Multiple Regression

www.crgraph.de

Datenanalyse für DoE und historische Daten

Inhalt

Voraussetzung und verwandte Themen

Für diese Beschreibungen sind Grundlagen der Statistik vorteilhaft. Weiterführende und verwandte Themen sind:

www.crgraph.de/Literatur

www.versuchsmethoden.de/Versuchsplanung.pdf

www.versuchsmethoden.de/PLS.pdf

Keywords:

Multiple-Regression, Regressionsanalyse, Stepwise regression, Datenanalyse, DoE-Auswertung, Korrelation, Multikollinearität, Wechselwirkungen, Varianzanalyse, Modell-ANOVA, p-value

Einführung

Bei einer Regression wird ein Zusammenhang zwischen Einflussgrößen, bzw. -Parametern zu eine Zielgröße hergestellt. Bei einem Parameter x verwendet man hierfür eine Ausgleichsgerade ⇨ $\hat{y} = b \cdot x + a$. Eine multiple Regression erweitert die Zusammenhänge auf mehrere Einflussparameter. Die Bestimmung der Größe der Einflüsse erfolgt über die Methode der kleinsten Fehlerquadrate.

Die multiple Regression ist die Standardmethode zum Auswerten von Versuchsplänen bzw. einer DoE, aber auch für allgemeine Datenauswertungen.

Ziel und Nutzen

Das Ziel ist es ein Modell zu erstellen, mit dem man die Zusammenhänge am besten beschreiben kann. Dabei sind Wechselwirkungen sehr wichtig. Mit der dann gewonnenen Modellgleichung können Vorhersagen und Optimierungen vorgenommen werden.

Grundlagen

Soll beispielsweise der Verbrauch eines Fahrzeuges (Zielgröße) in Abhängigkeit der Einflussgrößen Gewicht, Motorleistung und Luftwiderstand bestimmt werden, so wird zunächst folgender vereinfachter Ansatz, der auch Multiple lineare Regression genannt wird:

$$\hat{y} = b_o + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 \ldots\ldots$$ (b_o = Konstante)

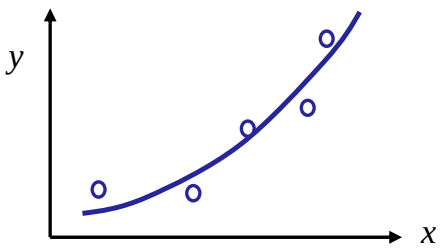
Beispiel Kraftstoffverbrauch:
x_3 = Luftwiderstand
x_2 = Leistung
x_1 = Gewicht
$\hat{y}$ = Verbrauch

Die Koeffizienten b stellen die Gewichtungen und somit die Einflussstärken dar. Sie werden über die Methode der kleinsten Fehlerquadrate bestimmt.

Weiterhin kann das Modell um quadratische Ansätze erweitert werden, um nichtlineare Zusammenhänge zu beschreiben (bekanntlich nimmt der Luftwiderstand im Quadrat zu):

$$\hat{y} = b_o + b_1 \cdot x_1 + b_2 \cdot x_1^2 + \ldots\ldots$$

quadratischer Term für Gewicht



Gibt es Wechselwirkungen, so ist das Modell um die Produkte der Einflüsse zu erweitern:

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

(der Verbrauch steigt bei gleichzeitiger Veränderung von Gewicht und Leistung überproportional mehr, als die einzelnen Einflüsse für sich). In Matrizenform schreibt man die Modellgleichung:

$$\hat{y} = X\, b$$

Hinweis: 1. Spalte in X steht für konstanten Anteil zur Bestimmung von b_o

Der gesuchte Vektor b mit den Koeffizienten bestimmt über die Matrizen-Operation:

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

In einem späteren Kapitel gibt es ein Rechenbeispiel zum Nachrechnen.

Schrittweise Regression (Stepwise regression)

Man wählt bei der DoE und bei der Regression ein Modell aus, z.B. mit allen Wechselwirkungen und allen quadratischen Termen.

Über die statistischen Hypothesentests zeigt sich als Ergebnis der sogenannten p-values, dass viele Modell-Terme nicht signifikant sind. D.h. sie lassen sich im Rahmen der Streuung nicht als Wirkung nachweisen. Deshalb sollten diese Modellterme nicht mehr verwendet werden. Es gibt zwei Methoden die Modelle aufzubauen:

Backward-Verfahren:
Bei Beginn sind alle Terme im Modell. Schrittweises Herausnehmen der nicht signifikanten Terme. Der Term mit dem schlechtesten p-value wird als erstes entfernt.

$$y = b_o + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + b_4 x_2^2 \ldots..$$

Forward-Verfahren:
Bei Beginn ist nur die Konstante im Modell Schrittweises Hereinnehmen der signifikanten Terme. Der Term mit dem besten p-value wird als erstes hineingenommen.

$$y = b_o + b_1 x_1 + b_2 x_1^2 + b_3 x_2 + b_4 x_2^2 \ldots..$$

Vorsicht: Es gibt u.U. unterschiedliche Modell-Ergebnisse zwischen beiden Vorgehensweisen, insbesondere bei nicht geplanten Daten.

Kategoriale Faktoren bzw. Parameter

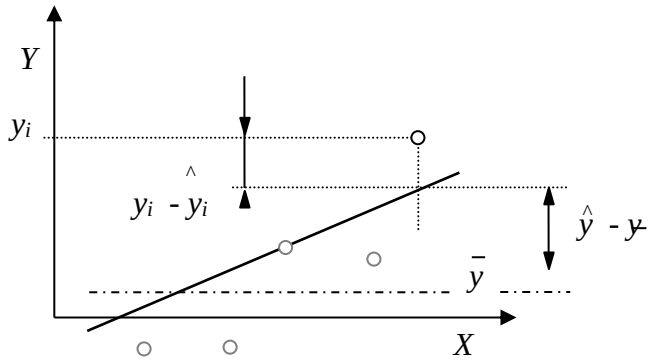
Kategoriale oder qualitative Größen, deren Variationen in Form von textlichen Benennungen angegeben werden, müssen in geeignete Zahlenform gebracht werden. Für zwei Einstellungen verwendet man -1 und 1 in einer Spalte. Ist der kategoriale Faktor F z.B. ein Bauteil von Lieferant a und Lieferant b, so erhält a den Wert -1 und b den Wert 1. Ab jedem weiteren Merkmal (Variation) wird eine zusätzliche Spalte angelegt.

	F[b]	F[c]	F[d]
a	-1	-1	-1
b	1	0	0
c	0	1	0
d	0	0	1

Die Einstellung a des allgemein genannten Faktors F stellt die Grundstufe dar. Die entsprechende Zeile enthält deshalb überall -1. Die anderen Variationen haben in ihrer „Spalte“ eine 1. Für die Anzahl der benötigten Versuche gilt die oben genannte Formel, wenn für p die Anzahl der Merkmale -1 eingesetzt wird. Versuchspläne mit kategorialen Faktoren haben aufbaubedingt teilweise Korrelationen von r=0,5 oder größer.

Das Bestimmtheitsmaß R²

Wie bei der normalen Regression interessiert zunächst der Korrelationskoeffizient r oder das Bestimmtheitsmaß R^2. Je besser dieser an dem Wert 1 liegt, desto besser wird das Merkmal y durch die Merkmale x beschrieben. Je kleiner R^2 ist, desto mehr streuen die Werte, oder es gibt gar keinen Zusammenhang zu y. R^2 besagt, wie viel Prozent der Streuung durch das Modell erklärt werden können.



$$SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SS_{Reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

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

Neben dem Bestimmtheitsmaß R² findet man häufig auch das adjustierte Bestimmtheitsmaß R²adj. Hierbei werden die entsprechenden Freiheitsgrade mitberücksichtigt:

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

SS : Sum of Squares
MS : Varianz

DF_{Reg} : Freiheitsgrad der Regression Anzahl X-Variablen im Modell $DF_{Reg} = z - 1$
(z = Anzahl Modellterme x_1, x_2, x_3, $x_1 \cdot x_2$, x_1^2 usw.)

DF_{Res} : Freiheitsgrad der Residuen $DF_{Res} = n - z - 1$
(n = Anzahl Versuche)

DF_{Total} : Freiheitsgrad Total $DF_{Total} = n-1$

Für große Stichprobenumfänge sind beide angenähert gleich. Je kleiner der Stichprobenumfang wird, desto größer ist die Abweichung. R^2 überschätzt bei kleiner Anzahl von Freiheitsgraden den Anteil erklärter Streuung mitunter erheblich. Große Unterschiede zwischen R^2 und R^2_{adj} deuten auf unnötige Terme im Modell hin.

Weitere wichtige Kenngröße sind:

- Vorhersagemaß des Modells Q²
- Lack of Fit (Modellschwäche)
- p-Value (Test der Regressionskoeffizienten)
- Standardabweichung des Gesamtmodell Root-Mean-Square RMS

Ausführliche weitere Beschreibungen und Verfahren zu

- Modell-ANAOVA (Streuungszerlegung)
- Transformation der Zielgröße mit Box-Cox
- Vertrauensbereiche

sowie deren mathatische Grundlagen sind im

Taschenbuch der statistischen Qualitäts- und Zuverlässigkeitsmethoden

zu finden, siehe Anhang.

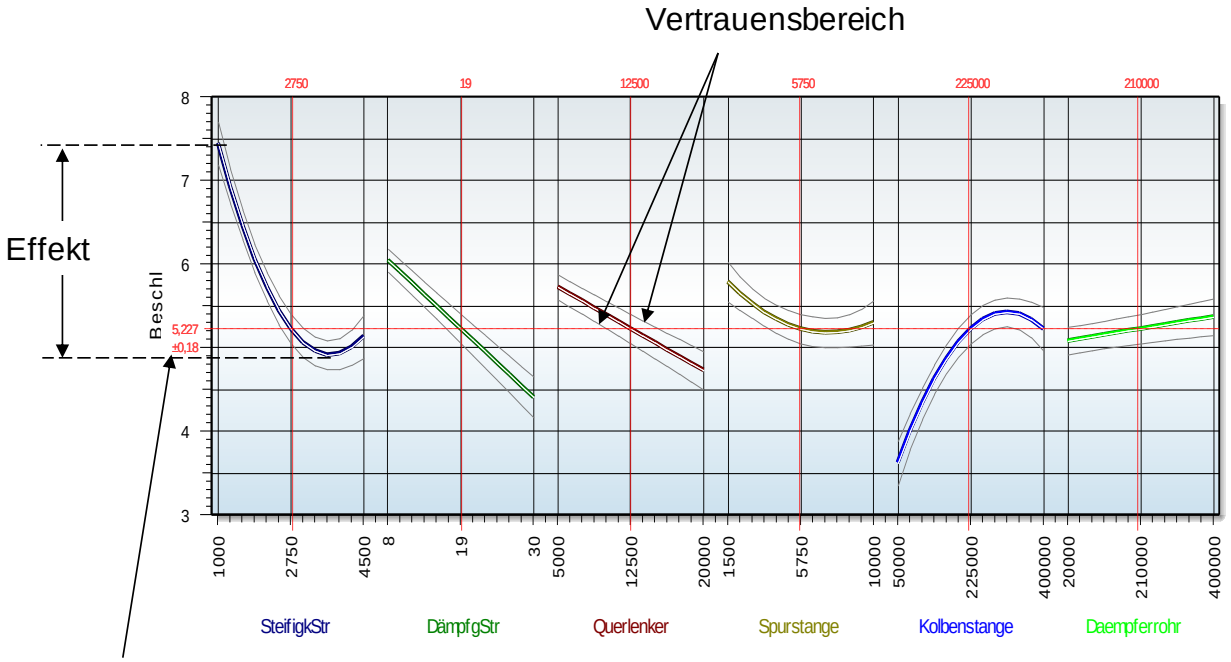
Grafische Darstellungen und weitere Kennwerte

Anstelle der reinen Modellgleichung ist das sogenannte Kurvendiagramm die beste Darstellungsform der Ergebnisse. Im folgenden Beispiel geht es um die Beschleunigung eines Schwingsystems mit den Einflüssen von Steifigkeiten und Dämpfungen.

In diesem Kurvendiagramm lassen sich sofort für jede Einstellung die Zielgrößen ablesen (gestrichelten Linien). Je stärker ein Einflussfaktor ist, desto steiler ist der Verlauf.

Als Effekt bezeichnet man innerhalb der Einstellungsgrenzen die Änderung der Zielgröße (Vertrauensbereich siehe späteres Kapitel *Vertrauensbereich für die Zielgröße*).

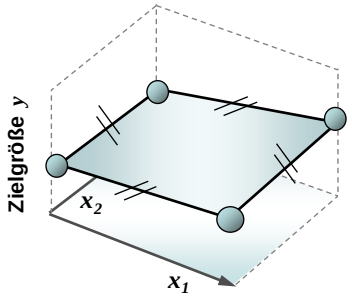
Tipp: Es sollte das Modell nicht nur aufgrund der statistischen Kennwerte beurteilt werden. Best Practice ist es, die Kurvenverläufe im oben dargestellten Diagramm auf Plausibilität und bekannte Zusammenhänge hin zu überprüfen. Z.B. kommt es oft vor, dass die Kurven in den negativen Bereich gehen, obwohl phyisikalisch keine negativen Werte möglich sind.



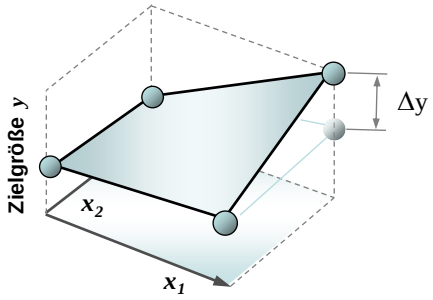
Modellwert für aktuelle Faktoreinstellungen

Grafische Darstellung der Wechselwirkung

Eine Wechselwirkung verursacht eine stärkere Veränderung der Zielgröße, als die Einflüsse der einzelnen Faktoren in Summe. Wechselwirkungen haben physikalische Ursachen:



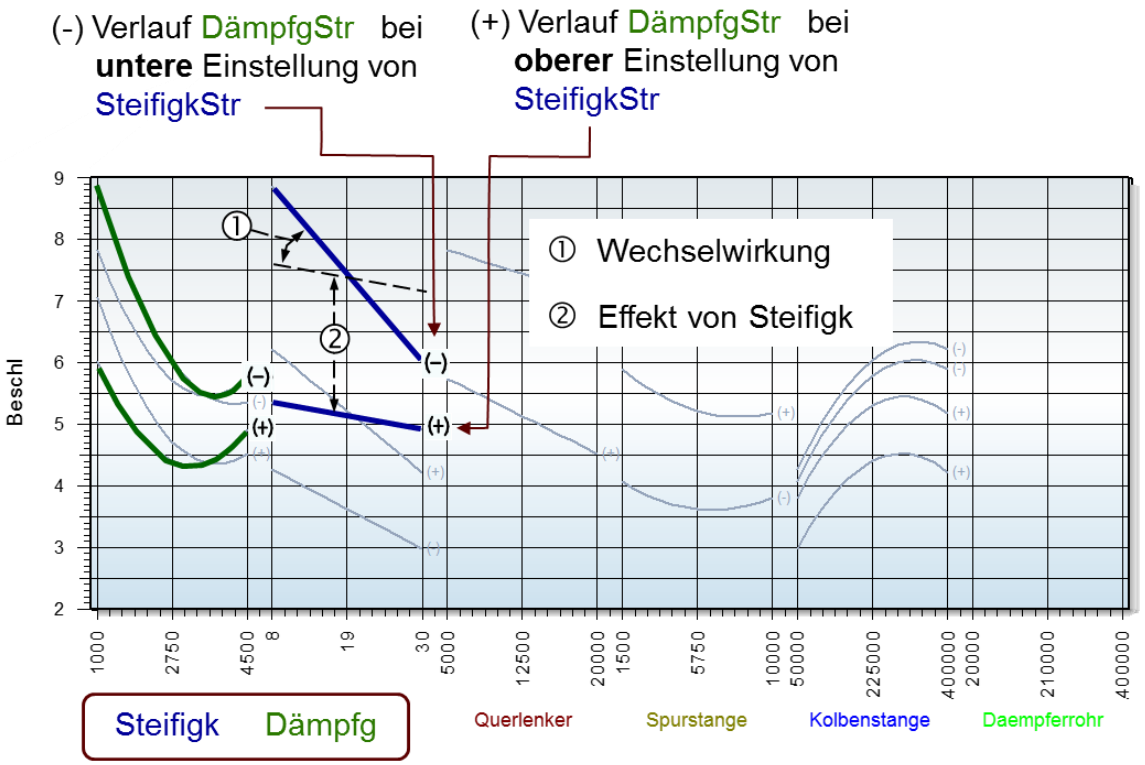
Keine Wechselwirkung
Steigungen verlaufen parallel



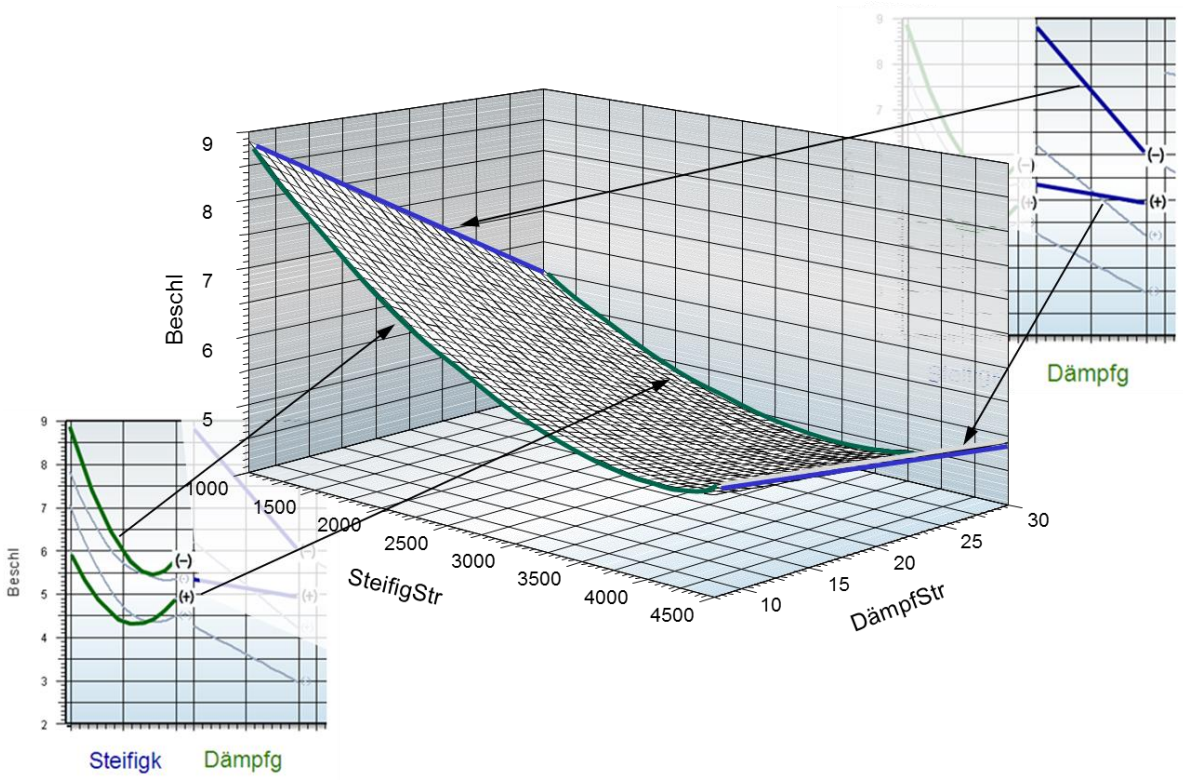
Wechselwirkung vorhanden
Steigungen verlaufen unterschiedlich

Analog zum Kurvendiagramm lassen Wechselwirkung auch als Kurvenpaare darstellen. Für jeden Faktor mit einer Wechselwirkung gibt es 2 Kurvenpaare.

