

# Wahl der Modellklasse: lineare Regression, Logit Modell etc.

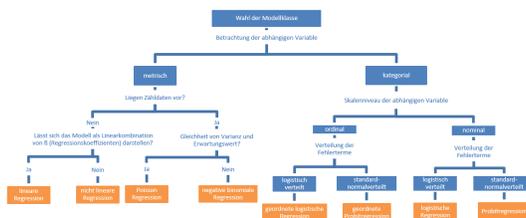
## Erläuterung der Problemstellung

Im Allgemeinen wird bei der Regressionsanalyse der Zusammenhang zwischen einer abhängigen Variable  $(Y)$  und einer  $(X_1)$  oder mehreren unabhängigen Variablen  $(X = (X_1, X_2, \dots, X_P))$  untersucht. Dieser Zusammenhang kann sich je nach Datenlage und untersuchter Fragestellung der empirischen Arbeit deutlich unterscheiden. In diesem Abschnitt "Wahl der Modellklasse", soll auf die gängigsten Modelle der Regressionsanalyse eingegangen werden. Anhand von bestimmten Kriterien wird erläutert, wann welche Modellierung am besten geeignet ist, um den Zusammenhang in den vorliegenden Daten darzustellen, ausgehend von dem funktionalen Zusammenhang  $(Y = f(X, \beta) + \epsilon)$ , der die abhängige Variable als Funktion von unabhängigen Variablen und zu schätzenden Parametern  $(\beta)$  zuzüglich eines Fehler- oder Residuenters  $(\epsilon)$  ausdrückt. Um zu entscheiden, welches Modell geschätzt werden soll, betrachtet man nun zuerst die abhängige Variable  $(Y)$  und kann aufgrund ihres Skalenniveaus eine Vorauswahl an Modellklassen treffen. Als zweiten Schritt betrachtet man den funktionalen Zusammenhang zwischen  $(Y)$  und  $(X)$  und wählt so die passende Modellklasse aus. Abschließend muss überprüft werden, ob die jeweiligen zusätzlichen Annahmen der gewählten Modellierung erfüllt sind.

Allerdings ist die additive Darstellung  $(Y = f(X, \beta) + \epsilon)$  für nicht metrische Zufallsgrößen  $(Y)$  **robust vereinfachend**, da die Addition für nicht metrische Merkmale gar nicht definiert ist. Für nicht metrische abhängige Größen muss man  $(Y)$  durch  $(Y^*)$ , die latente Ausprägung von  $(Y)$  im Sinne eines Schwellenwertmodells, ersetzen. Trotz dieser Einschränkung findet man eine derartige Systematisierung der unterschiedlichen Regressionsmodelle relativ häufig.

Zur Veranschaulichung der Gliederung dieses Abschnitts, sowie zum Einordnen der unterschiedlichen behandelten Modelle dient die folgende Grafik: Ausgehend von der Betrachtung der abhängigen Variablen findet hier eine erste Einteilung in unterschiedliche Modellklassen statt, durch weitere Betrachtungen kann eine Unterscheidung zwischen den hier vorgestellten Modellen vollzogen werden.

Unter diesem [Link](#) finden **SAS**-, **Stata**- oder **SPSS**-Nutzer eine nützliche praktische Anleitung zur Umsetzung einer Regression mit kategoriellen abhängigen Variablen.



**i** fu:stat bietet regelmäßig Schulungen für **Hochschulangehörige** sowie für **Unternehmen und weitere Institutionen** an. Die Inhalte reichen von Statistikgrundlagen (Deskriptive, Testen, Schätzen, lineare Regression) bis zu Methoden für Big Data. Es werden außerdem Kurse zu verschiedenen Software-Paketen gegeben. Auf Anfrage können wir auch gerne individuelle **Inhouse-Schulungen** bei Ihnen anbieten.

**i** **Übersicht zur Modellauswahl**

In dieser Grafik werden die hier behandelten Modelle in einer Baumstruktur angeordnet, sodass ausgehend von den vorliegenden Daten in dieser Struktur das benötigte Modell identifiziert werden kann.

- [Erläuterung der Problemstellung](#)
- [Metrische abhängige Variable](#)
- [Lineare Regression und nichtlineare Regression](#)
  - [Das lineare Regressionsmodell](#)
  - [Beziehung der linearen Regression zur Anova](#)
  - [Nichtlineare Regression](#)
- [Analyse von Zähldaten](#)
  - [Poisson Regression](#)
  - [Negative Binomial Regression](#)
- [Kategoriale abhängige Variable](#)
  - [Dichotome oder multinomiale abhängige Variable](#)
  - [Logistische Regression \(Logit-Modell\)](#)
  - [Probitregression](#)
- [Ordinale abhängige Variable](#)
  - [Geordnete Probitregression](#)
  - [Geordnete logistische Regression](#)

## Metrische abhängige Variable

Dieses Kapitel zu der Behandlung von metrischen abhängigen Variablen enthält im ersten Teil die lineare und nichtlineare Regression. Im zweiten Teil wird auf den Umgang mit Zähldaten eingegangen.

# Lineare Regression und nichtlineare Regression

## Das lineare Regressionsmodell

Das **lineare Regressionsmodell** kann gewählt werden, wenn für die abhängige Variable und für die unabhängige/n Variable/n folgendes Skalenniveau vorliegt:

abhängige Variable ( $y$ )	metrisch
unabhängige/n Variable/n ( $x$ )	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $x_1, \dots, x_P$ ) dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Liegen mehrere unabhängige Variablen vor, so spricht man von einer multiplen Regression.

Mit einer linearen Regression wird die abhängige Variable  $y_i$  anhand einer oder mehrerer unabhängiger Variablen  $(x_{i,1}, \dots, x_{i,P})$  erklärt:  $y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_P x_{i,P} + \epsilon_i$ . In Matrixschreibweise erhält man: 
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,P} \\ 1 & x_{2,1} & \dots & x_{2,P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,P} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
  $(\beta_1, \dots, \beta_P)$  beschreiben dabei die Steigung der zu fittenden Gerade,  $(\beta_0)$  den y-Achsenabschnitt. Es ist darauf zu achten, dass die Regressionskoeffizienten  $(\beta_p)$  mit  $(p \in \{0, \dots, P\})$  nur in erster Potenz in das Modell eingehen, die unabhängigen Variablen können allerdings auch in anderen Potenzen in das Modell eingehen.

### Wann wird das lineare Regressionsmodell gewählt?

Das lineare Regressionsmodell wird gewählt, wenn davon ausgegangen werden kann, dass ein linearer Zusammenhang zwischen der/den unabhängige/n Variable/n und der abhängigen Variable besteht.

Im einfachen linearen Regressionsmodell erkennt man einen linearen Zusammenhang dadurch, dass durch die Punktwolke der paarweisen Messergebnisse im **Scatterplot** (Streudiagramm) gut durch eine Gerade gefittet werden kann. Dabei sollten die Messergebnisse möglichst nah um diese Gerade verteilt liegen und die Abstände der Messergebnisse zu der Gerade bei hohen wie niedrigen Ausprägungswerten der unabhängigen Variablen im Mittel möglichst gleich bleiben. Aber auch in anderen Fällen, bei denen im Scatterplot nicht direkt ein linearer Zusammenhang festgestellt werden kann, könnte die lineare Regression die richtige Wahl sein. So lässt sich die lineare Regression auch auf folgenden Zusammenhang anwenden:  $(y_i = \beta_0 + \beta_1 x_{i,1}^2 + \epsilon_i)$ . In diesem Fall würde die Verteilung der Punkte im Scatterplot einen quadratischen Zusammenhang nahelegen. Der Grund dafür, dass dieser Zusammenhang auch mittels linearer Regression beschrieben werden kann, ist, dass die unabhängige/n Variable/n auch in Potenzen verschieden von 1 vorliegen können (hier:  $(x_{i,1}^2)$ ). Die Linearität bezieht sich folglich nur auf die Regressionskoeffizienten  $(\beta_p)$  mit  $(p \in \{0, \dots, P\})$ .

Bei der linearen Regression werden folgende **Annahmen** getroffen:

- Die Fehlerterme  $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$  sind Zufallsvariablen mit Erwartungswert 0  $(E(\epsilon_i) = 0)$  und der Varianz  $(\sigma^2)$   $(V(\epsilon_i) = \sigma^2)$  (Homoskedastie).
- $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)$  sind unabhängig
- $(\epsilon_i)$  und  $(x_i)$  sind unkorreliert.
- Für die Überprüfung der Signifikanz über einen Test wird die Normalverteilung der Fehlerterme benötigt.

Eine genaue Erklärung zum linearen Regressionsmodell mit Beispielen und ausführlichen Umsetzungen in unterschiedlichen Statistik-Programmen kann man [hier](#) finden.

## Beziehung der linearen Regression zur Anova

Die Voraussetzungen für die einfaktoriellen ANOVA entsprechen genau den Annahmen, die wir für das lineare Regressionsmodell treffen (siehe vorheriger Abschnitt). Bei der einfaktoriellen ANOVA wird darauf getestet, ob die Mittelwerte der Gruppen (bezüglich des Faktors) gleich sind. Die Nullhypothese lautet also  $H_0: \mu_1 = \mu_2 = \dots = \mu_J$ . Der Test auf diese Nullhypothese mittels ANOVA ist ein Spezialfall des F-Tests (dient der Überprüfung der Gesamtsignifikanz des linearen Regressionsmodells): Ist  $\beta_p = 0 \forall p \in \{1, 2, \dots, P\}$  erfüllt, so entspricht der F-Test genau der ANOVA.

## Nichtlineare Regression

Die lineare und nichtlineare Regression unterscheiden sich nicht in den Skalenniveaus der verwendeten Variablen.

abhängige Variable $(y)$	metrisch
unabhängig e/n Variable/n $(x)$	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $x_1, \dots, x_P$ ) dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Auch in der nichtlinearen Regression wird wie bei der linearen Regression von einer metrisch skalierten abhängigen Variablen ausgegangen, jedoch ist der funktionale Zusammenhang in dieser Modellklasse nicht mehr linear in den zu schätzenden Parametern  $\beta$ . Das heißt, auch in nichtlinearen Modellen gilt  $E(Y|X=x) = f(x, \beta)$  aber  $f(x, \beta)$  entspricht nicht mehr der Identität  $g(x) = x$  wie in der linearen Regression. Beispielsweise könnte  $f(x, \beta) = \frac{\beta_0 x}{\beta_1 + x}$  annehmen. Diese Funktion kann nicht mehr als Linearkombination der beiden  $\beta_p, p=0,1$  dargestellt werden. Wichtige nichtlineare Funktionen sind Exponentialfunktionen, logarithmische oder auch trigonometrische Funktionen.

Ein Eindruck der Beziehung zwischen  $X$  und  $Y$  kann wie beim linearen Modell über Scatterplots gewonnen werden. Streuen die Punkte nicht um eine Gerade, kann das auf ein nichtlineares Modell hindeuten. Es muss jedoch beachtet werden, dass auch ein Plot, der mit dem linearen Modell  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}^2$  beschrieben werden kann, eine nichtlineare Beziehung zwischen  $Y$  und  $X$  anzeigt. Deshalb sei noch einmal darauf hingewiesen, dass sich der Begriff "nichtlinear" auf die zu schätzenden Parameter bezieht, nicht auf die erklärenden Variablen.

Bei einigen funktionalen Zusammenhängen gibt es die Möglichkeit durch Transformation wieder ein lineares Regressionsmodell zu erzeugen. Beispielsweise kann aus  $Y = a \exp(X \beta) + v$  durch logarithmieren der Gleichung  $\ln(Y) = \ln(a) + X \beta + \epsilon$ , mit  $\epsilon = \ln(v)$  erzeugt werden. Wichtig ist es, zu beachten, dass diese Transformationen auch die Fehlerterme  $\epsilon$  betreffen. Allgemeine Modellannahmen der linearen Regression bezüglich der Fehler müssen geprüft werden.

Anwendungsformen nichtlinearer Modelle finden sich vielfach in den Naturwissenschaften. Wachstumsanalyse, (Enzym-)Kinetik, aber auch die (linearisierbare) Cobb-Douglas Produktionsfunktion sind Beispiele für Anwendungsgebiete nichtlinearer Regressionsmodelle.

## Analyse von Zähldaten

In Zähldatenmodellen liegt die abhängige Variable ganzzahlig vor und nimmt nur nichtnegative Werte an  $(y_i \in \mathbb{N}_0)$ . Derartige Modelle geben an, wie oft das Ereignis von Interesse innerhalb eines Zeitraums aufgetreten ist. Beispiele für Daten, die mit einem Zähldatenmodell untersucht werden können, sind die Anzahl der Patente, die von einer Firma im Jahr angemeldet werden, oder die Anzahl der Kinder, die in einem Monat in einer Stadt geboren werden. