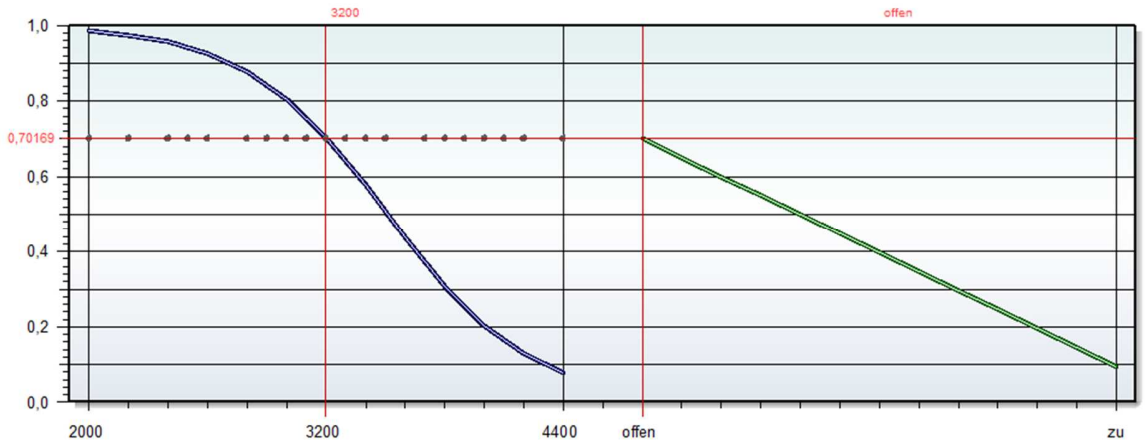
Diskrete Regression

Das Logit Modell



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/Multiple-Regression.pdf

www.versuchsmethoden.de/Poisson-Regression.pdf

Keywords:

Diskrete, logistische, Regression, Logit Modell, Logits, gut, schlecht, Maximum Likelihood, Wahrscheinlichkeit, Devianz-Test, pseudo R^2

Einführung

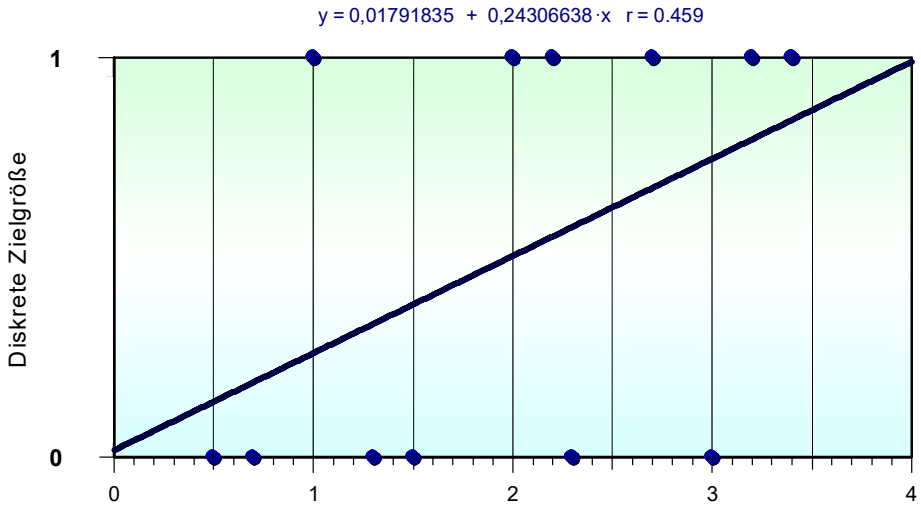
Unter einer diskreten oder logistischen Regression versteht man eine Auswertung mit Zielgrößen, die keinen stetigen Messwert, sondern qualitativen Charakter haben. Beispielsweise könnte das Ergebnis einer Untersuchung nur mit „gut“ oder „schlecht“ beurteilt werden, wie Riss vorhanden oder nicht. Diese Aussagen stellen das unterste Level der Auswertbarkeit dar.

Ziel und Nutzen

Das Ziel ist es mit der diskreten Regression eine Wahrscheinlichkeit für das Eintreten einer bestimmten „Eigenschaft“ vorherzusagen, z.B. wann ein Bauteil schlecht ist. Mit dieser Methode können Modelle bestimmt werden, deren Zielgröße nur auf zwei Stufen besteht, was mit der Multiplen Regression nicht möglich ist.