Durch die Farbzuordnung kann man sehen, welcher Wechselwirkungs-Partner zu dem Parameter gehört. Die ersten beiden Kurven sind grün, da der Partner für die Steifigkeit die blau Dämpfung ist.



Verlaufen die Steigungen der Kurvenpaare mehr und mehr unterschiedlich steil, umso höher ist der Einfluss der Wechselwirkung. Der Höhenunterschied der Kurven ist lediglich der Effekt des Faktors, der der Wechselwirkungspartner ist. In einem 3D-Digramm sieht das dann folgerndermaßen aus:

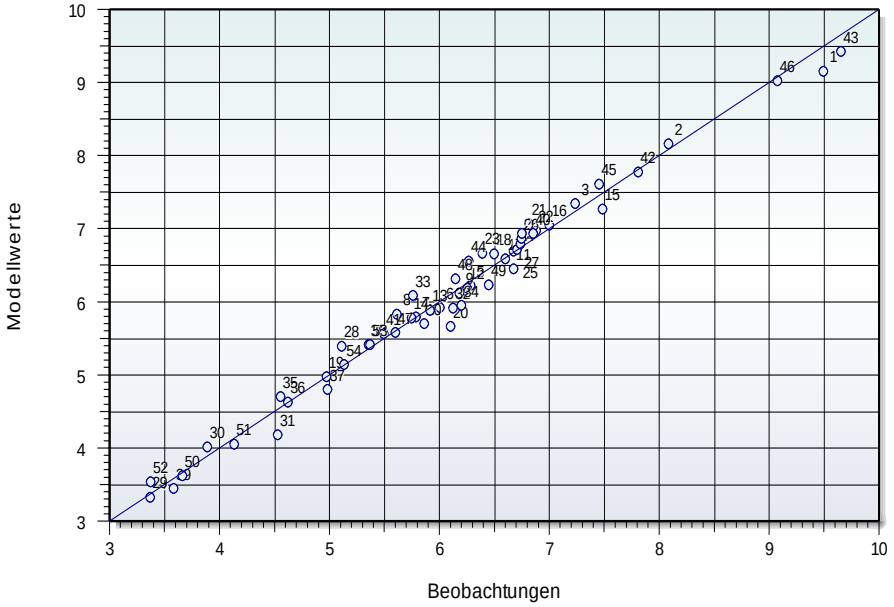


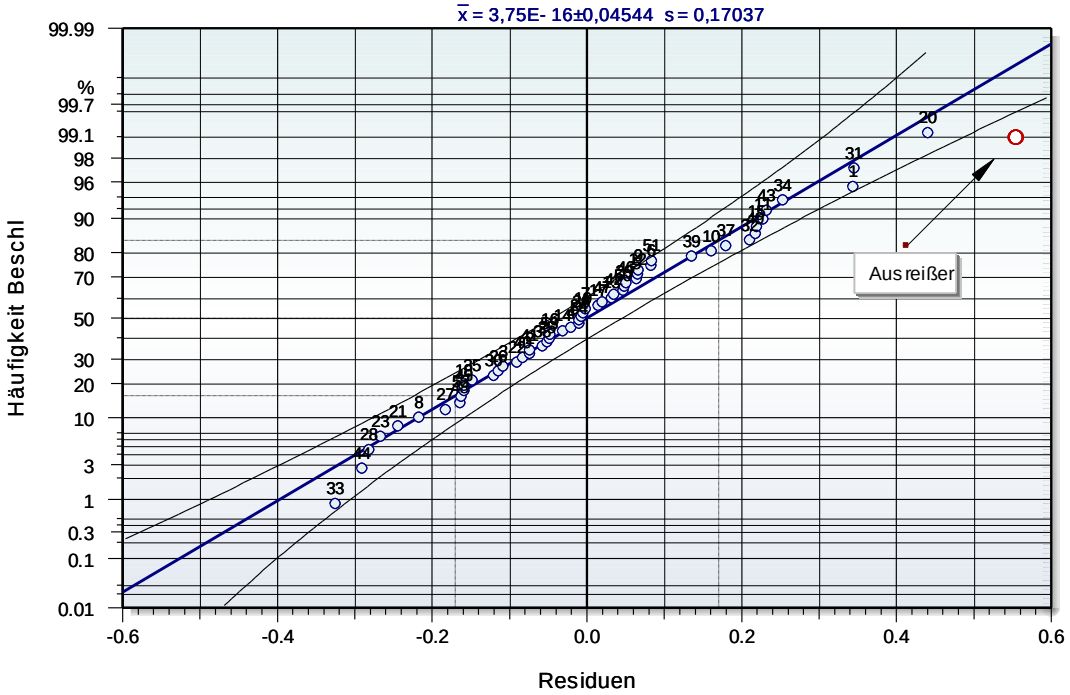
Residuenverteilung

Die „Güte“ eines Modells kann auch dargestellt werden, indem man die jeweiligen Rechenwerte (Funktionswerte) über die beobachteten Werte y_i aufträgt. Bezogen auf die tabellarischen Daten bedeutet dies, dass man für jede Zeile die Merkmale x in das Modell eingibt und $\hat{y}_i$ berechnet. Dabei wird dieser Rechenwert $\hat{y}_i$ (Modellwert) von dem Wert der Beobachtung (Messwert) y_i in der Tabelle mehr oder weniger abweichen. Diese Abweichungen stellen die so genannten Residuen dar.

Je besser das Modell ist, desto näher liegen die einzelnen Punkte auf der Mittellinie, die unter 45° verläuft.

Für diese Abweichungen kann man eine Häufigkeitsverteilung erstellen, um zu beurteilen, ob die Residuen normalverteilt sind.





Abweichungen von der Normalverteilung sind ein Indiz dafür, dass systematische Fehler vorliegen, oder andere Störgrößen einen Einfluss haben. Ist während der Messung z.B. die Temperatur hoch gegangen oder wurden Veränderungen am Messaufbau vorgenommen. Weiterhin sind Ausreißer hier markant zu erkennen (weit außerhalb der Geraden, statistische Bewertung siehe Grubbs-Test).

Probleme der Multiplen Regression

Die multiple Regression funktioniert gut, wenn die Daten aus einer Versuchsplanung kommen. Sobald die Daten jedoch mehr und mehr korrelieren, sind die Signifikanztests nicht mehr eindeutig – Problem der sogenannten Multikollinearitäten. Die Programmbeschreibung von Visual-XSel am Ende zeigt, wie man damit umgehen kann.

Über einen paarweisen Test auf Korrelationen, kann vorab geprüft werden, ob Parameter untereinander ein Problem haben. Gesamthaft gibt der Varianz-Inflations-Faktor, kurz VIF einen guten Überblick:

Wechselwirkungen, die eines der Hauptziele der DoE sind, können u.U. nicht ganz eindeutig bestimmt werden. Gibt es Messfehler, so können diese als Wechselwirkungen gedeutet werden, obwohl sie phyisikalisch eigentlich nicht relevant sind. Expertenwissen sollte hier immer eingebracht werden.

Eine Alternative bei hohen Multikollinearitäten ist das PLS-Verfahren, jedoch sind hier auch die Wechselwirkungen nicht sicherer zu bestimmen.

Grundsätzlich ist davon abzuraten, Datensätze ohne Vorwissen auszuwerten und die Ergebnisse als die „wahre Welt“ zu betrachten.

Vergleich der Multiplen Regression mit anderen Auswertemethoden

Insbesondere für die Auswertung von Versuchplänen ist die klassischen Auswertemethode eigentlich die Varianzanalyse, auch als ANOVA bekannt. Die Multiple Regression hat die ANOVA heute aber aufgrund viele Vorteile praktisch abgelöst. Folgende Tabelle, die wohl keinen Anspruch auf Vollständigkeit hat, zeigt eine Gegenüberstellung der wichtigsten Methoden:

	Vorteile	Nachteile
Multiple Regression	▪ Einfache Handhabung ▪ Interpretierbare Modelle ▪ Schrittweise Regression ▪ Beliebige Mischmodelle möglich	▪ Mehr Datenzeilen nötig, als Modell-Terme ▪ Probleme bei Koliniearitäten
Partial Least Square	▪ Ideal für Daten die korrelieren ▪ Auswertung für historische Daten ohne Versuchplanung ▪ Weniger Datenzeilen als Modell-Terme möglich ▪ Liefert weitere Kennzahl für Modell-Terme (VIP).	▪ Modell-Koeffizienten werden zu gering geschätzt. ▪ Ersetzt nicht die Multiple Regression
ANOVA	▪ Klassischer Hypothesentest ▪ Prozentuale Darstellung der Einflüsse ▪ Berücksichtigung abhängiger Gruppen bei nested ANOVA	▪ Probleme mit Koliniearitäten ▪ Begrenzte Modellbildung ▪ Homogenität der Varianz nötig ▪ Klassische ANOVA setzt Unabhängigkeit und Zufälligkeit voraus
Kreuztabellen Chi²	▪ Diskrete Zielgr. > 2 Auspräg. ▪ Aussagekräftige Paarvergleiche ▪ Vergleich von kategorialen Gruppen	▪ Liefert kein Modell ▪ Stat. Test über Chi² stark von Stichprobe abhängig
Neuronale Netze	▪ Umfangreiche Daten möglich ▪ Extr. nichtlineare Zusammenhäng darstellbar.	▪ Nicht interpretiere Modelle ▪ Benötigt große Datenmengen zum Antrainieren.

Weitere Auswerteverfahren für unterschiedliche Zielgrößen

- Logistische oder diskrete Regression
  (Zielgröße diskret auf zwei Stufen, z.B. gut/schlecht)

- Poisson-Regression
  Damit beschreibt man zählbare Ereignisse

Rechenbeispiel:

Es liegt ein Modell mit einer Wechselwirkung vor:

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

Versuchsplan:			Ergebnisse 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

Die einzelnen Schritte der Gleichung $b = \left(X^T X\right)^{-1} X^T y$ ergeben sich wie folgt:

$X' = X^T X$ mit $X = \begin{bmatrix} 1 & x_{11} & .. & x_{z1} \\ 1 & x_{12} & .. & x_{z2} \\ .. & .. & .. & .. \\ 1 & x_{1n} & .. & x_{zn} \end{bmatrix}$ $z+1$ Spalten und n Zeilen

Die jeweiligen Zellen berechnen sich nacheinander entsprechend mit:

$$x'_{j,i} = \sum_{k=1}^{n} x_{k,i}^{(T)} \, x_{j,k}$$ (erster Index = Spalte, zweiter Index = Zeilen)

Die erste Spalte mit jeweils einer 1 steht für die Konstante b_o. Die beiden weiteren für die Hauptfaktoren x_1 und x_2 und die letzte Spalte errechnet sich aus dem Produkt der Spalte 2 und 3 (Wechselwirkung von x_1 und x_2).

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

z.B. Zelle
J=1 i=1
$x'_{1,1} = (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) + (1)\cdot(1) = 5$

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

gesamthaft ergibt sich:

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

und invertiert:

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

und über den Zwischenschritt:

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

erhält man das Ergebnis für die gesuchten Koeffizienten:

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

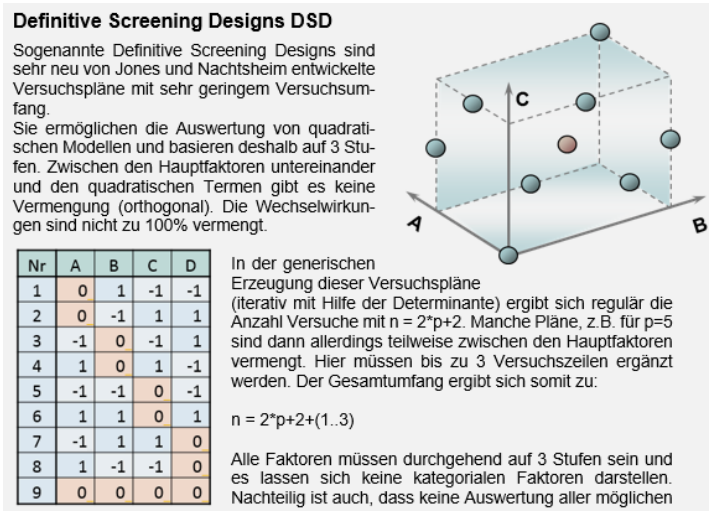
Die Gleichung vom Anfang lautet also:

$$\hat{y} = 6{,}4 + 1{,}5\, x_1 + 2{,}5\, x_2 + 0{,}5\, x_1\, x_2$$

Literatur - Weiterführende Beschreibungen

Ausführliche softwareunabhängige Beschreibungen zum Thema DoE und der dazugehörigen Auswertungen gibt es im

Taschenbuch der statistischen Qualitäts- und Zuverlässigkeitsmethoden



Definitive Screening Designs DSD

Sogenannte Definitive Screening Designs sind sehr neu von Jones und Nachtsheim entwickelte Versuchspläne mit sehr geringem Versuchsumfang.
Sie ermöglichen die Auswertung von quadratischen Modellen und basieren deshalb auf 3 Stufen. Zwischen den Hauptfaktoren untereinander und den quadratischen Termen gibt es keine Vermengung (orthogonal). Die Wechselwirkungen sind nicht zu 100% vermengt.

Nr	A	B	C	D
1	0	1	-1	-1
2	0	-1	1	1
3	-1	0	-1	1
4	1	0	1	-1
5	-1	-1	0	-1
6	1	1	0	1
7	-1	1	1	0
8	1	-1	-1	0
9	0	0	0	0

In der generischen Erzeugung dieser Versuchspläne (iterativ mit Hilfe der Determinante) ergibt sich regulär die Anzahl Versuche mit n = 2*p+2. Manche Pläne, z.B. für p=5 sind dann allerdings teilweise zwischen den Hauptfaktoren vermengt. Hier müssen bis zu 3 Versuchszeilen ergänzt werden. Der Gesamtumfang ergibt sich somit zu:

n = 2*p+2+(1..3)

Alle Faktoren müssen durchgehend auf 3 Stufen sein und es lassen sich keine kategorialen Faktoren darstellen. Nachteilig ist auch, dass keine Auswertung aller möglichen

Weitere Informationen und Leseproben:
crgraph.de/Literatur

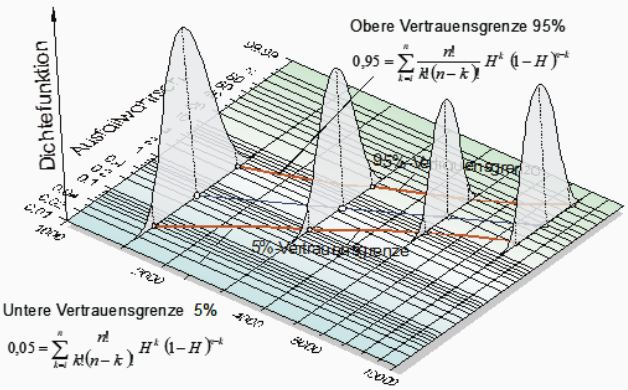
Speziell das Buch

Weibull & Zuverlässigkeitsmethoden

vertieft anwendungsbezogen die Statistiken und Methoden rund um Weibull und aller weitere Verteilungen. Die Versuchsplanung behandelt hier spezielle Lebensdauerfragen aufgrund unterschiedlicher Belastungen, Temperaturen, etc.

2.5.1 Vertrauensbereich der Weibull-Gerade

Bei der Weibull-Auswertung handelt es sich praktisch immer um eine Stichprobe. Die Gerade im Weibull-Diagramm entspricht also nur der Stichprobe. Je mehr Teile geprüft oder ausgewertet werden, desto mehr streuen die „Punkte" um die Weibull-Gerade. Man kann statistisch eine Abschätzung über den Bereich der Grundgesamtheit machen. Hierfür wird ein sogenannter „Vertrauensbereich" eingeführt. In der Regel gibt man diesen mit 90% an. Die obere Vertrauensgrenze entspricht dann einer Aussagewahrscheinlichkeit von P_A=95%.



Weitere Informationen und Leseproben:
crgraph.de/Literatur

Consulting & Schulungen

Bei unseren Inhouse- oder Online-Schulungen wird die praxisnahe Anwendung von statistischen Methoden vermittelt. Wir haben über 25 Jahre Erfahrung, insbesondere in der Automobilindustrie und unterstützen Sie bei Ihren Problemstellungen, führen Auswertungen für Sie durch, oder erstellen firmenspezifische Auswertevorlagen.

