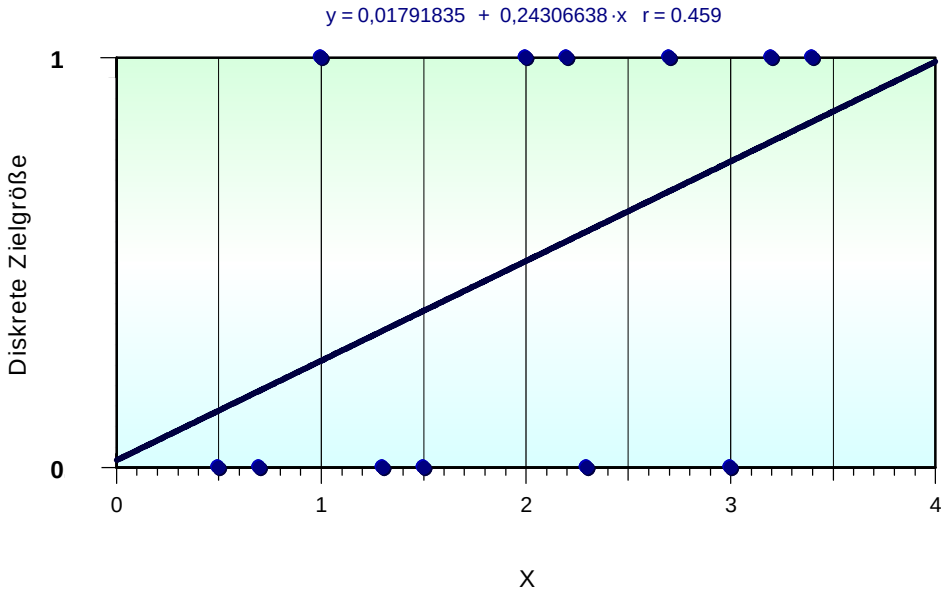
Diskrete Regressionsanalyse

$$LH = \prod_{i=1}^{n} \hat{p}_i^{y_i} \cdot (1-\hat{p}_i)^{1-y_i}$$

Unter einer diskreten 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 sollte es immer sein, die möglichst beste „Auflösung“ zu erhalten, d.h. zumindest eine Abstufung wie Anriss, Riss bis Mitte, Riss fast vollständig und abgerissen. Hierdurch ist, wenn auch mit einiger Unschärfe, eine Auswertung nach herkömmlicher Art immer noch möglich (dabei ist die Abstufung mit möglichst gleichen Abständen zu erreichen).
Wenn weiterhin nur eine Unterscheiden auf 2 Stufen (gut/schlecht, schwarz/weis, 0/1, usw.) möglich ist, kann man die folgenden Vorgehensweise anwenden. Gegeben sei folgender Zusammenhang,

x	0,5	0,7	1	1,3	1,5	2	2,2	2,3	2,7	3	3,2	3,4
y	0	0	1	0	0	1	1	0	1	0	1	1

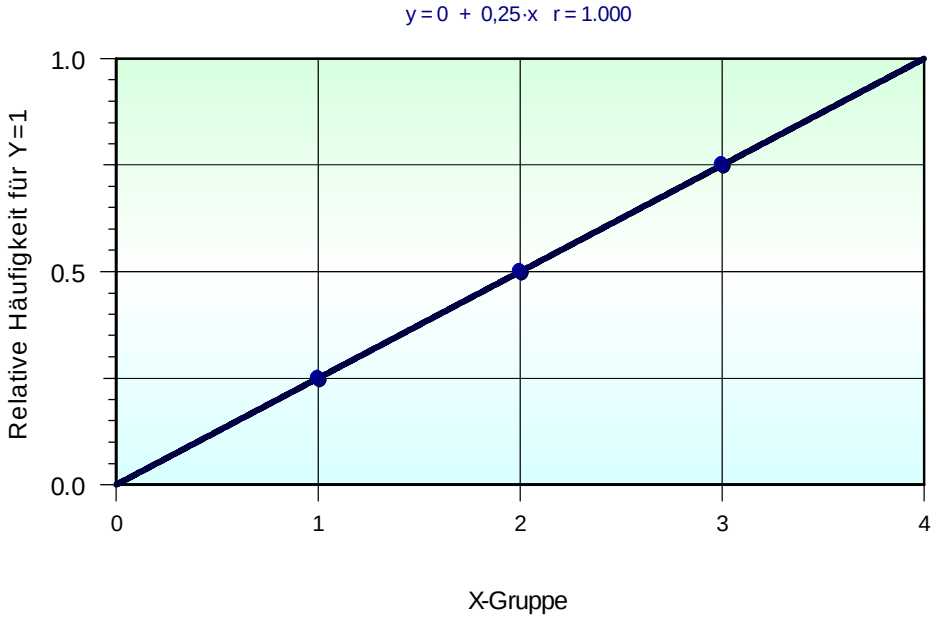
der zu der nicht befriedigenden folgenden Regression führt (Ausgleichsgerade):