Grundlagen

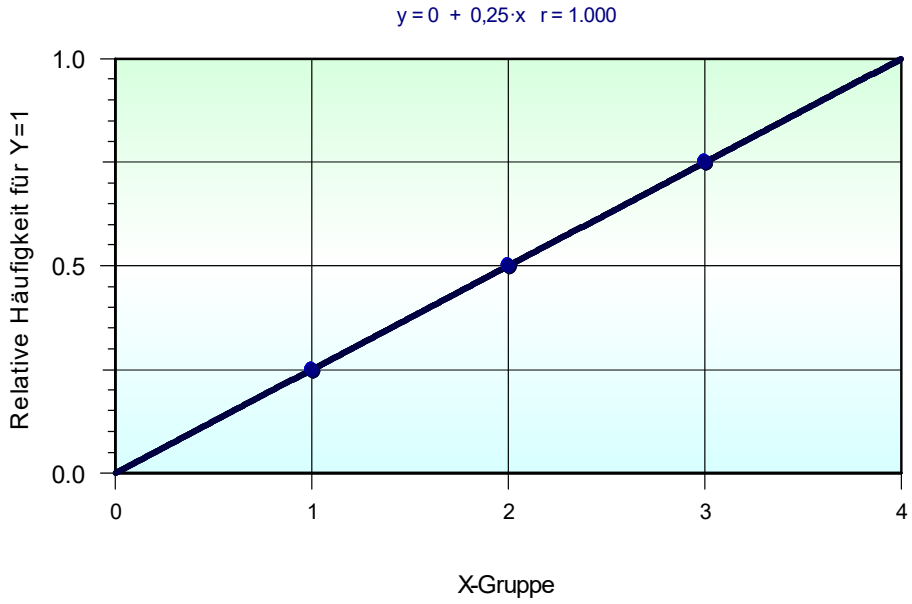
Wenn nur ein Unterscheiden auf 2 Stufen (gut/schlecht, schwarz/weiß, 0/1, usw.) möglich ist, kann man die folgende Vorgehensweise anwenden. Gegeben sei folgender Zusammenhang, der zu der nicht befriedigenden folgenden Regression führt (Ausgleichsgerade):



Sinnvoller ist es hier, statt der direkten Darstellung der Zielgröße, die Wahrscheinlichkeiten, dass ein „Zustand“ eintritt, darzustellen. Hierzu fasst man x-Bereiche zusammenfassen (Klassierung) um auf „zählbare Ereignisse“ zu kommen. Die Tabelle wird dann zu:

x (Originalwerte)	0,5	0,7	1	1,3	1,5	2	2,2	2,3	2,7	3	3,2	3,4
x-Gruppe (klassiert)	1,0				2,0				3,0			
y	0	0	1	0	0	1	1	0	1	0	1	1
n_i = Anzahl (y=1)	1				2				3			
Anz./Gruppengröße	1/4 = 0,25				2/4 = 0,5				3/4 = 0,75			

Die x-Werte werden den Gruppen 1, 2 und 3 zugeordnet (entsprechend einer mittigen Klassierung, hier auf ganze Zahlen). Innerhalb dieser Gruppen wird nun die Anzahl y=1 gezählt (bei Begriffen, wie „gut“ und „schlecht“ ist festzulegen, auf was sich das Zählen bezieht, z.B. auf „schlecht“). Hieraus lassen sich die relativen Häufigkeiten pro Gruppe errechnen. Stellt man diese dar, so ergibt sich eine erheblich bessere Beziehung:



Erkauft wird dies durch eine Reduktion der x-Informationen, d.h. für diese Auswertung werden deutlich mehr Beobachtungen gebraucht, als bei stetigen Messgrößen. In dem vorherigen Beispiel stehen anstelle der ursprünglich 12 Informationen nur noch 3 zur Verfügung, was ein entsprechender Nachteil ist. Unter Umständen stehen bei der Auswertung zu wenig Freiheitsgrade zur Bestimmung von möglichen Wechselwirkungen zur Verfügung. Da es sich hier meist aber um reine Beobachtungen handelt (nicht um geplante Versuche), liegen in der Regel auch ausreichende Daten vor.

Die Bildung der relativen Häufigkeiten sind gleichzeitig Schätzer für die Wahrscheinlichkeit p, dass y =1 wird. Es gilt, wie bereits im Beispiel verwendet (letzte Zeile):

$$p_i = \frac{n_i}{n_{Gruppe}}$$ n_i : Anzahl y =1, darf nicht 0 sein; Faustwert für $n_{Gruppe} \geq 5$

Für n_i<0 und n_i>4 ergeben sich allerdings unsinnige Wahrscheinlichkeiten von p<0 und p>1. Deshalb sind geeignete Transformationen notwendig, wie z.B. durch die Arcus-Sinus-Funktion. Bevor man zu der eigentlichen Regressionsanalyse geht, werden die relativen Häufigkeiten über folgende allgemeine Beziehung umgerechnet.

$$y' = \frac{2}{\pi} ArcSin\left(\sqrt{p}\right)$$

Danach wird das Regressionsmodell gebildet. Bei der Prognose von Wahrscheinlichkeiten aus dem gefundenen Modell wird über die Umkehrfunktion

$$\hat{p} = \sin\left(\frac{\pi}{2}\hat{y}\right)^2$$

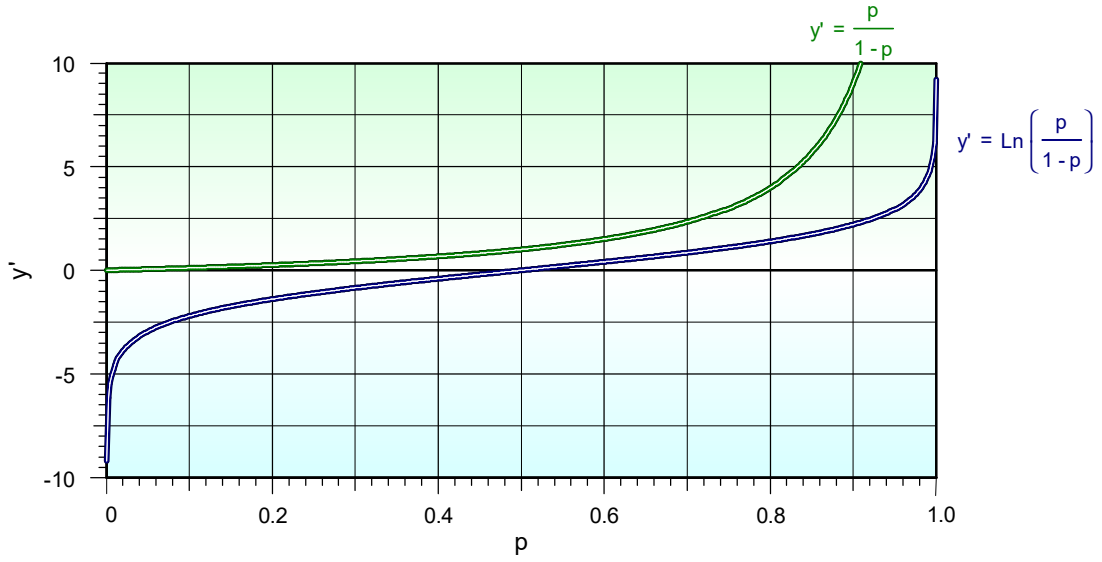
wieder auf Wahrscheinlichkeiten umgerechnet, wobei sichergestellt ist, dass Werte <0 und >1 nicht entstehen ($\hat{p}$ steht hier für den Schätzer der Wahrscheinlichkeit aus dem Regressionsmodell). Diese Art der Transformation wird insbesondere im Taschenbuch Versuchsplanung, Kleppmann empfohlen.

Eine für diese Problemstellung häufig verwendete Transformation ist das so genannte Logit-Modell:

$$y' = \ln\left(\frac{p}{1-p}\right) \quad \text{bzw.} \quad b_o + b_1 x_1 + ...b_z x_z = \ln\left(\frac{p}{1-p}\right)$$

Der Ausdruck *p/(1-p)* stellt Chancen dar (engl. odds) und hat die Bedeutung Eintrittswahrscheinlichkeit/Gegenwahrscheinlichkeit. Man spricht hier auch von Logits. Der Umgang mit Chancen und die Interpretation ist etwas ungewohnt, es sei denn man ist bei den Pferdewetten, denn die Chancen entsprechen hier den Quoten. Wichtig ist, dass die logistische Regression nicht Wahrscheinlichkeiten, sondern Wahrscheinlichkeitsverhältnisse behandelt.