Weitere Informationen finden Sie unter:
crgraph.de/schulungen

Sie haben ein konkretes Qualitätsproblem, oder wollen ein Produkt effizient und zuverlässig entwickeln? Sie wollen keine Statistik-Software anschaffen, weil diese voraussichtlich zu selten gebraucht wird, oder weil zu wenig Zeit zur Einarbeitung vorhanden ist? Dann sind unsere Q-Support Pakete genau das Richtige:

crgraph.de/consulting

Hotline

Haben Sie noch Fragen, oder Anregungen? Wir stehen Ihnen gerne zur Verfügung:

Tel. +49 (0)8151-9193638

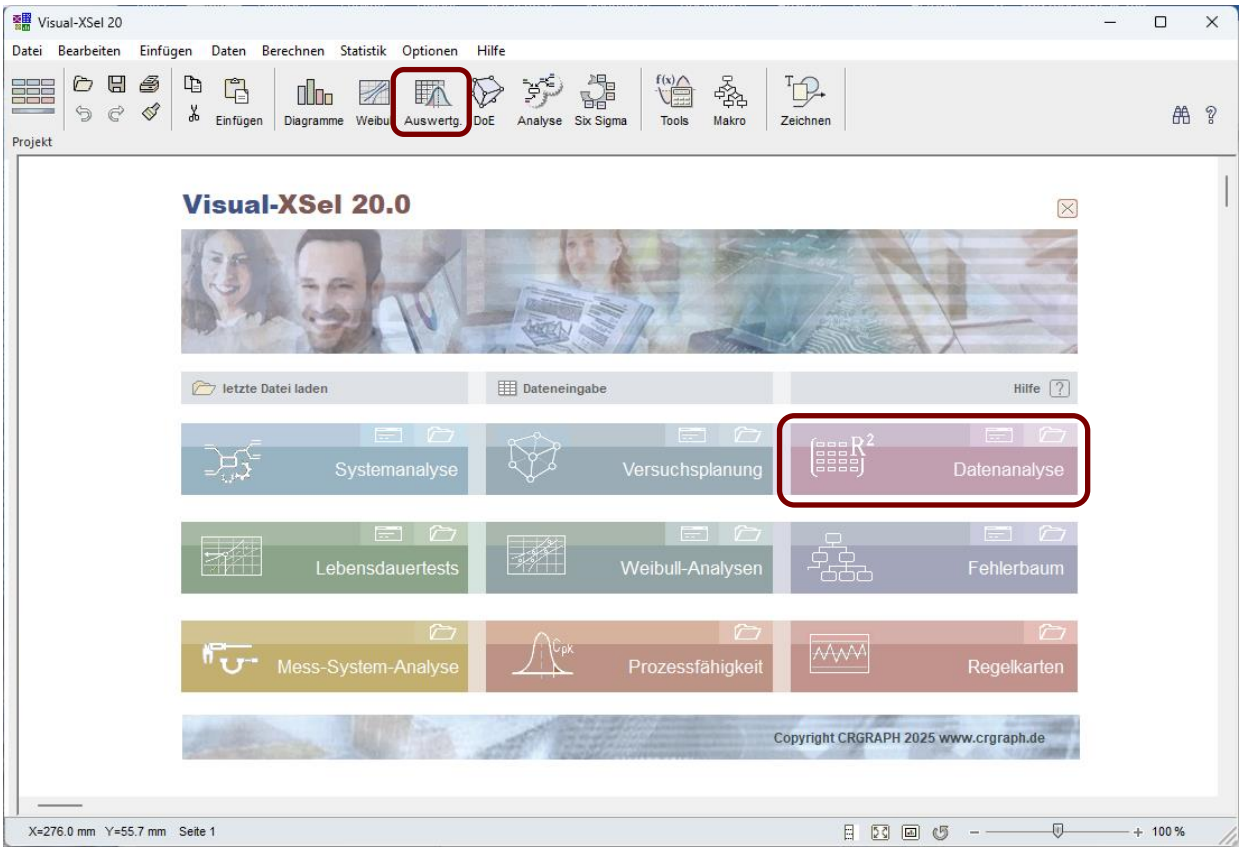
E-Mail: info@crgraph.de

Besuchen Sie uns auf unserer Home-Page: www.crgraph.de

Anwendung in Visual-XSel

www.crgraph.de

Unsere Software **Visual-XSel** ist ein leistungsfähiges Tool für alle wichtigen statistischen Qualitäts- und Zuverlässigkeitsmethoden. Verwenden Sie für den Einstieg die **Versuchsplanung** im Leitfaden (siehe auch crgraph.de/themen-index), oder die Ikone **DoE**.



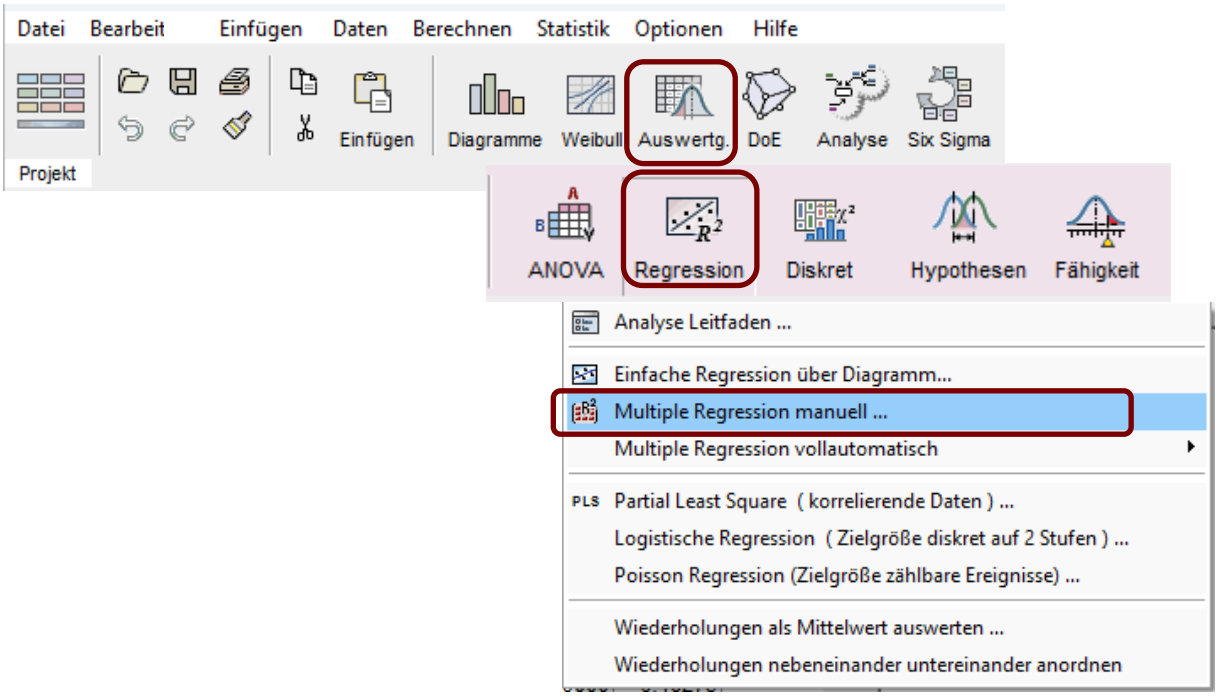
Hier finden Sie eine Übersicht und Einstiegsvideos:
crgraph.de/visual-xsel-software/

Nicht umsonst ist diese Software in vielen namhaften Firmen im Einsatz:
crgraph.de/Referenzen.

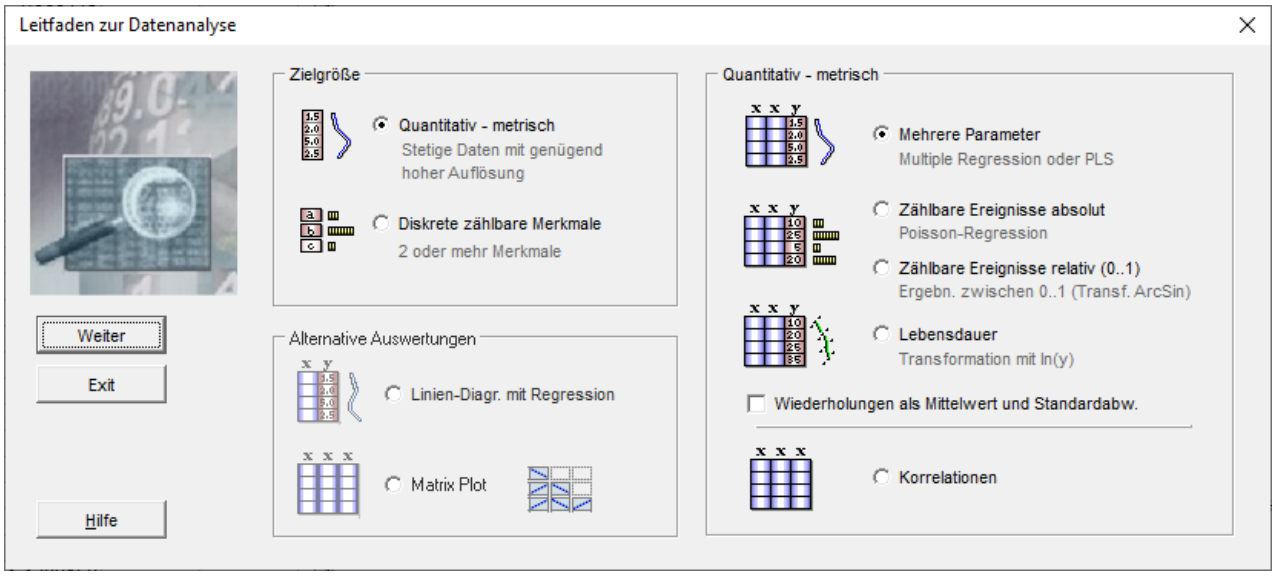
Die folgende Beschreibung ist eine Anleitung und Einführung in die Datenauswertung in Visual-XSel.

Visual-XSel ist besitzt eine sehr leistungsfähige Datenanalyse mit den beschriebenen Verfahren. Nicht umsonst verwendet der mit dem Ressourceneffizienzpreis 2021 ausgezeichnete Analyser® Visual-XSel für die Modellbildung (www.contech-analyser.de)

Bei entsprechenden Daten in der Tabelle, z.B. *…\Beispieldaten\Beispiel_Radaufhängung_MulReg.xls,* verwenden Sie für die Datenauswertung den Leitfaden oder die Ikone *Auswertung*



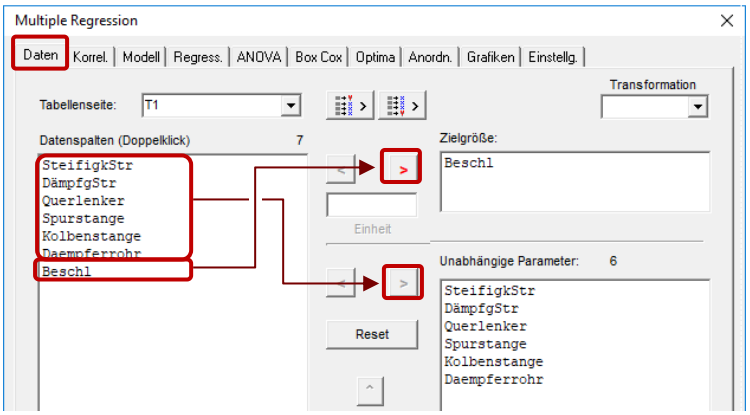
Bei Verwendung des Analyse Leitfadens erscheint folgende Auswahl:



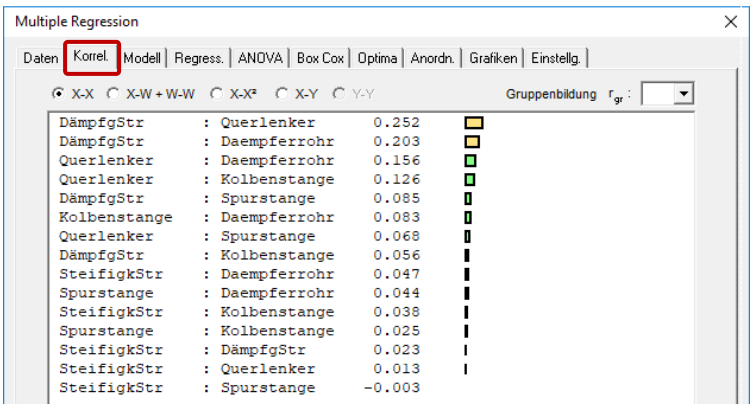
Hier wird festgelegt, was für eine Art der Zielgröße vorliegt. Die Standardeinstellung ist quantitativ bzw. sind metrische Daten. Bei zählbaren Ereignissen absolut wird die Poisson-Regression mit der Maximum-Likelihood-Methode angewendet. Bei zählbaren Ereignissen relativ wird die Zielgröße math. so transformiert, dass das Ergebnis nur zwischen 0…1 liegt. Bei Lebensdauerwerten müssen diese logarithmiert werden und Partial-Least-Square kann mit korrelierenden Daten umgehen, die nicht aus einer DoE stammten. Die zuletzt genannte Methode kann man auch später auswählen, da die Korrelation erst noch zu prüfen ist.

Ordnen Sie die in der Kopfzeile der Tabelle vorkommenden Titel jeweils als Zielgröße oder als Parameter zu.

Sowohl Zielgröße, als auch Parameter können hier transformiert werden.
Wurde vorher die Zielgröße als Lebensdauer angegeben, steht hier ln(y).
Bei ungekannten math. Zusammenhängen hilft später als Entscheidung die Box-Cox Analyse.

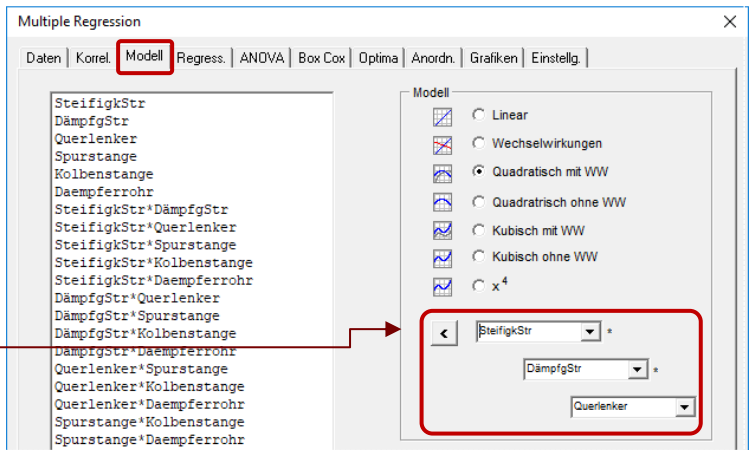


Unter dem nächsten Reiter kann eine mögliche kritische **Korrelation** erkannt werden. In der Regel sollten die Parameter keine Korrelation r > 0,60 haben. Für den Fall, dass die Korrelationen zu hoch sind, erscheint automatisch ein Dialog mit der Auswahl weiterer möglicher Maßnahmen. Weitere Beschreibungen hierzu weiter hinten unter Expertenwissen.

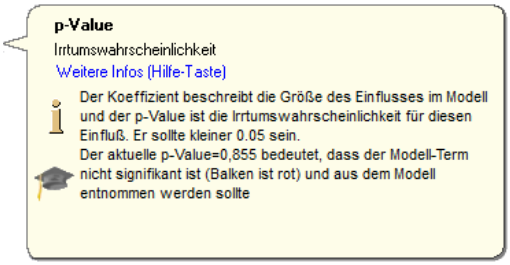


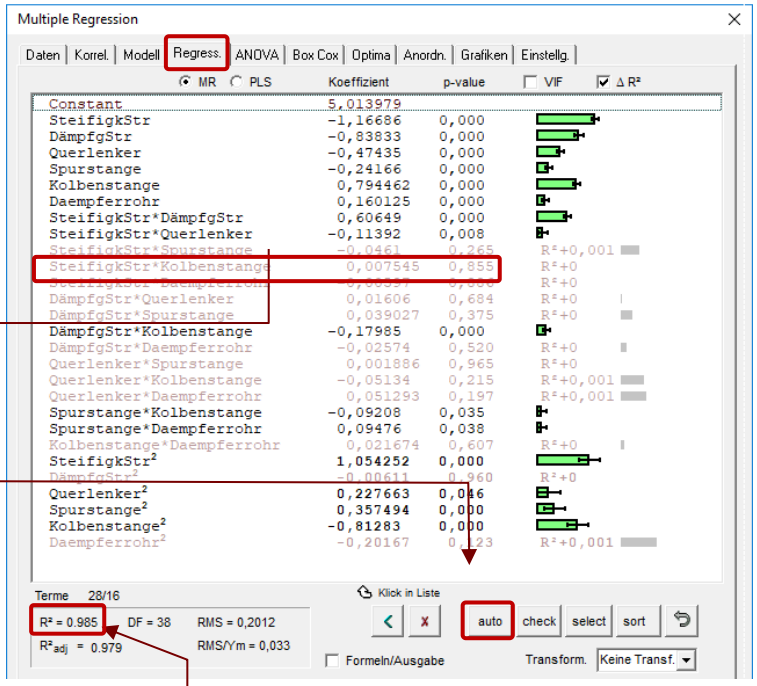
Unter dem Reiter **Modell** kann ausgewählt werden, ob ein rein lineares Modell, oder eines mit Wechselwirkungen, oder nichtlinearen Verläufen zu verwenden ist (Quadratisch). Für den Fall, dass historische Daten vorliegen (ohne Versuchsplan) ist evtl. ein quadratisches Modell ohne Wechselwirkungen zu empfehlen, da bei korrelierenden Daten Wechselwirkungen problematisch sind.

Es können auch gezielt 3-fach-Wechselwirkungen ausgewählt werden.



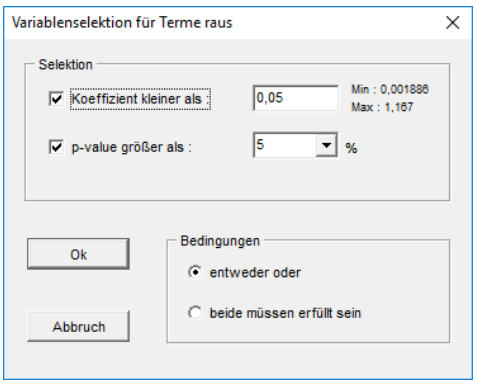
Das Ergebnis der multiplen **Regression** erscheint zunächst für das volle Modell auf dem nächsten Reiter. Darin sind zu Beginn immer Modellterme, die nicht signifikant sind. Klicken Sie auf einen beliebigen Term im Modell, um weitere Informationen zu erhalten.