Sinnvoller ist es hier statt der direkten Darstellung der Zielgröße die Wahrscheinlichkeiten, dass ein „Zustand“ eintritt, darzustellen. Hierzu fast 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 aber 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.

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

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

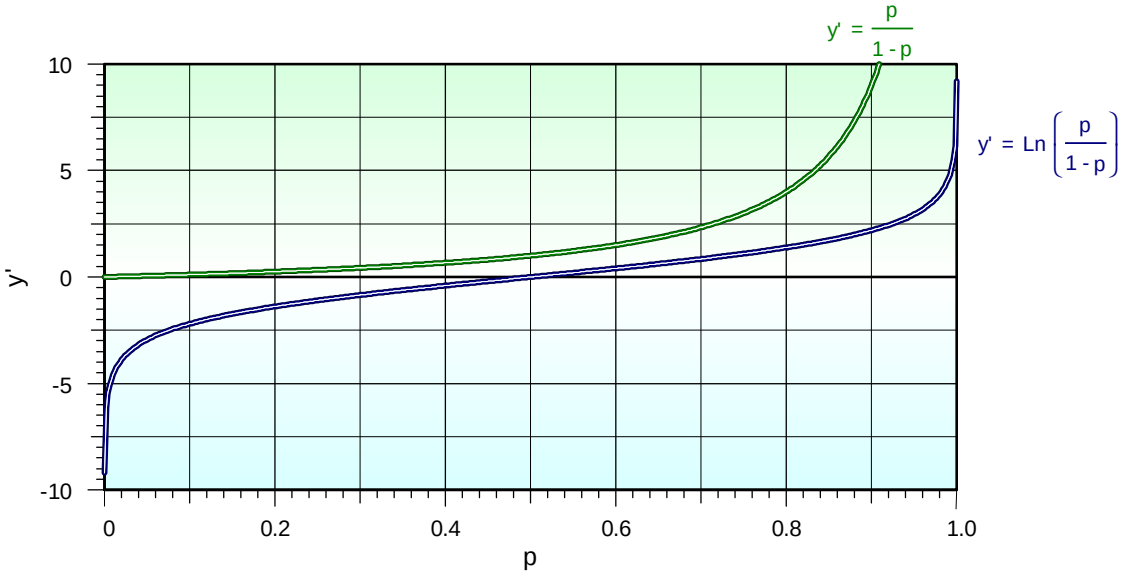
$$\hat{p} = \sin\left(\frac{\pi}{2}\hat{y}\right)^2$$

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 in /3/ empfohlen.
Eine für diese Problemstellung häufig verwendete Transformation ist das sogenannte Logit-Modell:

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

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 beim Pferdewetten, denn die Chancen entsprechen hier den Quoten. Wichtig ist, sich zu merken, dass die logistische Regression nicht Wahrscheinlichkeiten, sondern Wahrscheinlichkeitsverhältnisse behandelt.