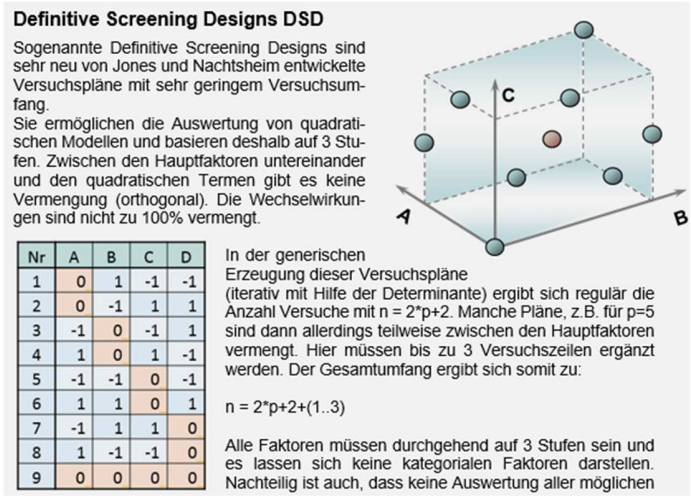
Um zusätzlich die Untergrenze des Wertebereiches zu beseitigen, werden die Chancen noch logarithmiert.



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

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

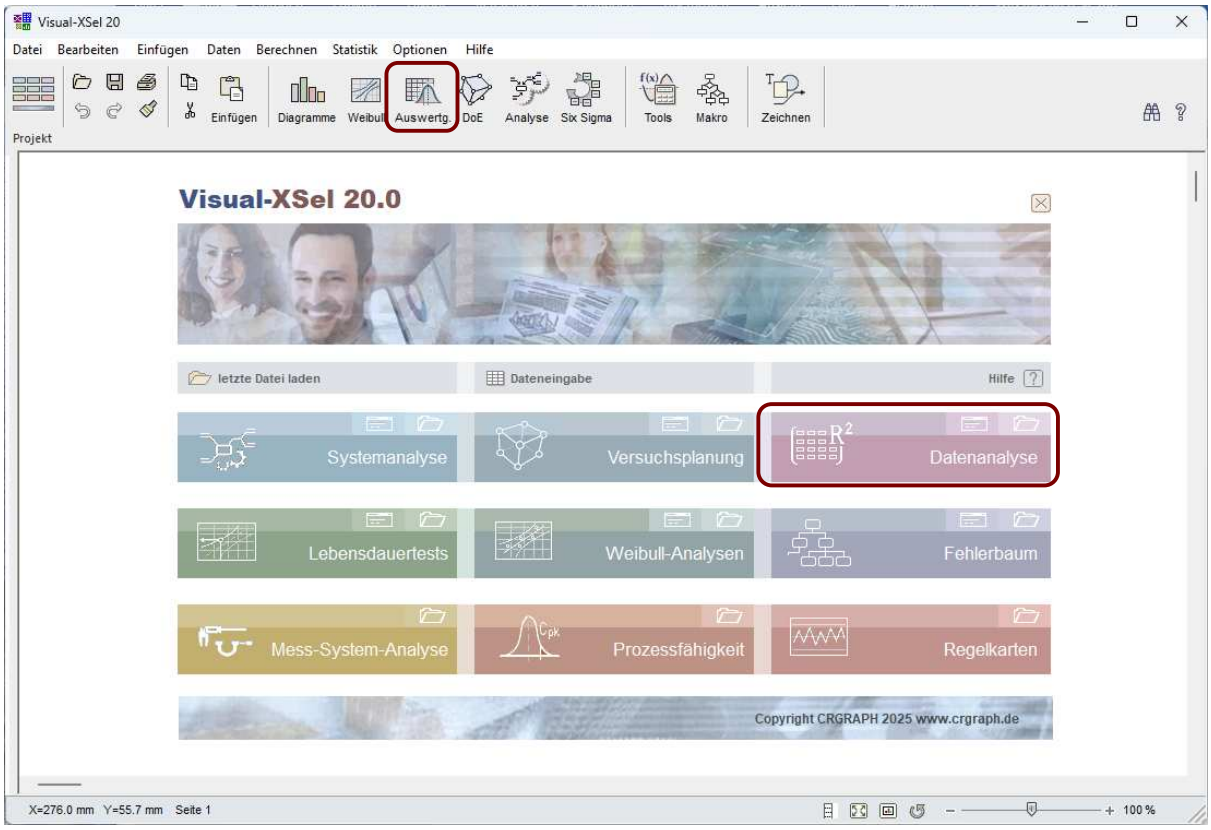
Hotline

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 **Datenanalyse** im Leitfaden (siehe auch crgraph.de/themen-index), oder die Ikone **Auswertung**.



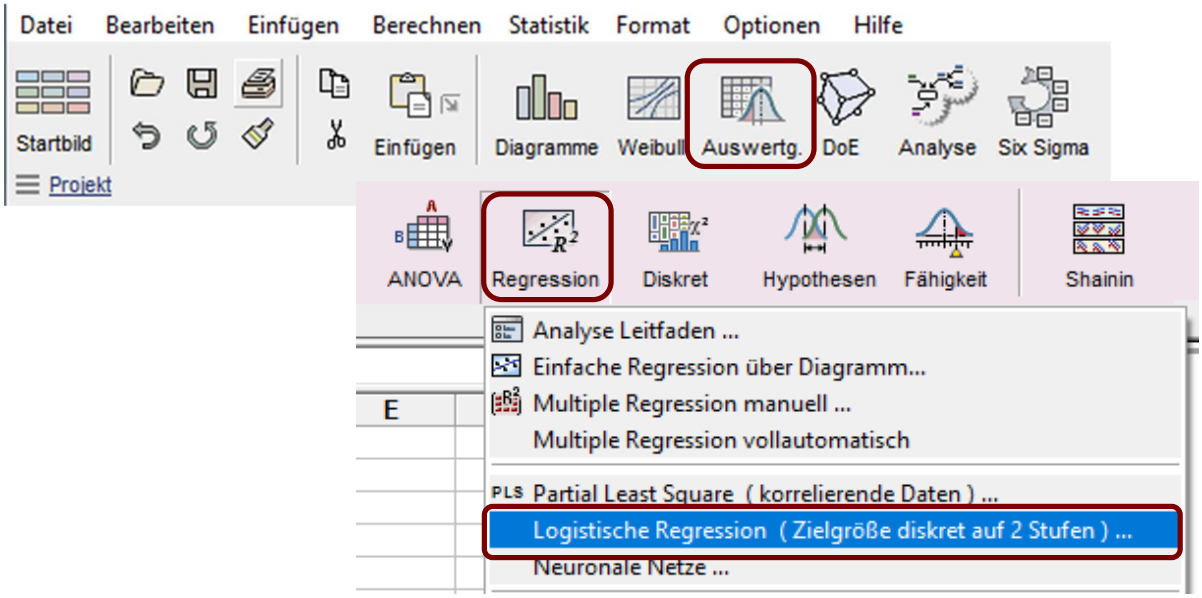
Hier finden Sie eine Übersicht und Einstiegsvideos:
crgraph.de/visual-xsel-software/

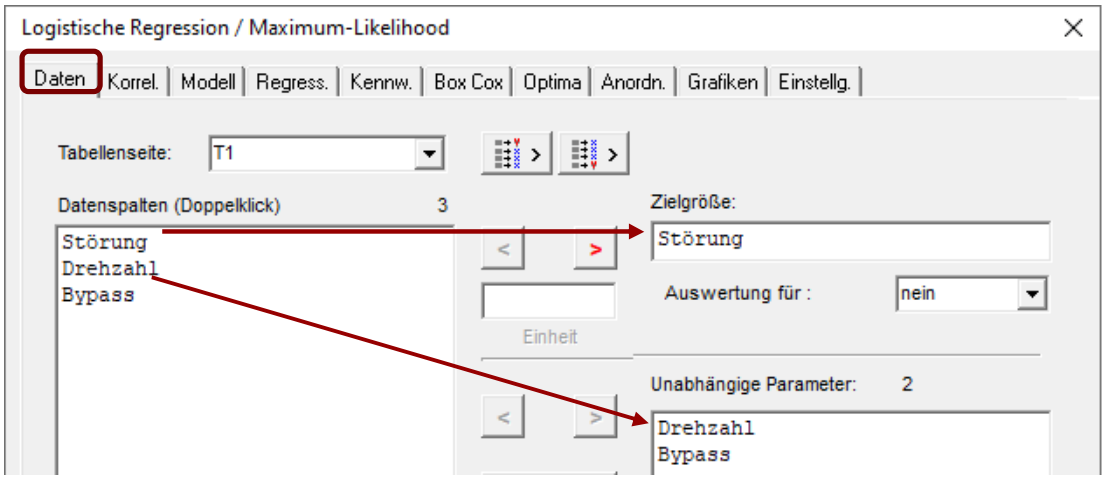
Youtube ein Video:

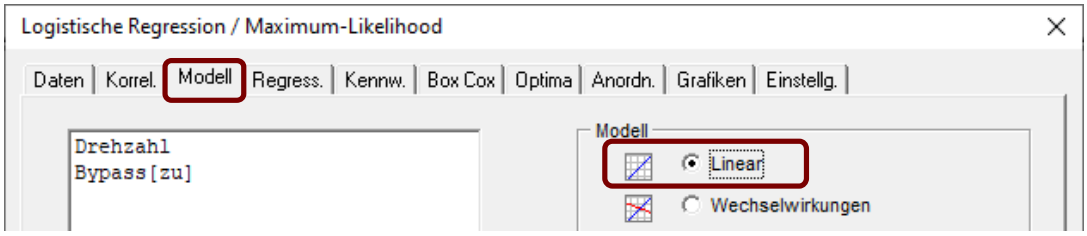
Nicht umsonst ist diese Software in vielen namhaften Firmen im Einsatz:
crgraph.de/Referenzen.

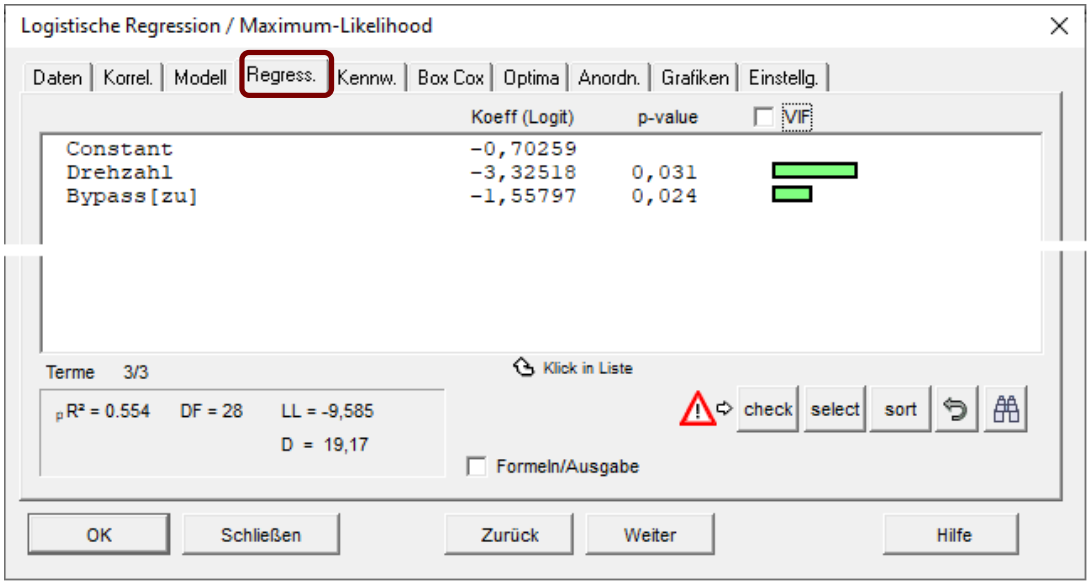
Die folgende Beschreibung ist eine Anleitung und Einführung in die Erstellung einer diskreten Regression in Visual-XSel.

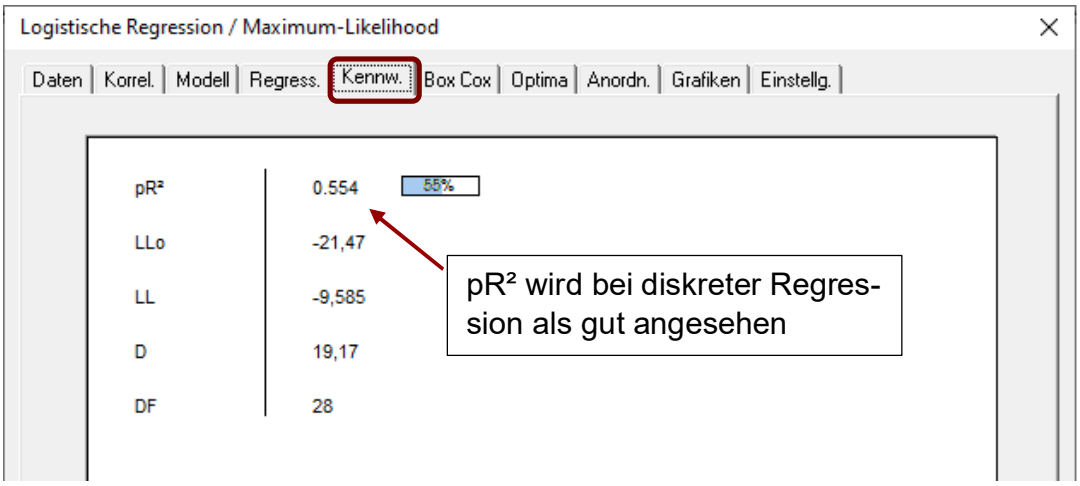
Diskrete Regression

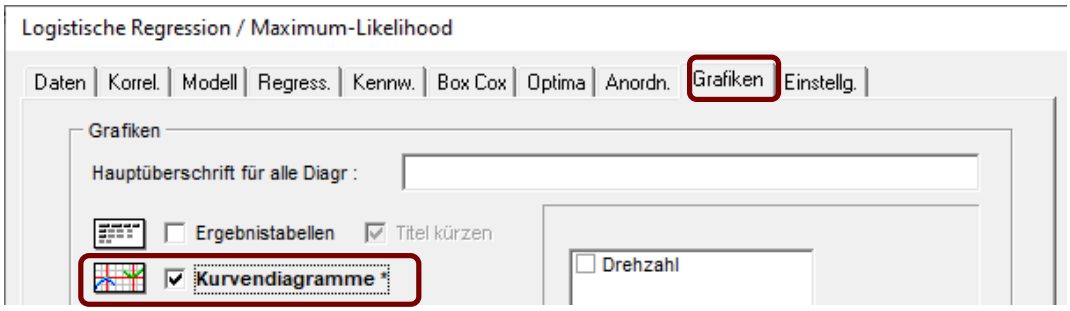


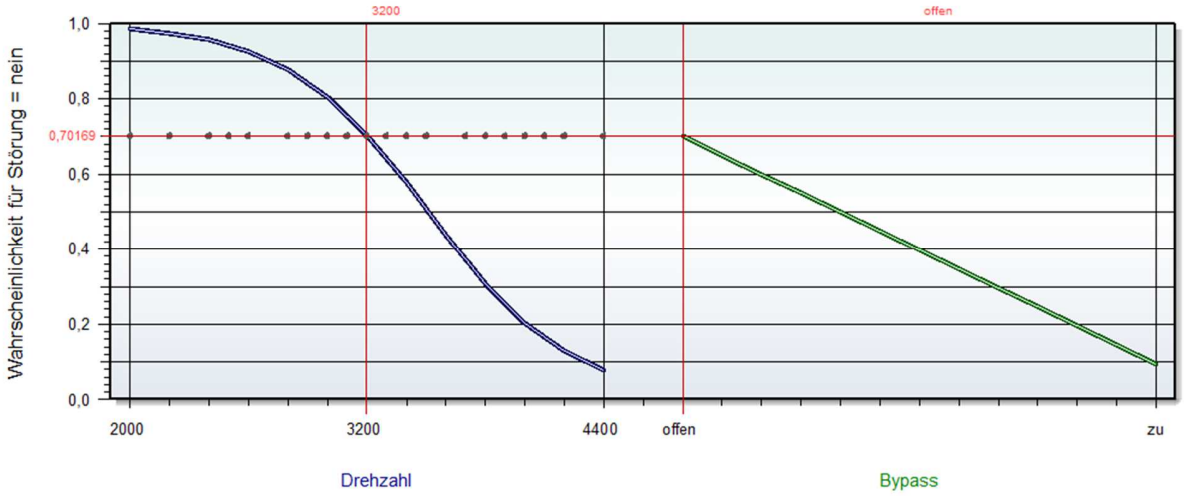












Das Beispiel zeigt, dass die Wahrscheinlichkeit für eine Störung am größten ist, wenn die Drehzahl klein und der Bypass offen ist.