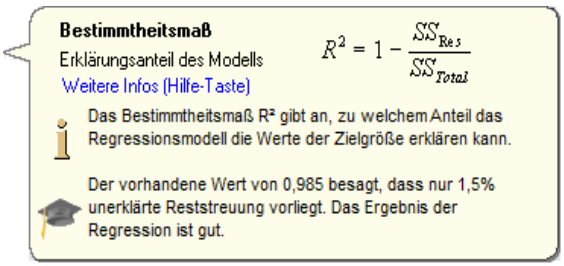




Um nicht signifikante aus dem Modell zu entfernen, kann die Taste auto verwendet werden. Danach erscheinen diese in grau und können manuell ins Modell zurückgeholt werden <, z.B. wenn trotz geringer Signifikanz fachlich begründet der Term doch im Modell bleiben soll.

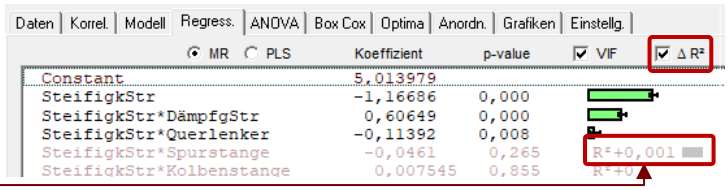
Soll für die Modell-Terme nicht nur der p-value ausschlaggebend sein, ist es über die Taste select möglich auch die Größe der Koeffizienten einzubeziehen.



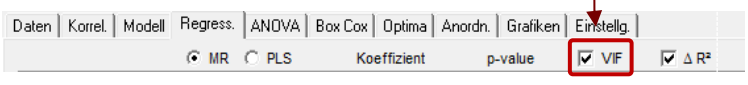


Eine wichtige Kenngröße zur Beurteilung des Modells ist das Bestimmtheitsmaß R². Fahren Sie mit der Maus über die beiden Kennzahlen um weiteres zu erfahren.

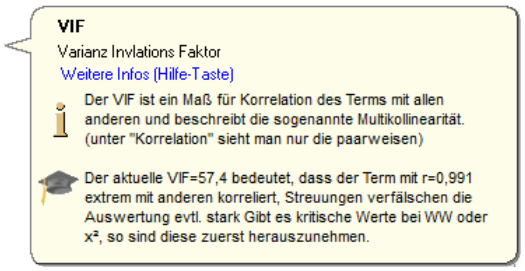
Weiterhin sollte man berücksichtigen, ob ein Term in grau wieder zurück in das Modell soll, welche Verbesserung auf das R² damit erreicht wird. Im Fall der ersten Wechselwirkung würde sich das R² nur um 0,001 erhöhen.



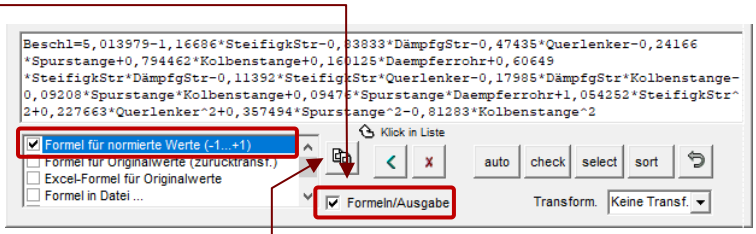
Die optionale Kennzahl **VIF** (Varianz-Inflations-Faktor) sagt aus, inwieweit die Modell-Terme mit allen anderen gemeinsam korrelieren. Je höher dieser ist, desto kritischer ist die Auswertung. Der VIF sollte nicht größer als 10



sein. Weitere Informationen zeigt die Sprechblase für jeden Term, wenn man in der Spalte VIF auf den jeweiligen Term klickt. Weitere Details unter:



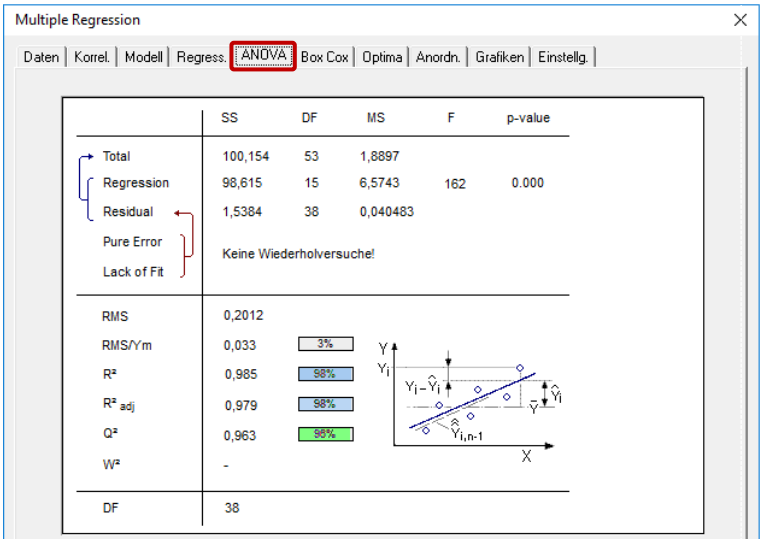
Setzt man den Haken unter Formeln, so kann man sich die Modellgleichung anzeigen lassen. In der Formel für normierte Werte muss man für die Parameterwerte den Bereich zwischen -1 .. +1 verwenden.
Bei der Variante mit Originalwerten können die ursprünglichen Einheiten eingesetzt werden.
Es ist auch möglich in dieser Auswahl die Formel, oder die gesamte Ergebnistabelle in eine Datei auszugeben.



Unter der Taste Kopieren kann diese Gleichung exportiert werden. Damit lassen sich Vorhersagen auch ohne diese Software machen

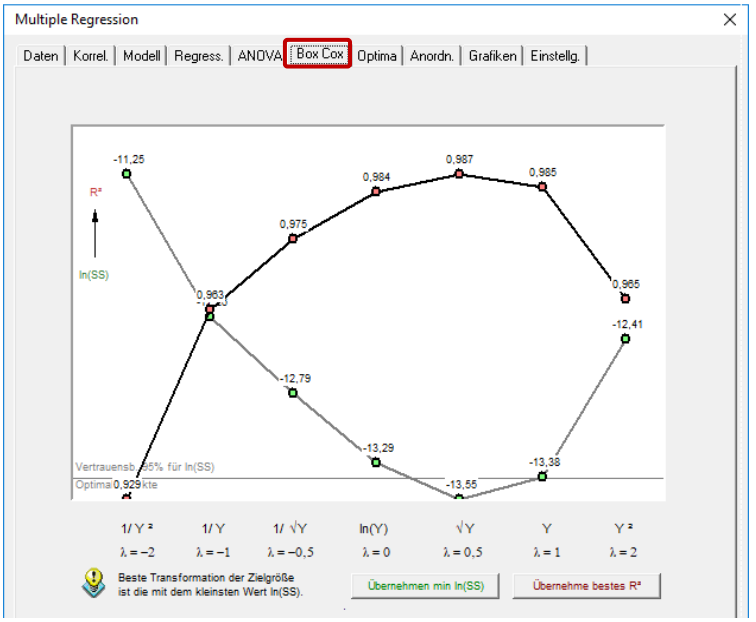
Unter der **ANOVA** werden weitere wichtige Kennzahlen zur Beurteilung des Modells gezeigt, wie unter den Grundlagen beschrieben. Die Wiederholbarkeit W² wird hier nicht angezeigt, da es keine Wiederholungen gibt. Somit ist hier auch kein Pure Error und das Lack of Fit bestimmbar.

Hinweis: Das Vorhersagemaß Q² wird in den Ergebnistabellen auf der Hauptseite des Programmes nur angezeigt, wenn die ANOVA explizit aufgerufen wurde. Dies gilt insbesondere bei mehreren Zielgrößen.

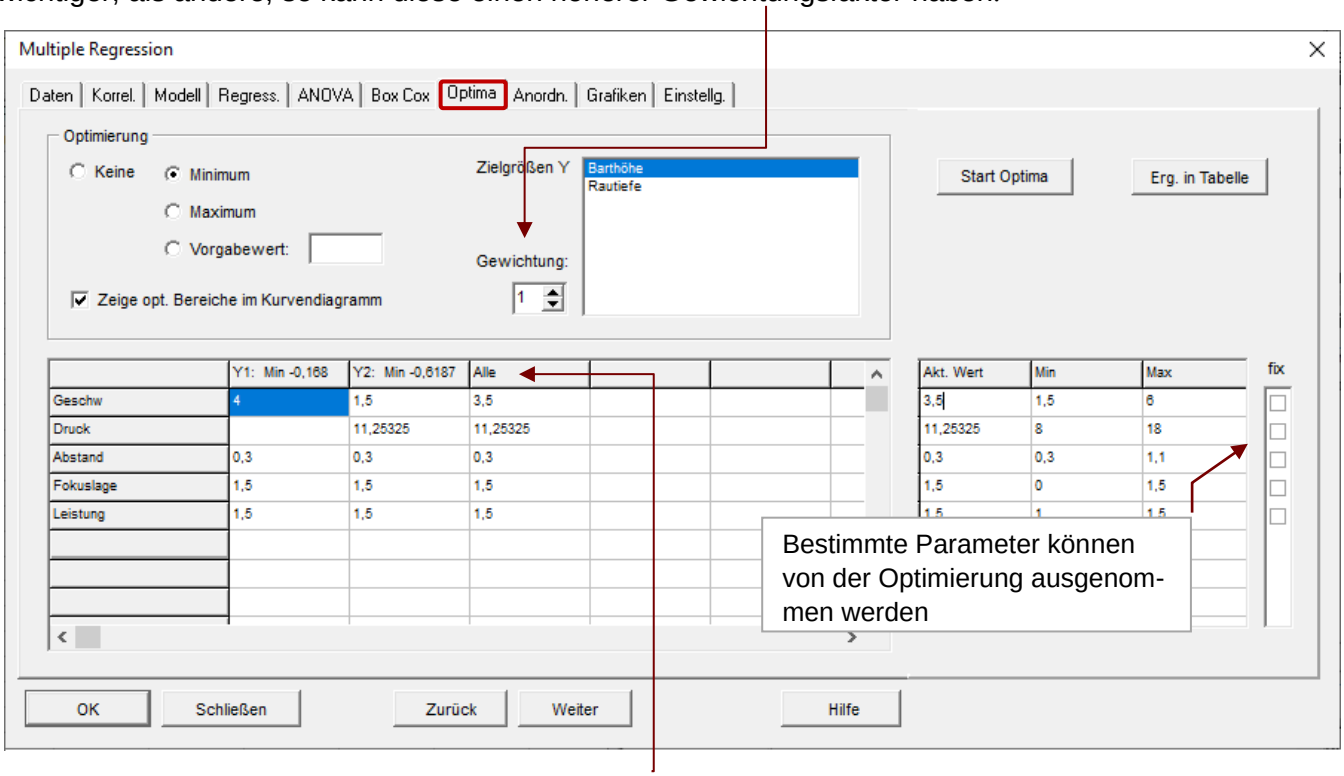


Die Zielgröße kann mit mathematischen Grundfunktionen transformiert werden. Das Ziel nach **Box-Cox** ist es hierdurch die Residuen ln(SS) in Hinblick auf die Normalverteilung, so klein wie möglich zu halten. Dieser Zusammenhang wird durch die grünen Punkte dargestellt.
Bei welcher Transformation das beste R^2 entsteht, zeigen die roten Punkte.
Zu beachten ist, dass es möglich ist, dass beide optimalen Transformationen nicht die selben sind.
Die jeweils gewünschte Umsetzung kann durch die Tasten unten erreicht werden.

Hinweis: Box-Cox ist nur möglich, wenn die Daten der Zielgröße nicht kleiner gleich 0 sind.



Unter Rubrik ***Optima*** können eine oder mehrere Zielgrößen und deren Modell optimiert werden. Für jede Zielgröße ist ein Minimum, ein Maximum, oder ein Vorgabewert definierbar. Ist eine bestimmte Zielgröße wichtiger, als andere, so kann diese einen höherer Gewichtungsfaktor haben.

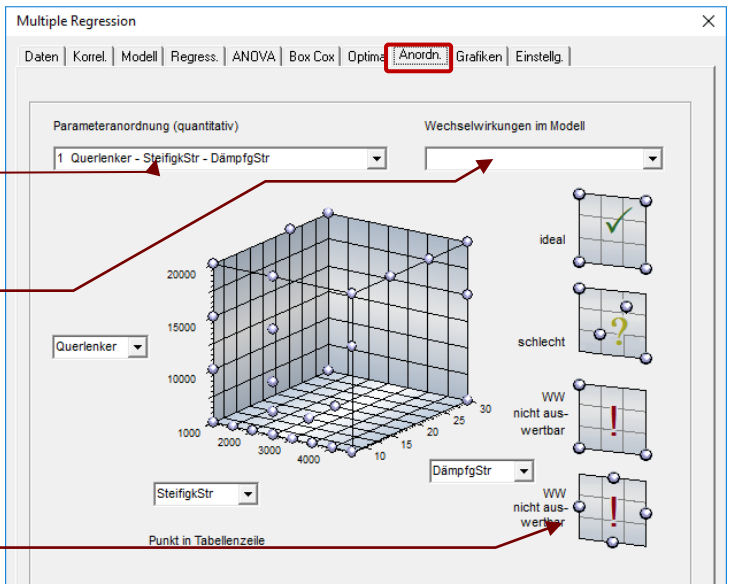


Zum Starten der optimalen Einstellungen, wählen Sie Start. Bei mehreren Zielgrößen werden zunächst die Einzeloptima aufgelistet und das gemeinsame Optima als Kompromis, berechnet durch Minimierung der Abweichungen im Quadrat zu den Einzeloptima.

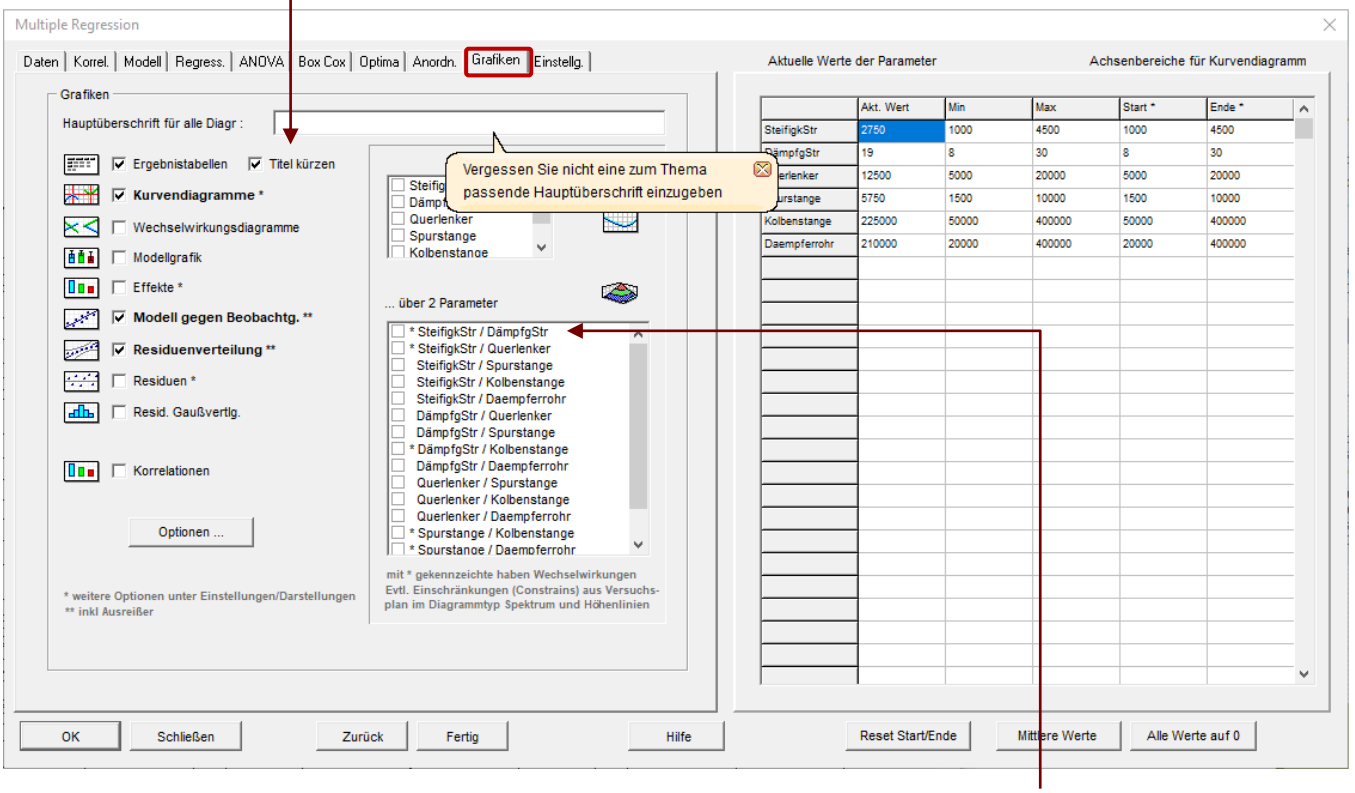
Unter Rubrik *Anordnung* lassen sich jeweils 3 quantitative Parameter als 3D-Grafik darstellen.
Alle möglichen 3fach-Kombinationen können in der oberen Auswahl gewählt werden.

Alternativ lassen sich die im Modell befindlichen 2-fach-Wechselwirkungen auswählen.

Stammen die Daten nicht aus einem Versuchsplan, so kann hier leicht beurteilt werden, wo nachträgliche Versuche noch durchzuführen sind, nämlich insbesondere an den Ecken. Günstige und kritische Verteilungen in den Ansichten sind auf der rechten Seite dargestellt.



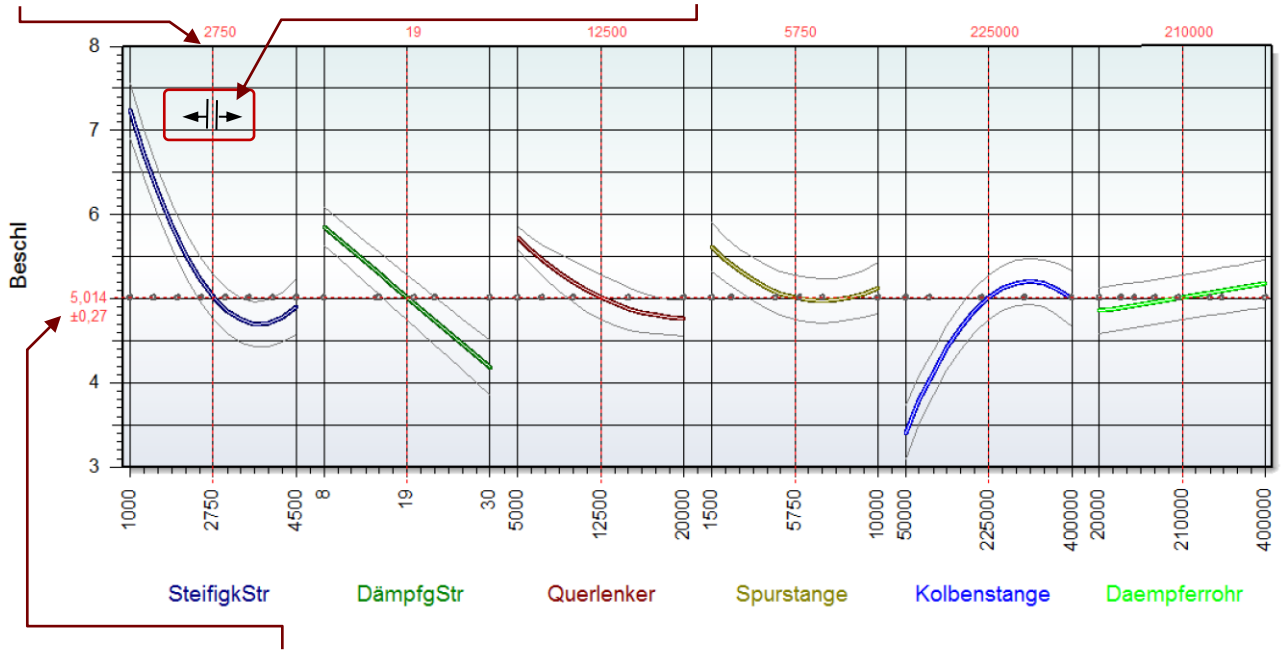
Für die **grafische Ausgabe** der Ergebnisse empfehlen sich insbesondere die 3 fett gekennzeichneten Diagramme. Die Ergebnistabelle wird im Hauptfenster und in der Tabelle (MRKoeff1) ausgegeben. Aus Platzgründen werden Titel, die über 12 Zeichen lang sind gekürzt wenn hier der Haken gesetzt ist.



Die Grafiken *über 2 Parameter* rechts sind 3-Diagramme. Diejenigen mit einem * sind Wechselwirkungen im Modell. Nur für diese ist eine Diagrammdarstellung interessant, siehe Kapitel *Grafische Darstellung der Wechselwirkung* unter den Grundlagen.

Das wohl wichtigste Diagramm für die Visualisierung des Regressionsmodells ist das Kurvendiagramm, das ebenfalls in den Grundlagen am Anfang bereits beschrieben wurde.

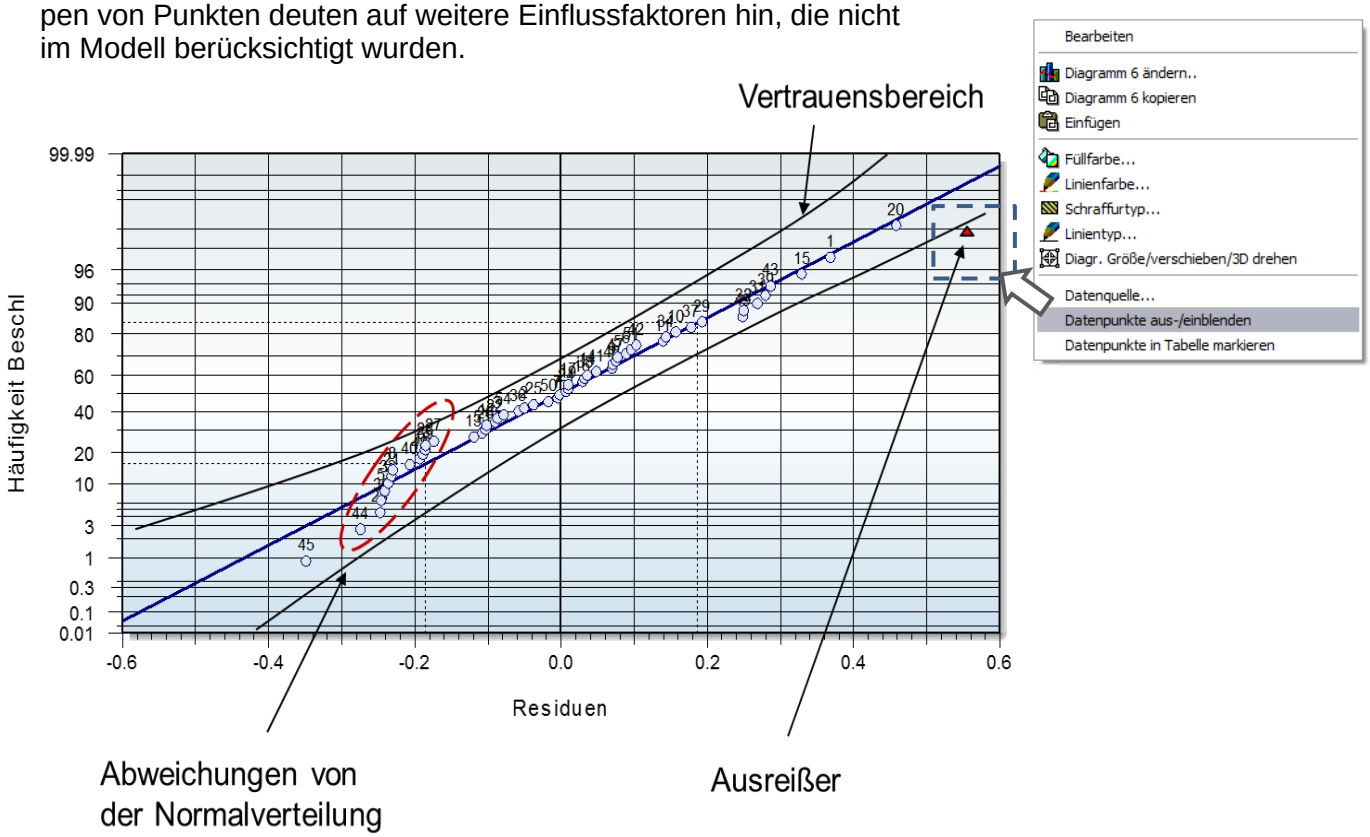
Die vertikalen roten Linien stellen die momentanen Einstellungen dar, mit ihren Werten über dem Diagramm. Die aktuellen Einstellungen können mit der Maus verändert werden.



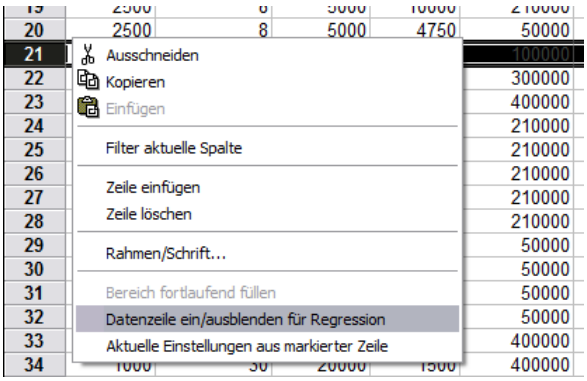
Die waagrechte rote Line zeigt das Modellergebnis mit dem konkreten Zahlenwert und seinem Vertrauensbereich. Hinweis: Stammen die Daten aus einem Experiment mit Einschränkungen (Constrains), so können evtl. nicht alle Einstellungen unabhängig voneinander variiert werden.

Residuenverteilung

Das Beispiel zeigt Abweichungen von der Normalverteilung der Residuen sowie rechts einen Ausreißer. Systematische Abweichungen bei ganzen Gruppen von Punkten deuten auf weitere Einflussfaktoren hin, die nicht im Modell berücksichtigt wurden.



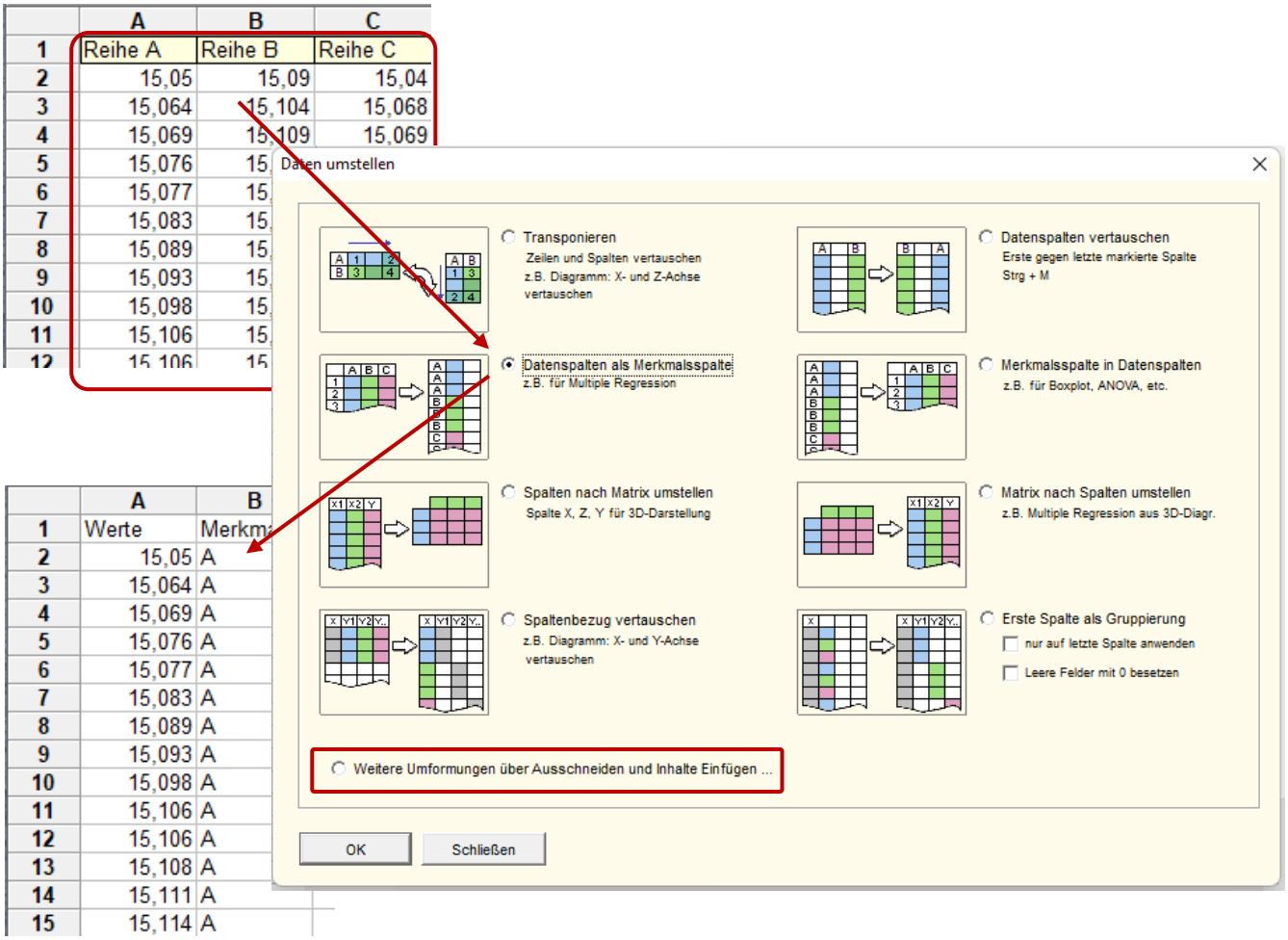
Ausreißer, aber auch beliebige Punkte, können durch Anklicken, oder durch „Einfangen“ mit der Maus ausgeblendet werden. Diese Punkte sind dann in der Tabelle ausgegraut, ebenso im Diagramm *Modell gegen Beobachtung*. In der Residuenverteilung sind herausgenommene Punkte nicht mehr sichtbar, da sie für die Verteilung nicht herangezogen werden dürfen. Nach dem Ausblenden wird das Dialogfenster der Regression neu geöffnet, denn es kann sein, dass bestimmte Terme nicht mehr signifikant sind, andere evtl. wieder ins Modell hineingenommen werden können. Um das Modell an die geänderten Daten anzupassen, sollte wieder die Taste *Auto* verwendet werden. Möchte man die Punkte wieder zurückholen, kann man die entsprechende Zeile in der Tabelle markieren und diese wieder einblenden. Gesamthaft ist das für mehrere ausgeblendete Punkte auch über den Menüpunkt *Statistik/ Datenauswertung /Alle Datenzeilen wieder zurück ins Modell* möglich.



Das Markieren der Punkte und das Ausblenden ist auch im Diagramm *Modell geben Beobachtung* möglich. Weitere Diagramme sind unter Grundlage im mittleren Teil erläutert.

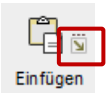
Vorbereitung der Daten

Liegen z.B. die Datenreihen in Spalten vor, müssen diese als Merkmalsspalte umgestellt werden, wie es für die Multiple Regression erwartet wird. Die entsprechenden Daten müssen vorher markiert sein. Hierfür ist in Version 17.0 der Menüpunkt *Bearbeiten / Daten umstellen* und in Version 20.0 der Menüpunkt *Daten / Daten umstellen* aufzurufen.



	A	B	C
1	Reihe A	Reihe B	Reihe C
2	15,05	15,09	15,04
3	15,064	15,104	15,068
4	15,069	15,109	15,069
5	15,076	15,	
6	15,077	15,	
7	15,083	15,	
8	15,089	15,	
9	15,093	15,	
10	15,098	15,	
11	15,106	15,	
12	15,106	15,	

	A	B
1	Werte	Merkm
2	15,05	A
3	15,064	A
4	15,069	A
5	15,076	A
6	15,077	A
7	15,083	A
8	15,089	A
9	15,093	A
10	15,098	A
11	15,106	A
12	15,106	A
13	15,108	A
14	15,111	A
15	15,114	A



Befinden sich die Daten vorher z.B. in Excel, so ist *Bearbeiten / Inhalt* einfügen zu verwenden, was die gleichen Funktionen zum Umformen bietet.

Expertenwissen

Expert

Ergebnistabelle im Hauptfenster

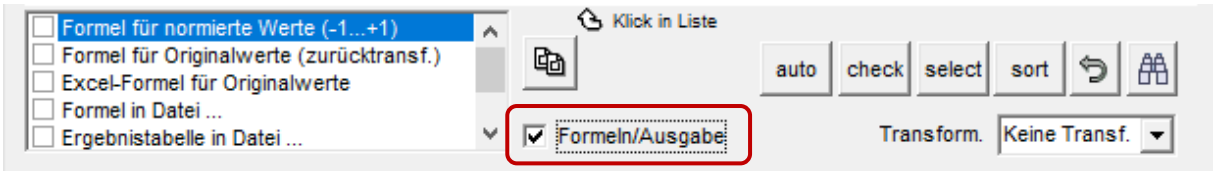
Ist unter dem Reiter Grafiken die Ergebnistabelle gewählt, werden folgende Daten ausgegeben:

Y	Koeffizient	StdAbw	t-Wert	p-value	VIF	Min	Max	AktWert	Transf	Normiert
Constant	1,134615									
X1	2,115385	0,262273	8,065591	0,001	1,0	1	2	1,5	Keine	((X1-1,5)/0,5)
X2	2,865385	0,262273	10,92521	0,000	1,0	3	5	4	Keine	((X2-4)/1)
X3	4,115385	0,262273	15,69124	0,000	1,0	4	6	5	Keine	((X3-5)/1)
X1*X2	4,884615	0,262273	18,62418	0,000	1,0					

R^2=0,995	DF=4	
Radj²=0,989	RMS=0,7721	RMS/Ym=1,2

Zielgröße-Transf Keine

Wenn im Dialogfenster der Regression unter *Einstellungen / Regression* die Parameter Skalierung normiert ist (Standardeinstellung), so beziehen sie die Koeffizienten nicht auf die Originalwerte, in der Ergebnistabelle als Min / Max zu sehen, sondern auf -1 .. +1. So ist es auch in der Formelausgabe der Fall, wenn die erste Option gewählt wird:



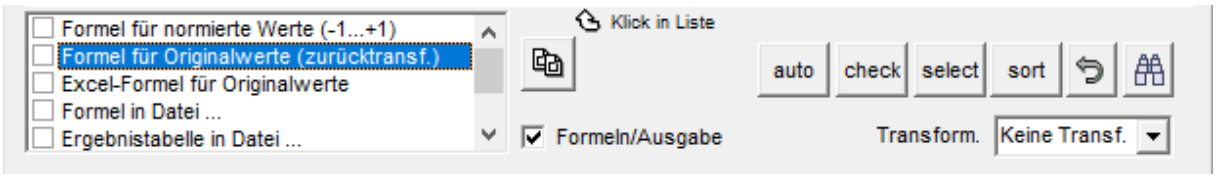
Y=1,134615+2,115385*X1+2,865385*X2+4,115385*X3+4,884615*X1*X2

Möchte man diese Formel später verwenden, so sind nicht die Originaleinheiten zu verwenden, sondern vorher die Normierung zu berücksichtigen. Beispiel für den Faktor X2. Sein Wertebereich ist zwischen 3 und 5. Soll nun die Formel für X2=3 berechnet werden, so ist hier

((X2 - 4)/1) = (3 - 4)/1 = -1

zu verwenden.

Wählt man gleich die Formel für die Originalwerte, ist diese Umrechnung hier bereits enthalten

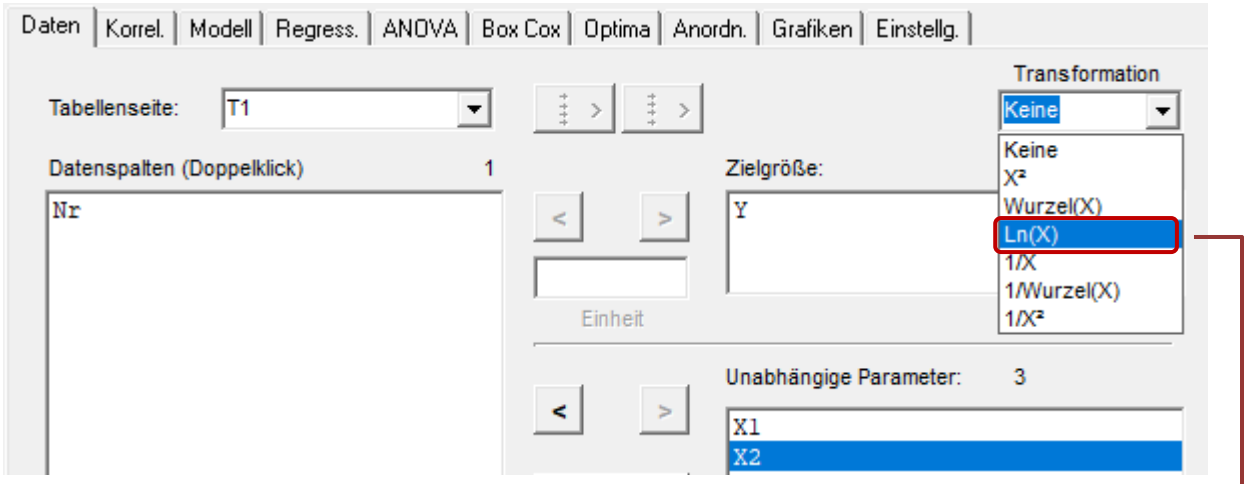


Was jedoch zu einer wesentlich längeren Formel führt:

Y=1,134615+2,115385*((X1-1,5)/0,5)+2,865385*((X2-4)/1)+4,115385*((X3-5)/1)+4,884615*((X1-1,5)/0,5)*((X2-4)/1)

In der Excel – Formel ist diese Rücktransformation immer enthalten.

Sind die Faktoren transformiert, z.B. X2 mit dem Ln unter dem Reiter Daten

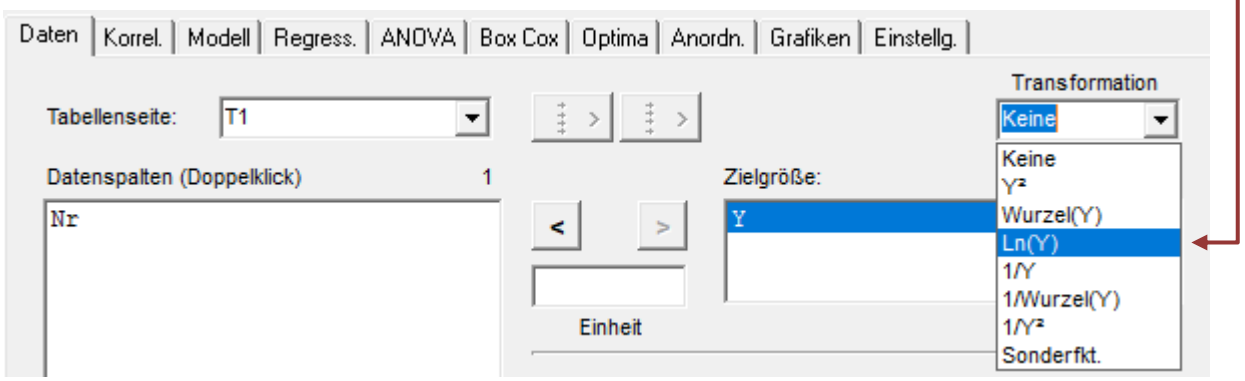


so kommt diese math. Umformung noch dazu:

Y=1,134615+2,115385*((X1-1,5)/0,5)+2,865385*((**ln(X2)**-1,354025)/0,255413)+4,115385*((X3-5)/1)+4,884615*((X1-1,5)/0,5)*((ln(X2)-1,354025)/0,255413)

wobei sich die Werte für die Normierung mit ändern.

Ist die Zielgröße transformiert, z.B. auch hier der ln



so wird hier in der Formel immer die Umkehrfunktion eingesetzt, denn die Regression wird mit den logarithmierten Werten durchgeführt und das Ergebnis muss sich wieder auf die Originaleinheiten beziehen:

Y=e^(…)

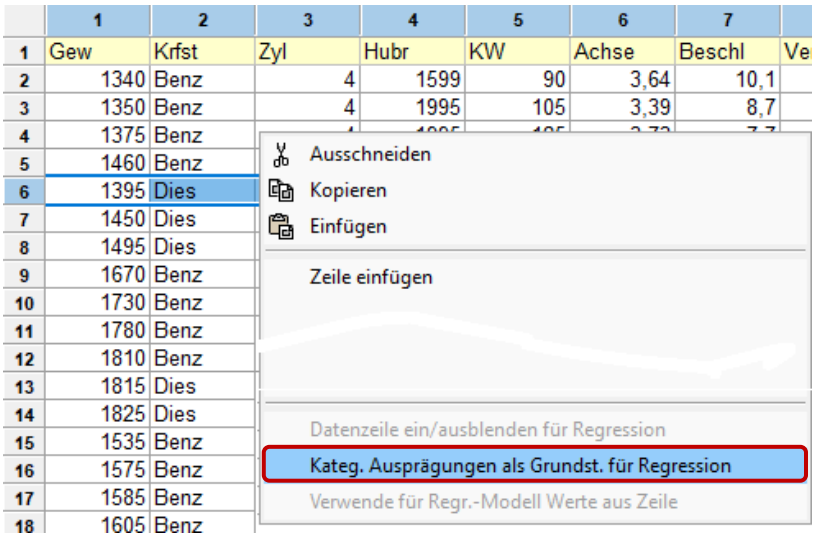
In diesem Beispiel ist allerdings die ln-Transformation nicht möglich, da die Y-Werte auch negativ sind.

Ausprägungen von kategorialen Faktoren bzw. Parametern

Bei kategorialen Faktoren bzw. Parametern wird mathematisch die „Grundstufe“ nicht als Term im Modell dargestellt. Es erscheinen nur die weiteren Ausprägungen bzw. Varianten in der Modellgleichung. Visual-XSel betrachtet die erste Ausprägung in der jeweiligen Datenspalte der ersten Zeile als Grundstufe. Beispiel Spalte „Krfst“ in der Beispieldatei *Beispiel_Verbrauch.vxgn* (Menüpunkt *Datei / Öffnen Daten, Mess-/Prozess, Beispieldaten*…)

	1	2	3	4	
1	Gew	Krfst	Zyl	Hubr	KW
2	1340	Benz	4	1599	
3	1350	Benz	4	1995	
4	1375	Benz	4	1995	
5	1460	Benz	6	2996	
6	1395	Dies	4	1995	
7	1450	Dies	4	1995	
8	1495	Dies	4	1995	
9	1670	Benz	4	1995	
10	1730	Benz	6	2996	

Hier ist demnach „Benz“ die Grundstufe und der Effekt wird bestimmt durch die Differenz zum Verbrauch bei „Dies“. Es ergibt sich, wie zu erwarten ist, eine Verbrauchsreduzierung. Möchte man aber den Diesel als Ausgangspunkt verwenden und die Verbrauchserhöhung beim Umstieg auf den Benziner wissen, so muss eine Datenzeile mit „Dies“ in der ersten Datenzeile stehen (bzw. Zeile 2 im Excel-Datenblatt). Es ist also hierzu diese Zeile 2 mit einer Zeile mit „Dies“ manuell zu vertauschen. Ab Version 20.0004 gibt es hierzu einen einfachen Menüpunkt über die rechte Maustaste. Vorher ist die gewünschte Zeile links anzuwählen, bzw. zu markieren.
Dies kann mehrfach hintereinander gemacht werden, um bei mehr als 2 Ausprägungen noch eine gewünschte Reihenfolge festzulegen, was für das spätere Kurvendiagramm sinnvoll sein kann.



	1	2	3	4	5	6	7	
1	Gew	Krfst	Zyl	Hubr	KW	Achse	Beschl	Ve
2	1340	Benz	4	1599	90	3,64	10,1	
3	1350	Benz	4	1995	105	3,39	8,7	
4	1375	Benz						
5	1460	Benz						
6	1395	Dies						
7	1450	Dies						
8	1495	Dies						
9	1670	Benz						
10	1730	Benz						
11	1780	Benz						
12	1810	Benz						
13	1815	Dies						
14	1825	Dies						
15	1535	Benz						
16	1575	Benz						
17	1585	Benz						
18	1605	Benz						

Numerische Datenspalten als kategorial deklarieren

Die Spalte „Zyl“ ist ganzzahlig und wird von Visual-XSel dadurch mathematisch als stetige numerisch Größe betrachtet, selbst dann, wenn die Zellen als Text formatiert wurden. Da aber hinter der Zylinderzahl technisch noch mehr steckt, als nur die reine Zahl (der 8 Zylinder ist ein V-Motor, der viel komplexer aufgebaut ist), könnte es hier Sinn machen, die Zylinderzahl als Kategorie zu definieren.

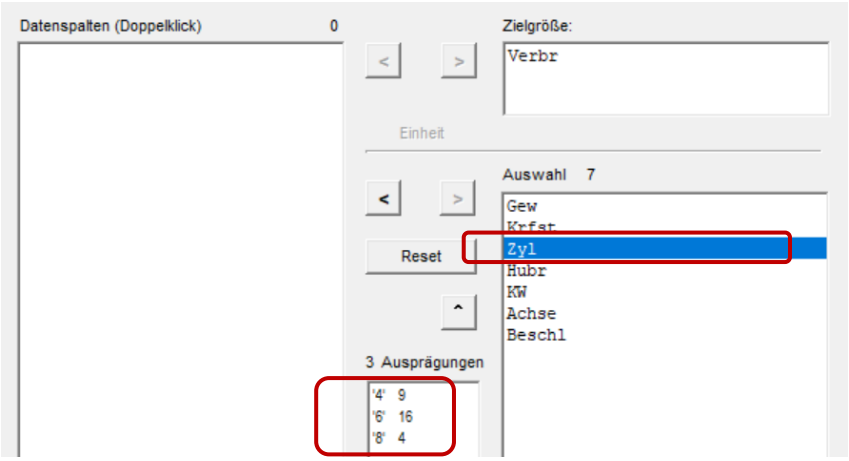
2	3	4	5
Krfst	Zyl	Hubr	KW
Benz	4	1599	90
Benz	4	1995	105
Benz	4	1995	125
Benz	6	2996	195
Dies	4	1995	105
Dies	4	1995	130
Dies	4	1995	150
Benz	4	1995	125
Benz	6	2996	160
Benz	6	2996	200
Benz	6	2979	225

Hierzu kann man nach der Auswahl der Datenspalten Zyl rechts anzuklicken und in der Mitte „Set kategorial“ drücken. Hierdurch werden die Zahlen in Anführungszeichen gesetzt, sodass sie als Text interpretiert werden.

Es erscheinen nun die Ausprägungen in der Mitte, mit der jeweiligen Anzahl.

Ist die Reihenfolge der Kategorien nicht wie gewünscht, kann entsprechend dem vorhergehenden Abschnittes vorgegangen werden.

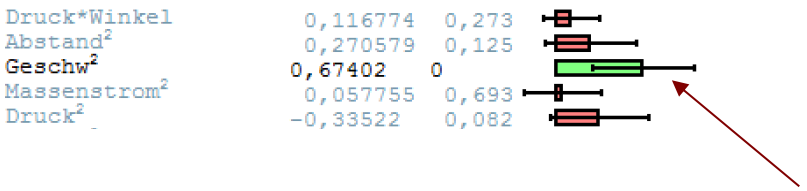
Achtung: Man kann diese Umformung in der Tabelle rückgängig machen. Das Modell muss dann jedoch komplett zurückgesetzt werden.

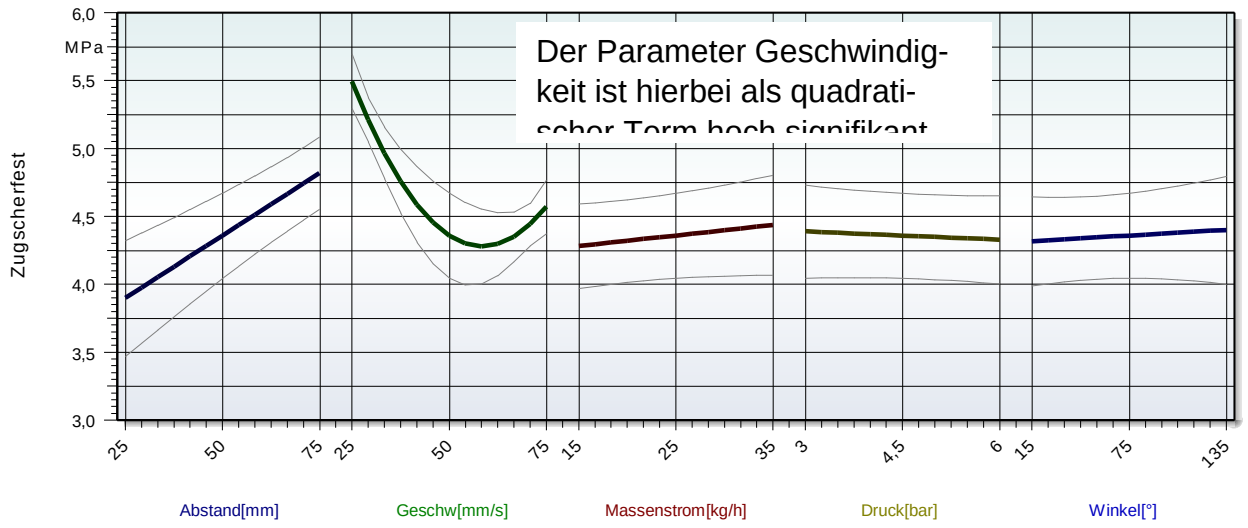


Pseudoquadratische Modelle

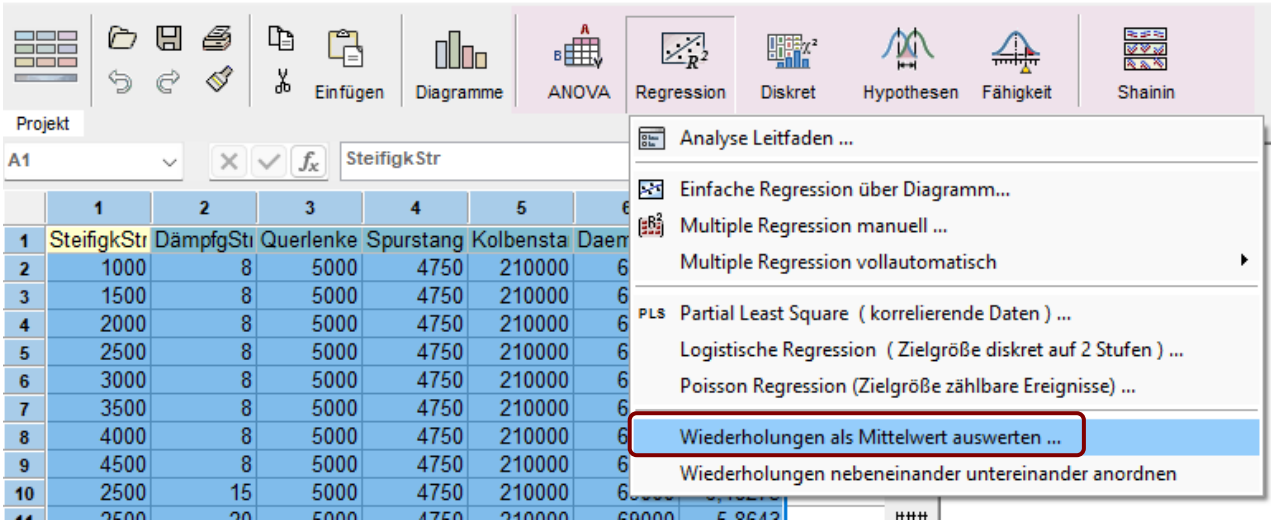
In einer anderen Auswertung *Beispiel_Zugfestigkeit.vxg* wurde ein Versuchsplan mit 3 Wiederholungen durchgeführt. Die Auswertung erfolgte hier über jeden Einzelmesswert (je 3 Messwerte stehen für identische Parametereinstellungen in der Tabelle untereinander).

	A	B	C	D	E	F	G
1		stand[mm]	hw[mm/s]	trom[kg/h]	Druck[bar]	Winkel[°]	rfest[MPa]
2		75	75	15	6	90	3,559
3		75	75	15	6	90	3,533
4		75	75	15	6	90	4,314
5		25	75	15	3	90	4,163
6		25	75	15	3	90	3,472
7		25	75	15	3	90	3,725



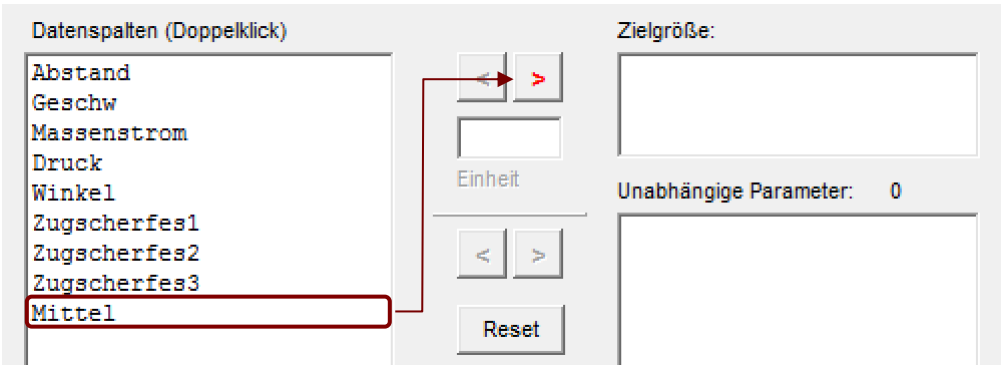


Eine zweite Auswertung auf Basis der Mittelwerte jeder Wiederholung soll zeigen, ob der quadratische Ansatz für die Geschwindigkeit Bestand hat. Durch die Mittelwertbildung reduzieren sich Messfehler. Es sind alle Spalten inkl. der Zielgröße zu markieren und die Ikone Auswertung und der folgende Menüpunkt auszuwählen:

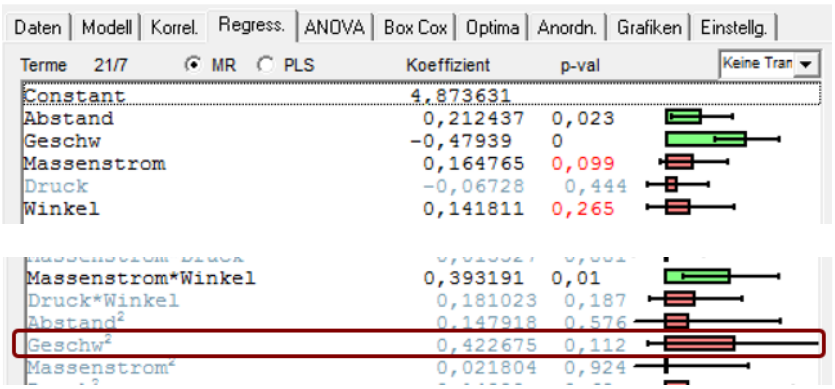


	1	2	3	4	5
1	SteifigkStr	DämpfgStr	Querlenke	Spurstang	Kolbensta
2	1000	8	5000	4750	210000
3	1500	8	5000	4750	210000
4	2000	8	5000	4750	210000
5	2500	8	5000	4750	210000
6	3000	8	5000	4750	210000
7	3500	8	5000	4750	210000
8	4000	8	5000	4750	210000
9	4500	8	5000	4750	210000
10	2500	15	5000	4750	210000

Hinweis: Der Menüeintrag „Wiederholungen…“ erscheint nur, wenn nicht bereits eine Auswertung durchgeführt wurde (hier dann Menüpunkt „Modell verwerfen“ nutzen).
In Version 17.0 nutzen Sie den Menüpunkt *Statistik / Datenauswertung*.
Als Zielgröße ist jetzt der Mittelwert zu verwenden:



Hier zeigt sich, dass der quadratische Term für die Geschwindigkeit nicht mehr signifikant ist und offensichtlich nur durch einzelne Messfehler zu Stande kam.

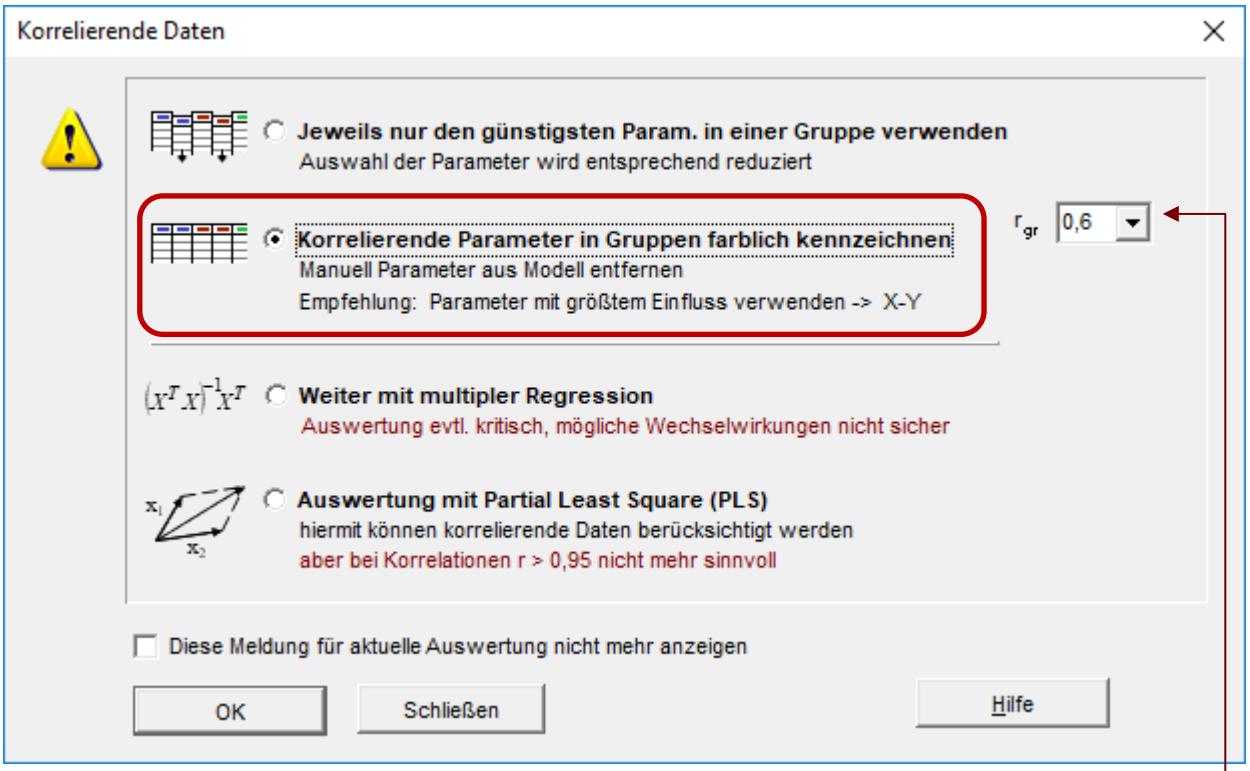


Probleme mit korrelierenden Daten

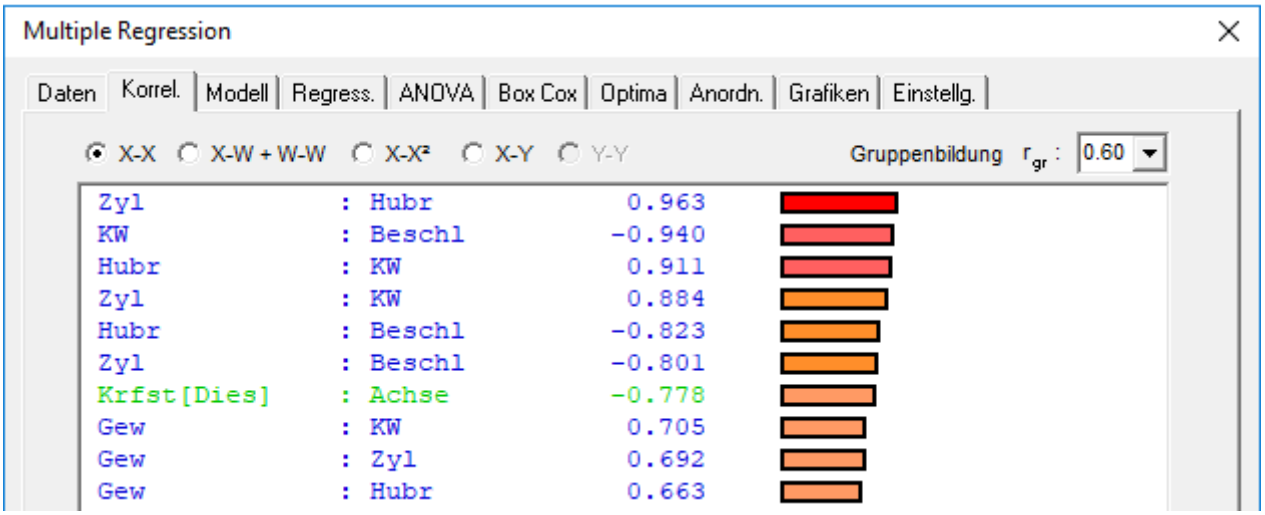
Wenn die auszuwertenden Daten nicht einer Versuchsplanung stammen, können die Parameterspalten korrelieren, was für die multiple Regression u.U. kritisch werden kann, insbesondere für die Bewertung von Wechselwirkungen. Die folgende Beschreibung zeigt die Vorgehensweise am Datensatz *…./Beispieldaten/Beispiel_Verbrauch.vxt*. Wie zu erwarten ist, hängen insbesondere die Zylinderzahl, Hubraum und Leistung zusammen.

	A	B	C	D	E	F	G	H
1	Gew	Krfst	Zyl	Hubr	KW	Achse	Beschl	Verbr
2	1340	Benz	4	1599	90	3,64	10,1	5,8
3	1350	Benz	4	1995	105	3,39	8,7	5,9
4	1375	Benz	4	1995	125	3,73	7,7	6,4
5	1460	Benz	6	2996	195	3,46	6	8,3
6	1395	Dies	4	1995	105	3,07	8,9	4,5
7	1450	Dies	4	1995	130	2,56	7,5	4,8

Nach Auswahl der Parameterspalten A-G und der Zielgröße Verbrauch erscheint bei Aufruf der Rubrik ***Korrelation*** folgende Meldung:

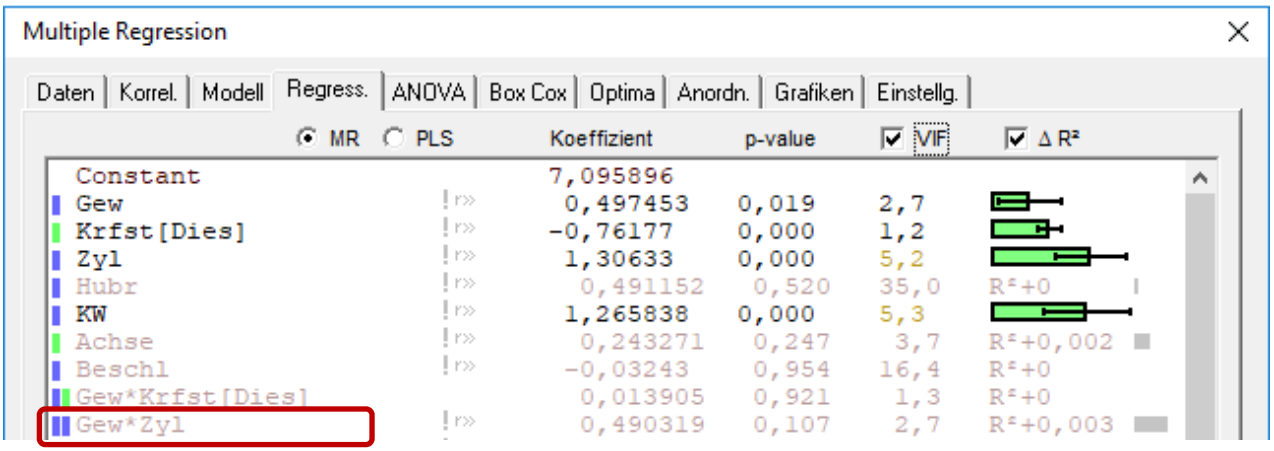


Bei diesem Datenbeispiel sollte die Korrelationsgrenze r_{gr} =0,6 anstelle von 0,9 gewählt werden. Ansonsten würden keine Gruppen gefunden werden.
Bestätigen Sie die Empfehlung, die korrelierenden Parameter in Gruppen farblich zu kennzeichnen. Es gibt hier zwei Gruppen, die blaue, die mit der Leistung zusammenhängt und die grüne, da Dieselfahrzeuge eine längere Hinterachsübersetzung als Benzinfahrzeuge haben (Achse).



Die Gruppen werden hier auf Basis der Korrelationsmatrix und einer Art Clusteranalyse gebildet. Das Limit, ab welcher Gruppen zusammengefasst werden ist hier auf r=0,60 festgelegt und kann frei geändert werden (oben rechts).

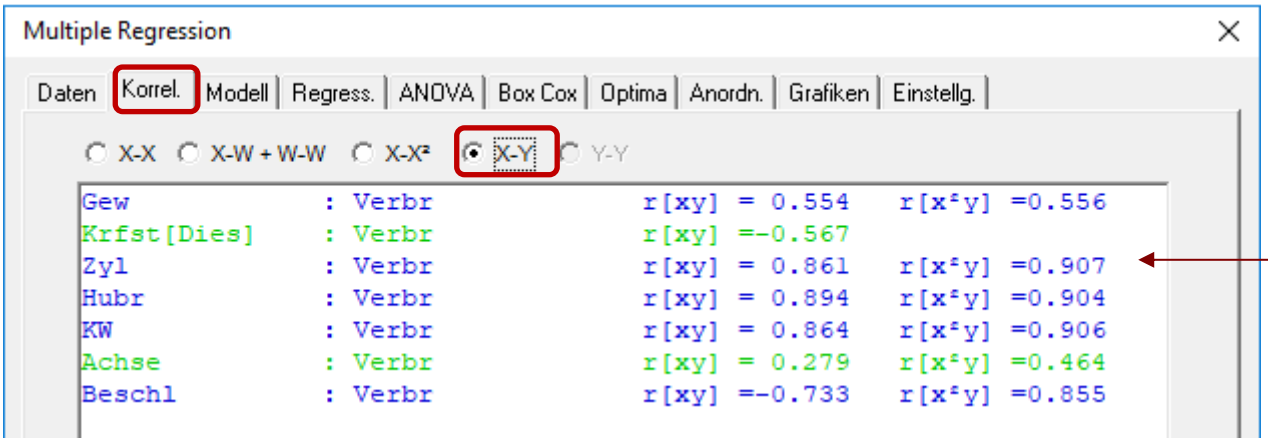
Geht man zunächst auf den Reiter Regression, so sind die Gruppen hier durch Farbbalken gekennzeichnet.



Wechselwirkungen sind bei korrelierenden Daten immer kritisch. Man sollte deshalb nie eine Wechselwirkung in das Modell mit aufnehmen, bei der beide Parameter aus derselben Gruppe stammen. Das ist hier bei Gew*Zyl der Fall. Das gezeigte Modell wurde durch ein quadratisches Modell mit Wechselwirkungen und anschließender „Modellbereinigung“ über die Taste *auto* erstellt. Hierdurch werden nie Wechselwirkungen mit in das Modell genommen, selbst wenn sie signifikant wären.

Obwohl das gezeigt Beispiel trotz hoher Korrelationen ein plausibles Modell ergibt, ist es auch sinnvoll, die Auswertung mit nur einem Parameter aus jeder Gruppen neu aufzubauen, insbesondere wenn wenig Hintergrundwissen über die Zusammenhänge besteht. Die jeweils anderen Parameter werden in Abhängigkeit der Korrelation durch diese zum großen Teil ausgedrückt. Die Frage ist nur für welchen Parameter man sich jeweils entscheidet. Es sollte derjenige sein, der am besten mit der Zielgröße zusammenhängt und zwar nicht nur linear, sondern falls möglich, auch als quadratisches Modell.

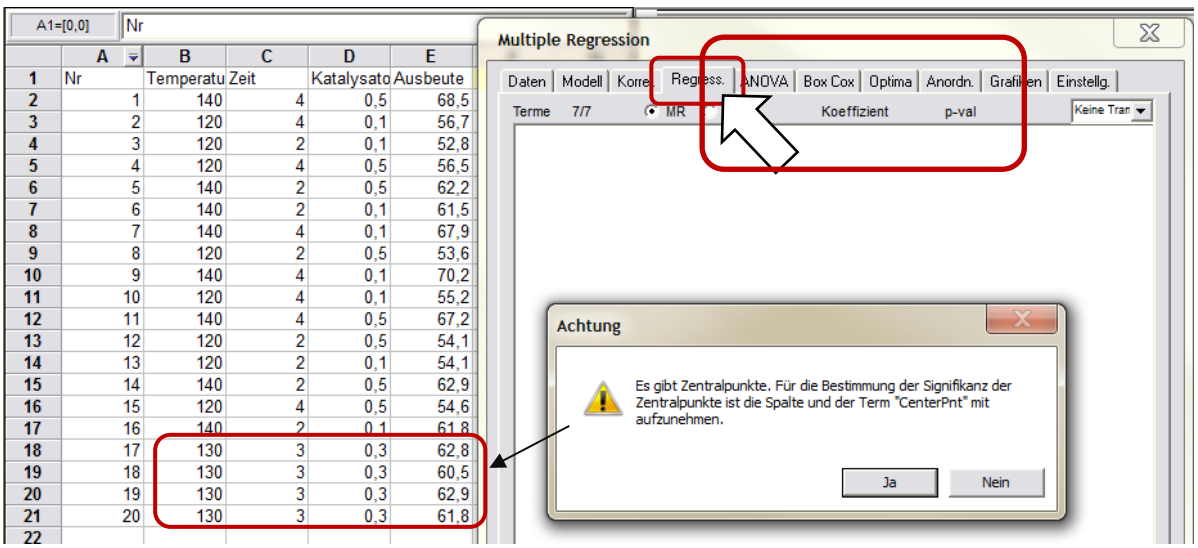
Eine Übersicht dieser Zusammenhänge ist über die Option X-Y möglich.



Die Liste zeigt, dass ein quadratischer Zusammenhang von Zyl am meisten die Zielgröße Verbrauch beschreibt, da die Korrelation hier am höchsten innerhalb der blauen Gruppe ist. Dabei gibt es einen nichtlinearen Zusammenhang, was durch [x²y] ausgedrückt wird. Die Leistung KW hat aber einen ähnlich hohen Wert. Aufgrund der „feiner" abgestuften Zahlenwerte von KW wird man sich evtl. deshalb eher hierfür entscheiden. Der nächste Schritt ist nun zurück auf die Rubrik ***Daten*** zu gehen und unter unabhängige Parameter eine reduzierte Auswahl zu verwenden. Hinweis: Aufgrund der deutlich geringen Korrelation von Gewicht zu KW, Zyl und Hubr soll dieser als zweiter Parameter mit verwendet werden. Es sind also von 7 Parametern hier nur noch KW, Krfst und Gew zu verwenden.

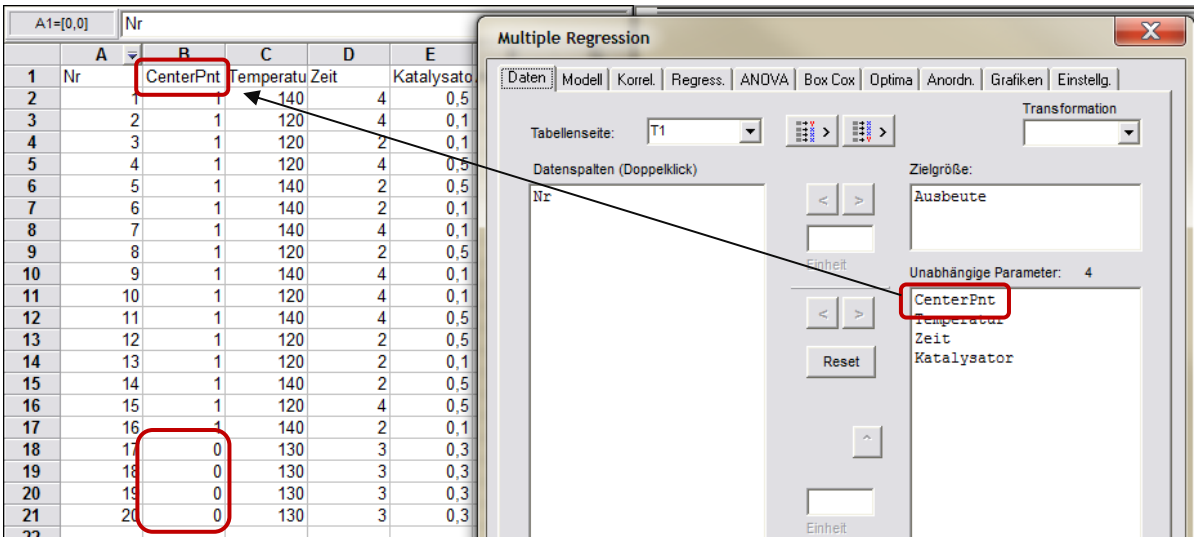
Prüfung der Nichtlinearität über Zentralpunkte

Für ein lineares Modell kann über Zentralpunkte eine evtl. Nichtlinearität geprüft werden. Im Beispiel für Zentralpunkte aus dem Taschenbuch Versuchsplanung von Prof. Kleppmann gibt es 4 Zentralpunkte. Beim erstmaligen Aufruf der Regression werden diese ermittelt und es folgt eine Abfrage, ob eine zusätzliche Auswertung über eine Kennzeichnungsspalte „CenterPnt" erfolgen soll.



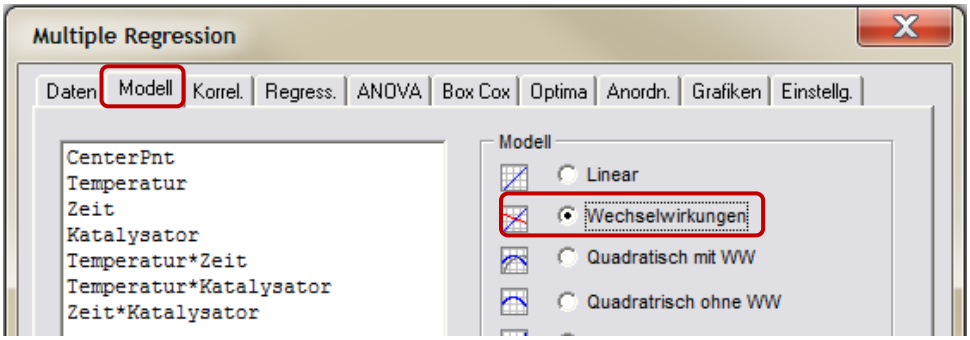
	A	B	C	D	E
1	Nr	Temperatu	Zeit	Katalysato	Ausbeute
2	1	140	4	0,5	68,5
3	2	120	4	0,1	56,7
4	3	120	2	0,1	52,8
5	4	120	4	0,5	56,5
6	5	140	2	0,5	62,2
7	6	140	2	0,1	61,5
8	7	140	4	0,1	67,9
9	8	120	2	0,5	53,6
10	9	140	4	0,1	70,2
11	10	120	4	0,1	55,2
12	11	140	4	0,5	67,2
13	12	120	2	0,5	54,1
14	13	120	2	0,1	54,1
15	14	140	2	0,5	62,9
16	15	120	4	0,5	54,6
17	16	140	2	0,1	61,8
18	17	130	3	0,3	62,8
19	18	130	3	0,3	60,5
20	19	130	3	0,3	62,9
21	20	130	3	0,3	61,8
22					

Bestätigen Sie diese Abfrage mit Ja und es wird eine neue Spalte eingefügt:

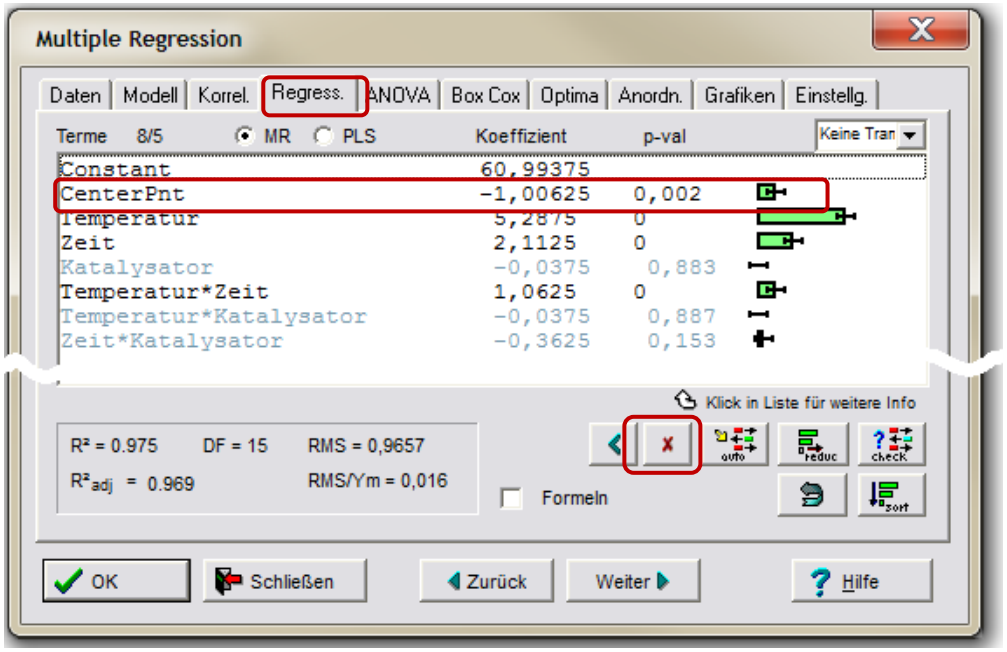


Überall, wo eine 0 vorkommt, gibt es Zentralpunkte. Übernehmen Sie CenterPnt als Parameter für die spätere Auswertung.

Unter dem Reiter Modell werden über die Auswahl Wechselwirkungen keine Paare zwischen CenterPnt und den anderen Termen angelegt. Der Titel CenterPnt sollte deshalb nicht geändert werden.



Nach Auswahl der Taste Auto bleibt der CenterPnt signifikant.



Es ist deshalb also davon auszugehen, dass das Modell nichtlinear ist. Für jeden Parameter müssen in weiteren Versuchen getrennt quadratische Zusammenhänge untersucht werden. In dieser späteren Auswertung braucht der Termin CenterPnt dann nicht mehr im Modell mit aufgenommen werden. Die Auswertung des CenterPnt's ist nur für das lineare Modell sinnvoll.

Hinweise:

In anderen Programmen wird als Constant der Werte ausgegeben, der hier die Summe von Constant + CenterPnt darstellt.

Fazit:

Die statistischen Kenngrößen, insbesondere der p-value sind ein Hilfsmittel zur Entscheidung, welche Modellterme im Modell bleiben und welche nicht. Die p-values sind aber keine Garantie dafür, dass die wirklichen Zusammenhänge so sind. Sehr wichtig ist vor allem das technische Wissen, um die Zusammenhänge zu Plausibilisieren.

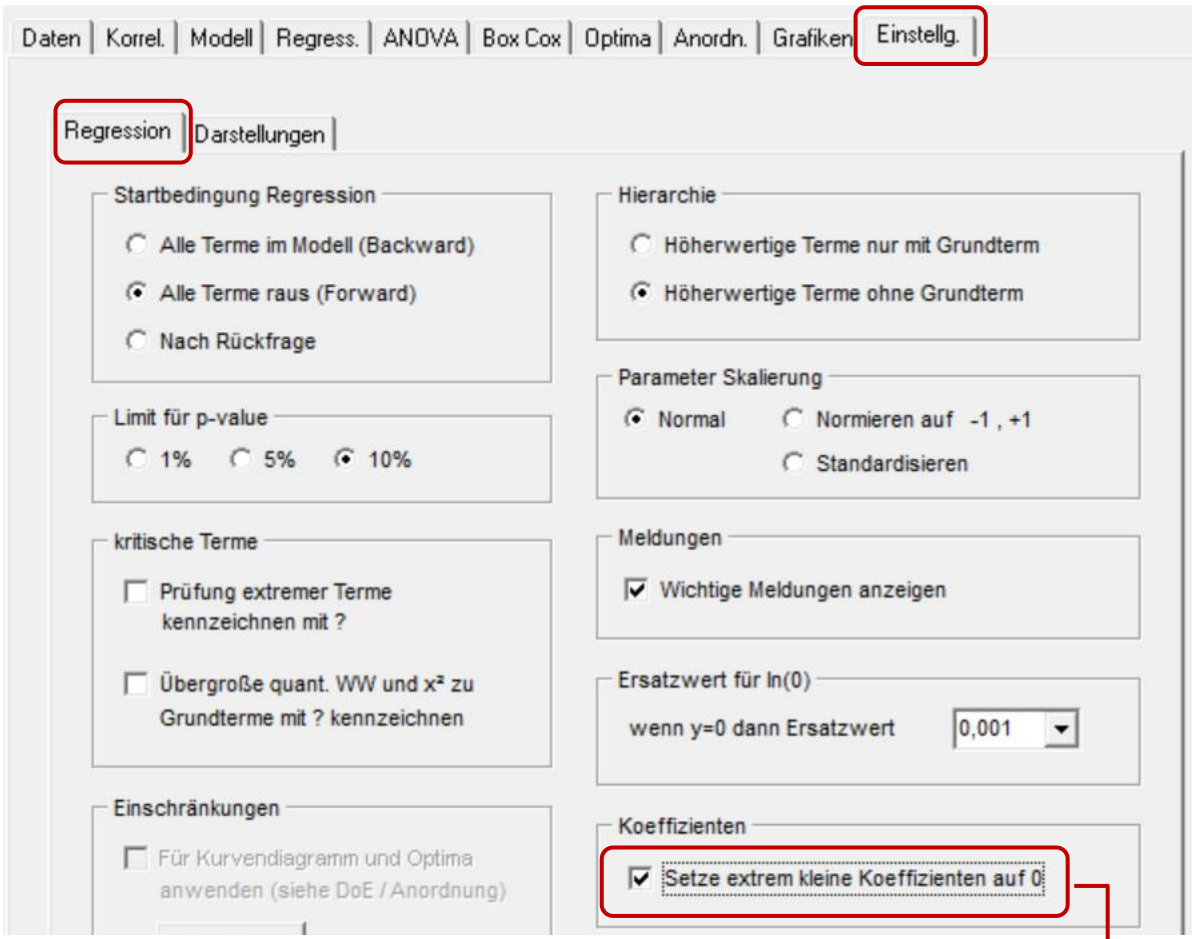
Ergänzende Versuche zur Bestätigung von optimalen Einstellungen sind unbedingt zu empfehlen.

Extrem kleine Koeffizienten

Gibt es gegenüber anderen Parametern welche mit extrem kleinen Koeffizienten,

Constant	1	
A	1,054E-8	0,027
B	1,054E-8	0,027
C	2	0,000
A*B	3	0,000
A*C	-6,14E-10	0,917
B*C	-6,14E-10	0,917

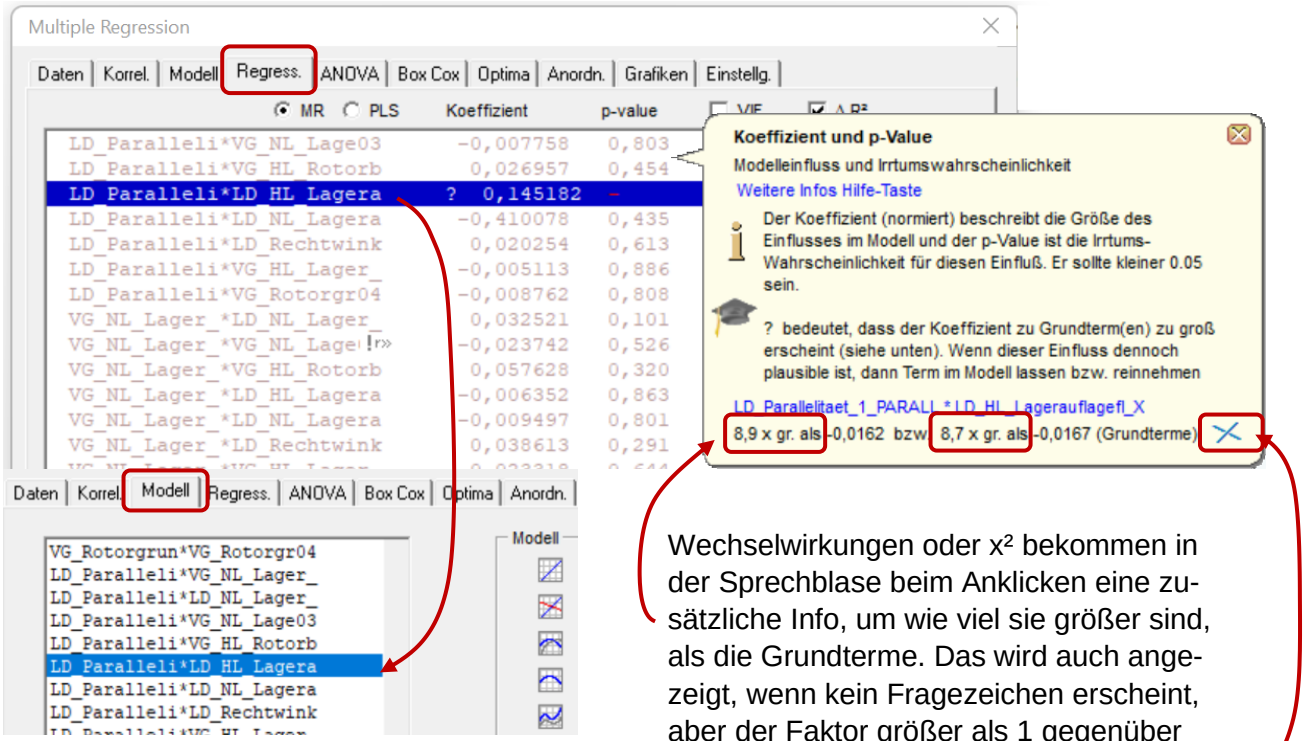
so können diese für eine bessere Übersichtlichkeit auf 0 gesetzt werden, was auch die Standardeinstellung ist:



Constant	1	
A	0	0,027
B	0	0,027
C	2	0,000
A*B	3	0,000
A*C	0	0,917
B*C	0	0,917

Weitere Funktionalitäten

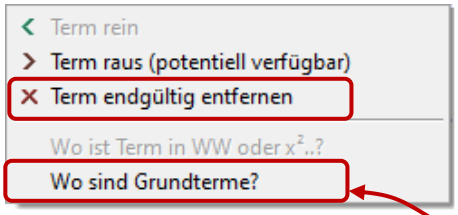
Angeklickte Terme unter dem Reiter Regression sind auch unter Reiter Modell ausgewählt

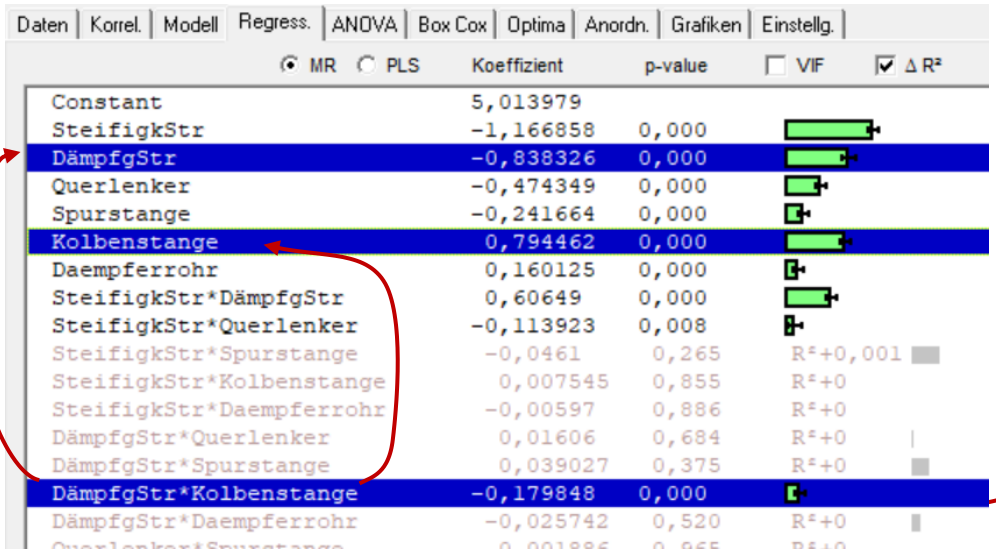


Wechselwirkungen oder x² bekommen in der Sprechblase beim Anklicken eine zusätzliche Info, um wie viel sie größer sind, als die Grundterme. Das wird auch angezeigt, wenn kein Fragezeichen erscheint, aber der Faktor größer als 1 gegenüber den Grundtermen ist. Außerdem erscheint ein Symbol für große WW

Unter Modell sind die gleichen Terme mit den gekürzten Namen gewählt.

Mit Hilfe der rechten Maustaste und dem neuen Popup-Menü können auch im Reiter Regression höherwertige Terme endgültig aus dem Modell gelöscht werden. Dies kann sinnvoll sein, wenn Terme keine Freiheitsgrade haben oder grundsätzlich nicht als real angesehen werden.
Mit Hilfe der **rechten Maustaste** kann nun angezeigt werden, wo sich die Grundterme befinden





Mit dem darüberliegenden Menüpunkt, kann bei Anklicken eines Grundterms gezeigt werden, in welchen WW- oder x^2-Termen es vorkommt.