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



Auch hier wird nach der Bestimmung der Modellparameter für die Zurückrechnung auf Wahrscheinlichkeiten die Umkehrfunktion benötigt:

$$\hat{p} = \frac{1}{1+e^{-\hat{y}}}$$

Diese wird auch als „logistische“ Verteilungsfunktion bezeichnet. In der Literatur findet man auch das sogenannte Probit-Modell. Hierbei wird die Wahrscheinlichkeitsfunktion der Standardnormalverteilung angesetzt:

$$\hat{p} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\hat{y}^2}$$

Die klassische Anwendung des Probit-Modells ist im Bereich der Dosis-Wirkungs-Analyse, z.B. bei Giften und Medikamenten. Der Nachteil dieses Modells ist, dass sich die Verteilungsfunktion nicht nach y auslösen lässt (die eigentliche Transformation müsste über die inverse Funktion erfolgen bzw. benötigt die Quantile der Normalverteilung). Abgesehen von den Randbereichen liefern Logit- und Probit-Modell praktisch gleiche Werte, weshalb es für die Thematik ausreicht ausschließlich mit dem Logit-Modell zu arbeiten.

Die Grenzen p = 0 und p = 1 sind über das Logit nicht darstellbar. Die Anzahl n_i pro Gruppe sollte ohnehin nicht 0 sein. Ist dies der Fall, so sind die Gruppen evtl. weiter zu fassen.
Bei der Regression mit stetigen Zielgrößen ist die Voraussetzung für die Methode der kleinsten Fehler-Quadrate zur Abschätzung der gesuchten Koeffizienten b, dass die Fehlerabweichungen identische Varianz haben. Dies ist hier nicht der Fall. Deshalb muss eine gewichtete Regression angewendet werden. Hierzu wird ein Schätzer für die Varianz benötigt. Zur Ermittlung der Koeffizienten bei nicht gewichteter Regression wurde bisher die bereits eingeführte Beziehung

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

verwendet. Bei der logistischen Regression besteht das Problem, dass die Varianzen der Modellfehler nicht konstant sind. Dadurch können über die Methode der kleinsten Fehlerquadrate die Varianzen der Modellschätzer nicht minimiert werden. Das Problem lässt sich aber durch eine gewichteten Regression beseitigen. Hierzu werden die Varianzen jeder Beobachtung gebraucht, die durch

$$\hat{s}_i^2 = \hat{p}_i (1 - \hat{p}_i)$$

definiert sind. Die Regressionskoeffizienten bestimmen sich dann durch:

$$b = \left(X^T \delta X\right)^{-1} X^T \delta \; y'$$

mit $\delta = diag(s_1^2, s_2^2 \; ....., s_n^2)$

wobei y' der Vektor der entsprechenden Logits ist. Es entsteht jedoch ein neues Problem. Die Schätzer für $\hat{p}_i$ bestimmen sich erst aus dem Ergebnis der Berechnung. Es muss also eine iterative Berechnung erfolgen.

Eine andere Möglichkeit zur Bestimmung der Modellparameter ist die Maximum-Likelihood, kurz ML-Methode. Das Grundprinzip ist relativ einfach. Die Parameter werden so gewählt, dass die geschätzten Variablen den Beobachtungen im Datensatz am ähnlichsten sind (Likelihood). Die Ähnlichkeit wird durch die sogenannte Likelihood-Funktion beschrieben, die sich aus dem Produkt der Likelihoods aller Fälle des Datensatzes zusammensetzt:

$$LH = \prod_{i=1}^{n} \hat{p}_i^{y_i} \cdot (1 - \hat{p}_i)^{1-y_i}$$

y_i stammt aus den n Beobachtungen, $\hat{p}_i$ aus dem Modell. Die Koeffizienten des Modells sind nun so zu suchen, dass LH maximal wird. Da die Likelihood eines einzelnen Falles

etwas ähnliches ist wie eine Wahrscheinlichkeit, kann sie einen Wert zwischen 0 und 1 annehmen. Das Produkt vieler Zahlen zwischen 0 und 1 wird allerdings sehr klein, deshalb wird auch hier LH logarithmiert und es entsteht das Log Likelihood[1] kurz LL:

$$\ln(LH) = LL = \sum_{i=1}^{n} y_i \cdot \ln(\hat{p}_i) + (1 - y_i) \cdot \ln(1 - \hat{p}_i)$$

Für beide Varianten gibt es keine analytische Lösung. Die Koeffizienten müssen ebenfalls iterativ bestimmt werden, wobei man zunächst einen beliebigen Startwert wählt. Mit diesen können die Logits und die ersten Schätzwerte der Wahrscheinlichkeiten $\hat{p}_i$ bestimmt werden. Damit wird für jede Datenreihe das Produkt der LH-Funktion oder die Summe der LL berechnet. Das Gleiche muss mit variierten Koeffizienten solange wiederholt werden, bis sich kein größerer LH- bzw. LL-Wert finden lässt (siehe Schrittweise Optimierungsverfahren).

Der wichtigste Vorteil der Maximum-Likelihood-Methode ist, dass zur Bestimmung der Koeffizienten keine Gruppenbildung der Daten erforderlich ist (die evtl. 0 Ereignisse enthalten können, womit das Logit nicht berechenbar ist).

Eine Maßzahl für die Güte der gefundenen Lösung ist die sogenannte Devianz (=Abweichung):

$$D = -2\,LL$$

Da der logarithmierte Wert zwischen 0 und 1 immer negativ ist wird das Vorzeichen geändert. Hierdurch erhält man zusätzlich mit dem Faktor 2 einen χ^2-verteilten Wert, der besagt, wie schlecht das Modell die Daten beschreibt. Deshalb ist es umso besser, je kleiner dieser Wert ist.
Bei der normalen multiplen Regression wird vor allem das Bestimmtheitsmaß R^2 für die Güte des Modells angegeben. Hier gibt es keine direkte Entsprechung, von McFadden wurde aber ein pseudo-R^2 definiert:

$$R_{MF}^2 = \frac{LL_0 - LL_1}{LL_0} = 1 - \frac{LL_1}{LL_0}$$

LL_o : Log-Likelihood des Modells, das nur aus der Konstanten $y'=b_o$ besteht
LL_1 : Log-Likelhood des konkreten Modells $y'=b_{o+}\ b_1x_1$ + ...

R_{MF}^2 kann nicht den Wert 1 erreichen, Werte von 0,2 – 0,4 werden in der Regel schon als gute Modellanpassung betrachtet.

Zur Bewertung der Signifikanz der einzelnen Koeffizienten (Faktoren) wird der sogenannte Devianz-Test empfohlen. Dabei wird geprüft, ob das Modell mit dem jeweiligen Faktor gegenüber dem ohne diesen einen signifikanten Unterschied aufweist. Zur Prüfung eines Faktors wird die Differenz der Devianzen gebildet:

$$\Delta D_F = -2\,LL_1 - (-2\,LL_F) = -2\left(LL_1 - LL_F\right)$$

[1] Hinweis: Man beachte die Sprachregelung Log Likelihood, obwohl für die Summe der Likelihoods der *ln* verwendet wird

wobei der Index $_F$ für das Modell ohne den zu betrachtenden Faktor gegenüber dem Ausgangsmodell mit dem Index $_1$ steht (siehe pseudo R^2).
Mit Hilfe der χ^2-Verteilung und dem Parameter ΔD_F, sowie mit dem Freiheitsgrad df=1 wird der p-Value = 1-α bestimmt.
Analog hierzu kann auch das Gesamtmodell gegenüber dem „Null-Modell“ getestet werden (siehe wiederum pseudo R^2). Die Differenzdevianz ist:

$$\Delta D_G = -2\left(LL_1 - LL_0\right)$$

mit den Freiheitsgrad: $df = z =$ Anzahl Faktoren, Wechselwirkungen usw.

Diese Vorgehensweise bedeutet jedoch einen relativ hoher Rechenaufwand, denn es muss für jeden zu prüfenden Faktor die ML-Iteration durchgeführt werden. Alternativ hierzu kann der häufig genannte Wald-Test verwendet werden. Dieser ist ähnlich dem t-Test bei der normalen Regression. Die Testgröße ist für jeden Faktor:

$$\chi_j^2 = \left(\frac{b_j}{s_{b,j}}\right)^2$$

mit $s_{b,j} = \sqrt{X^*_{j,j}}$ und $X^* = \left(X^T \delta X\right)^{-1}$ wobei δ die bereits eingeführte Diagonalmatrix aus den Varianzen jeder Beobachtungsreihe war.
Der p-Value berechnet sich auch hier über 1-α aus der χ^2-Verteilung mit df=1.

Im folgendem werden Verfahren beschrieben, mit denen es möglich ist die iterative Maximum-Likelihood-Methode zu behandeln.