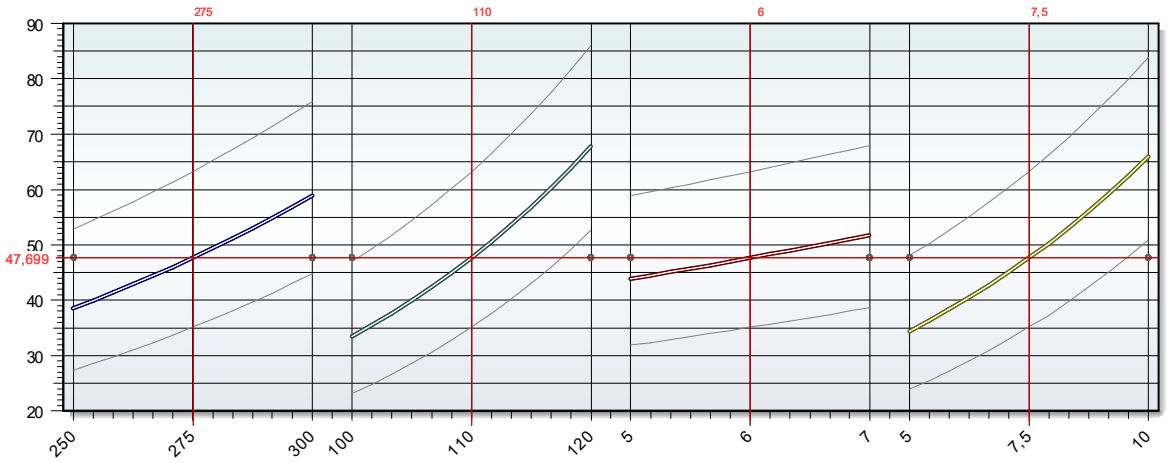
Poisson Regression

Auswertung zählbarer Ereignisse



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/Poisson-Verteilung.pdf

www.versuchsmethoden.de/Multiple-Regression.pdf

www.versuchsmethoden.de/Diskrete-Regression.pdf

Keywords:

Poisson-Regression, zählbare Ereignisse, Fehleranteil, Generalized Linear Model, GLM, Maximum-Likelihood-Estimation, MLE

Einführung

Mit der Poisson-Verteilung beschreibt man zählbare Ereignisse, die in einem definierten Zeitintervall auftreten. Dies kann z.B. die Anzahl von Fehlermeldungen innerhalb einer Woche sein. Die Poisson-Regression wird im Zusammenhang mit dem sogenannten verallgemeinerten linearen Regressionsmodell behandelt (Generalized Linear Model, kurz GLM).

Ziel und Nutzen

Das Ziel ist es, ein Modell für zählbare Ereignisse auf Basis der Poisson-Verteilung zu bestimmen, die hierfür gut geeignet ist.

Grundlagen

Die Poisson-Dichtfunktion (Wahrscheinlichkeit, dass die Anzahl Fehler y auftritt) ist definiert als (siehe: www.versuchsmethoden.de/Poisson-Verteilung.pdf):

$$g(x) = \frac{\lambda^y}{y!}\, e^{-\lambda}$$

y : Anzahl Ereignisse oder Fehler ganzzahlig
λ : Poisson- oder Fehlerrate

Bei der Poisson-Verteilung wurde die Anzahl Fehler als x definiert. Da jedoch x bei der Regression schon für den Einflussparameter steht, soll hier y verwendet werden. Wie auch bei Lebensdauermodellen, lässt sich hier der Zusammenhang für die Poisson-Rate am besten logarithmisch darstellen:

$$\ln(\lambda) = b_o + b_1 x_1 + b_2 x_2 + \cdots + b_z x_z$$

b : Koeffizienten des Modells
x : Einflussparameter

somit ist:

$$\lambda = e^{b_o + b_1 x_1 + b_2 x_2 + \cdots + b_z x_z}$$

In der Annahme, dass die beobachteten Ereignisse y_i Poisson-verteilt sind, gilt für den Erwartungswert $E[y_i] = \lambda_i$. Jede Beobachtung muss voneinander unabhängig und zufällig sein. Der beste Schätzer für die Koeffizienten b ist gegeben, wenn das Produkt aller Wahrscheinlichkeiten (Likelihood ⇨ L) jeder i-ten Beobachtung ein Maximum ist. Hierfür gilt:

$$L(b_o, b_1, b_2, \ldots b_z) = \prod_{i=1}^{n} \frac{{\lambda_i}^{y_i}}{y_i!}\, e^{-\lambda_i}$$

