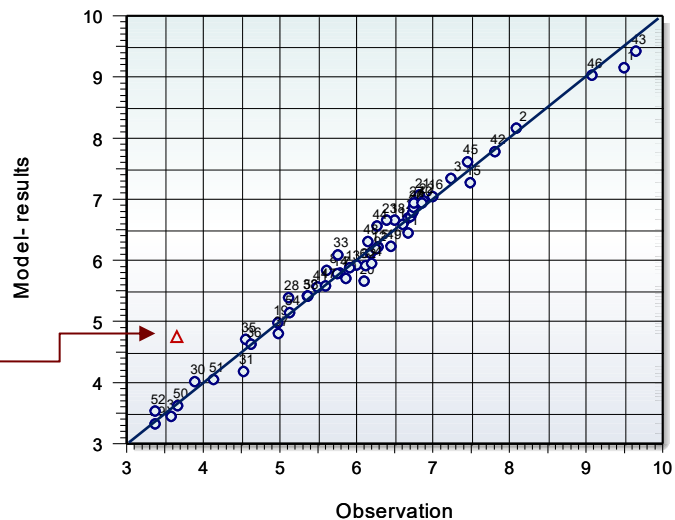
The vertical red lines represent the actual parameter adjustments with their values on the top. The horizontal red line shows the result of in the response axis together with a confidence range. Move the vertical red lines to change the parameter sets and to calculate the new result of the model.



The next diagram “Model versus Observation” gives one an overview where are the deviations between the regression model (function) and the measured values .

The best model is, if all points lie on the straight line. In this case the coefficient of determination  $R^2$  would be 1.

If outliers exists they are marked in red (not including in the example data set)



In addition there are a lot of further charts, which are not described here.

If there are any suggestions or hints about this short introduction, please give us a feedback to

info@crgraph.de