Die Bestimmung der Koeffizienten kann nicht analytisch, wie bei der Multiplen Regression erfolgen, denn y steht sowohl im Exponenten, als auch im Nenner mit Fakultät. Damit wird deutlich, dass sich das Ergebnis von der Methode der kleinsten Fehlerquad-

rate unterscheiden muss. Allerdings sind die Unterschiede gering und die hiermit bestimmten Koeffizienten können als Startbedingung der iterativen Berechnung verwendet werden. Am häufigsten wird für diese iterative Berechnung das sogenannte Newton-Raphson Verfahren genannt. Für die Suche der maximalen Wahrscheinlichkeiten kann auch die logarithmierte Form verwendet werden, was die sogenannte Log-Likelihood-Funktion LL ergibt:

$$LL(b_o, b_1, b_2, \dots b_z) = \sum_{i=1}^{n} y_i \ln \lambda_i - \lambda_i - \ln(y_i!)$$

Es werden die Koeffizienten gesucht, für die die Summe der rechten Seite ein Maximum ist, weshalb die Methode als Maximum-Likelihood-Estimation kurz MLE bekannt ist. Je nach Abbruchkriterium, bzw. Anzahl gewählter Iterationsschritte, können leicht unterschiedliche Ergebnisse entstehen!

Die Streuung der Koeffizienten s_b wird bestimmt durch:

$$s_{b,i} = diag(\sqrt{(X^T W)^{-1}})$$

mit der Gewichtung

$$w_{j,i} = x_{j,i} \cdot \mu_i \quad \text{und} \quad \mu_i = \exp(b_o + b_1 x_{1,i} + b_2 x_{2,i} + \cdots)$$

Zur Bestimmung, ob die Koeffizienten signifikant sind ergibt sich der *p-value* durch

p-value = 2·*VertlgNormal*(-|z|)

mit

$$z = \frac{b_i}{s_{b,i}}$$

Der Parameter gilt als signifikant, wenn *p-value* < 0,05, siehe:
www.versuchsmethoden.de/Hypothesentests.pdf

Zur Beurteilung des Gesamtmodells wird die Summe der Abweichungsquadrate als sogenannte Deviance berechnet. Die Modellstreuung ist:

$$D_{(b)} = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{e^{Xb}}\right) - (y_i - e^{Xb}) \qquad \text{mit } Xb =$$

und die bei der nur die Konstante b_o im Modell ist:

$$D_{(bo)} = 2 \sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{e^{bo}}\right) - (y_i - e^{bo})$$

womit sich das Bestimmtheitsmaß ergibt mit:

$$R^2 = 1 - \frac{D_{(b)}}{D_{(bo)}}$$

Anstelle des adjustierten Bestimmtheitsmaßes, wie bei der multiplen Regression, wird hier ein korrigiertes R^2_{kor} angegeben:

$$R^2_{kor} = R^2 - \frac{z}{D_{(bo)}}$$

Poisson-Regression mit Intercept

Wenn für bestimmte Kombinationen der Einflussparameter keine Ereignisse auftreten, gibt es das Problem, dass ln(y_i=0) nicht möglich ist. Deshalb sollte ein Offset verwendet werden, der hier Intercept genannt wird und der für jede Beobachtung individuell festgelegt werden kann. Zu empfehlen ist diesen Intercept nur für den Fall von yi=0, z.B. auf 0,01 zu setzen.

Normiertes Modell

Wie auch bei der Multiplen Regression ist es auch hier zu empfehlen, dass die Einfluss Parameter x normiert werden:

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

Hierdurch ist der Wertebereich zwischen -1 .. +1 und die Bestimmung des Modells wird stabiler, die Signifikanzen werden eindeutiger. Im Gegensatz zu Visual-XSel verwenden vielen Statistikprogramme als Standardeinstellung die Original-Wertebereiche, weshalb sich die Koeffizienten unterscheiden und die Modelle evtl. verschieden sind.

Weitere Kennzahlen

Eine weitere sehr häufig genannte Kennzahl für Likelihood-basierte Modelle und somit auch für die Poisson-Regression ist das nach dem Japaner Hirotsugu Akaike benannte **A**kaike **I**nformation **C**riterion, kurz AIC:

$$AIC = -2\,LL + 2(z+1)$$

Je kleiner AIC ist, desto besser, jedoch ist AIC alleine nicht interpretierbar, denn es gibt keinen allgemeingültigen Grenzwert für einen Bestwert. AIC dient somit eher für Vergleiche zwischen verschiedenen Modellen. Wie beim R^2 auch, liefern Modelle mit mehr Termen, wie z.B. Wechselwirkungsterme, meist bessere Kennwerte, obwohl z zunimmt und mit dem Faktor 2 eingeht. Diese Terme dürfen aber nur im Modell bleiben, wenn sie auch signifikant sind (siehe $p\text{-}value$).

Eine weitere Kennzahl ist das sogenannte **B**ayesian **I**nformation **C**riterion, kurz BIC (benannt nach dem englischen Statistiker Thomas Bayes). Diese Kennzahl ist dem AIC sehr ähnlich und berücksichtigt zusätzlich die Anzahl der Beobachtungen n:

$$BIC = -2\,LL + (z+1)\cdot\ln(n)$$

Ab n > 7 wird BIC größer als AIC und straft komplexere Modelle stärker ab. Der Einfluss von n soll aber nicht so verstanden werden, dass möglichst kleine Datensätze verwendet werden sollten. Mehr Informationen sind ja für die Modellbildung grundsätzlich von Vorteil.

Obwohl AIC und BIC oft als wichtige Kennzahlen genannt werden, erscheint der praktische Nutzen wegen der genannten Nachteile begrenzt.

Vertrauensintervalle

Für die Vertrauensintervalle der Modell-Vorhersagen kann die χ^2-Verteilung verwendet werden. Bei einem üblichen Vertrauensbereich von 90% ist die Irrtumswahrscheinlichkeit α = 10%. Der zweiseitige Vertrauensbereich für die Anzahl (Fehler)-Ereignisse y ist:

$$\frac{1}{2}\,\chi^2_{\frac{\alpha}{2},2y} \;\le\; y \;\le\; \frac{1}{2}\,\chi^2_{1-\frac{\alpha}{2},2(y+1)}$$

Beispiel Ausschuss in einem Herstellprozess

In einem Versuchsplan wurde der Ausschuss von Bauteilen in Abhängigkeit von Temperatur, Druck, Zeit und Partikel-Korngröße im Material gezählt.

Nr	Temp[C]	Druck[bar]	Zeit[min]	Korngröße	Ausschuß
1	250	100	5	5	7
2	250	100	5	10	36
3	250	100	7	5	14
4	250	100	7	10	45
5	250	120	5	5	39
6	250	120	5	10	74
7	250	120	7	5	48
8	250	120	7	10	84
9	300	100	5	5	26
10	300	100	5	10	59
11	300	100	7	5	34
12	300	100	7	10	69
13	300	120	5	5	62
14	300	120	5	10	99
15	300	120	7	5	71
16	300	120	7	10	109

Das Modell ergibt folgende Koeffizienten

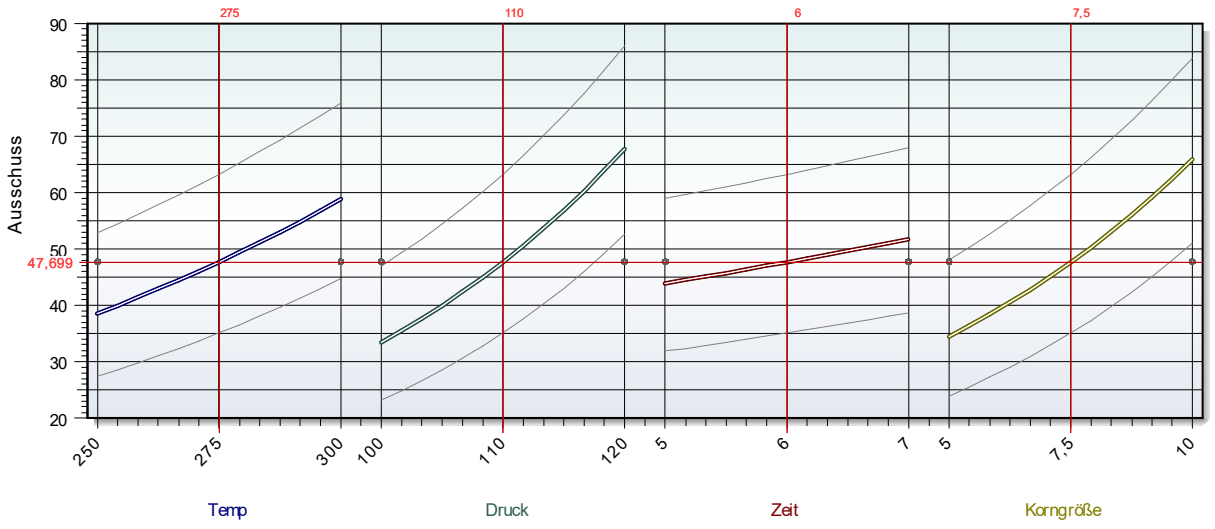
	Koeffizient	p-value
Constant	−3,79657	
Temp	0,008435	0,000
Druck	0,035221	0,000
Zeit	0,082689	0,015
Korngröße	0,129531	0,000

R^2 = 0.926 DF = 11 AIC = 119.3

R^2_{kor} = 0.910 BIC = 123.1

Alle Parameter sind signifikant (*p-value* < 0,05).

Das folgende Kurvendiagramm zeigt die Zusammenhänge als stetige Funktion, deren Ergebnisse auf- oder abzurunden sind.



Die Vertrauensbereiche sind relativ stark unsymmetrisch, siehe Kapitel Vertrauensintervalle.

Eine interessante Frage ist, wie sich das Modell verändert, wenn man den Beobachtungszeitraum vergrößert. Für eine Verdoppelung der Beobachtungszeit soll für diese Fragestellung angenommen werden, dass sich der Ausschuss ebenfalls jeweils verdoppelt (in der Realität würde das durch Schwankungen nicht exakt zutreffen). Das Ergebnis ist für den doppelten Ausschuss:

	Koeffizient	p-value
Constant	-3,103423	
Temp	0,008435	0,000
Druck	0,035221	0,000
Zeit	0,08269	0,001
Korngröße	0,129532	0,000

Es verändert sich offensichtlich nur die Konstante, die weiteren Modell-Koeffizienten und deren *p-values* bleiben gleich. Aussagen über die Zusammenhänge sind also theoretisch unabhängig vom Beobachtungszeitraum. Ist dieser jedoch zu klein, kann eine größere Zahl mit 0-Ereignissen vorkommen, was von großem Nachteil ist. Der Beobachtungszeitraum ist deshalb so zu wählen, dass auch „genügend“ Ereignisse zu erwarten sind.

Hinweis: Für dieses Beispiel wurden die Parameterskalierung nicht normiert, was normalerweise nicht zu empfehlen ist, insbesondere bei Modellen mit Wechselwirkungen und weiteren höherwertigen Termen. Die Koeffizienten unterscheiden sich zwischen Originalwerten (normal) und normierte Skalierung.

Literatur

Generalized Linear Model, 2nd Edition, McCullagh, J.A. Nelder, Chapman and Hall

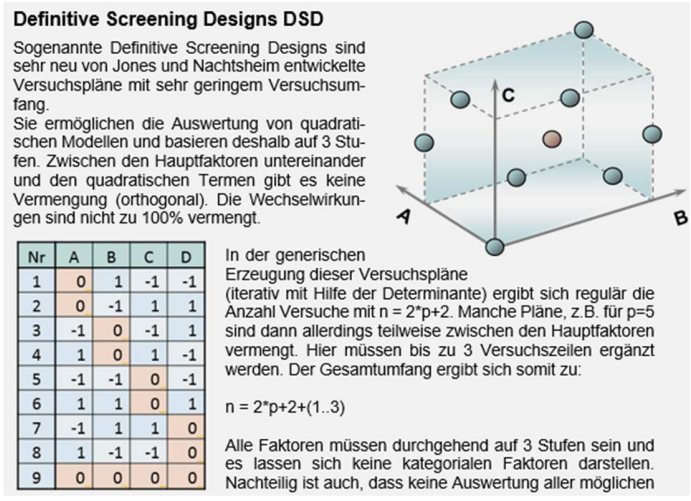
Generalisierte Lineare Modelle, H. Friedl, Techn. Universität Graz, 2014

Methoden der Statistik, Lehrbuch Version 1.0 2017, Humboldt-Universität zu Berlin

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

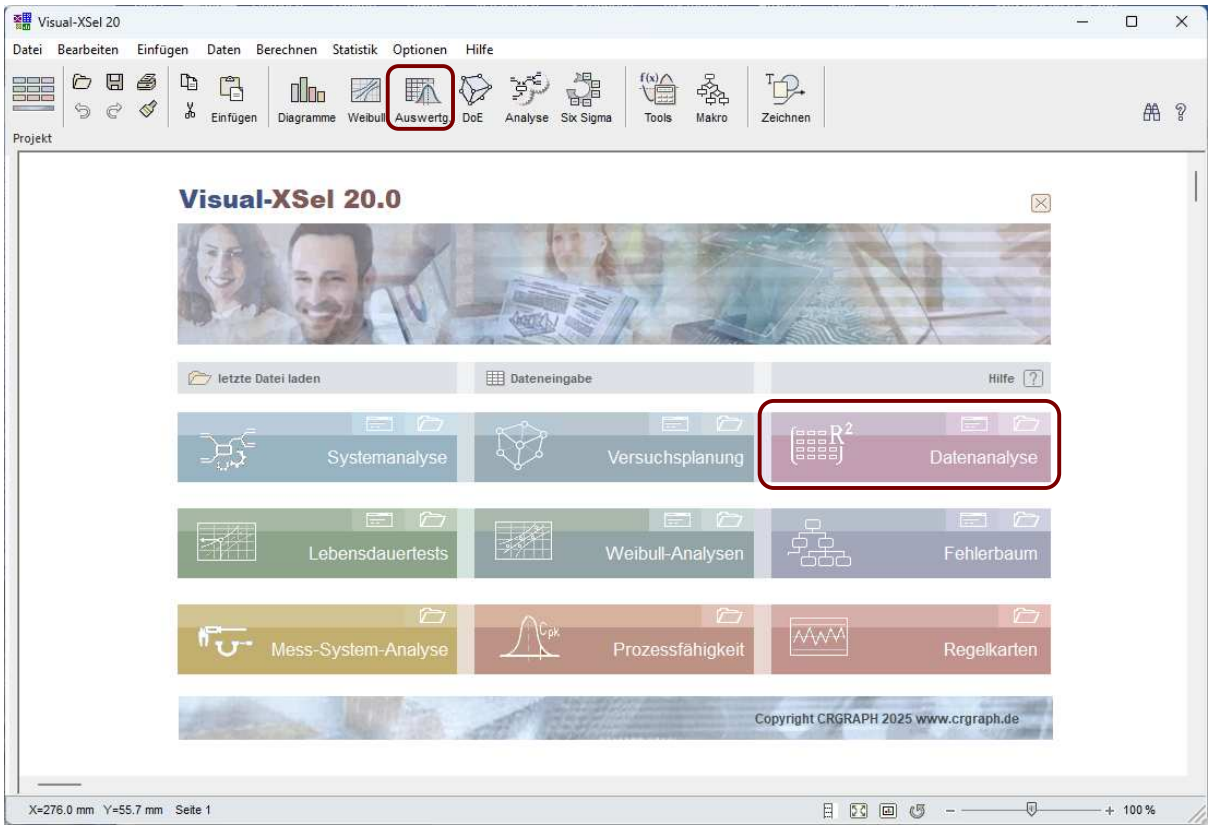
Hotline

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**.

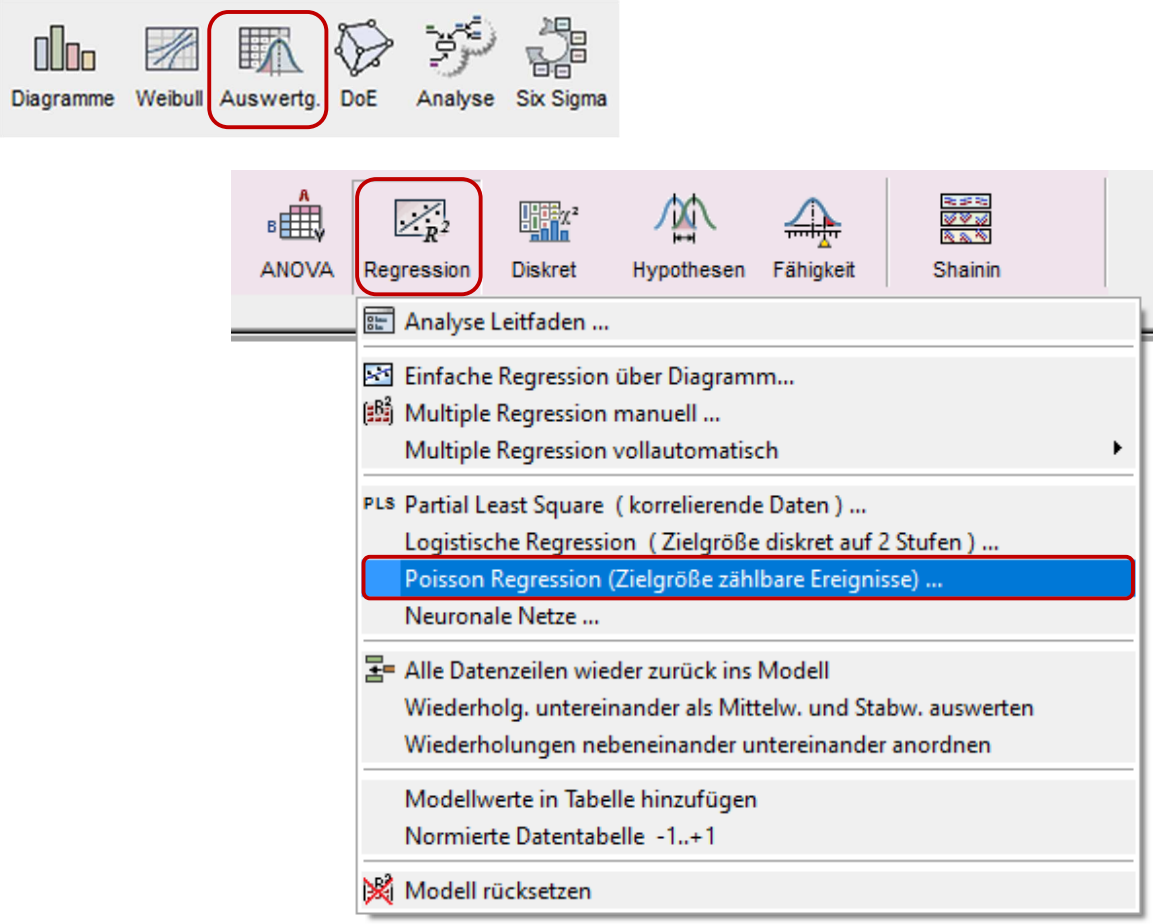


Hier finden Sie eine Übersicht und Einstiegsvideos:
crgraph.de/visual-xsel-software/

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 der Poisson-Regression in Visual-XSel.



Die weiteren Schritte entsprechen der gleichen Vorgehensweise, wie in der Programm-beschreibung *www.versuchsmethoden.de/Multiple-Regression.pdf*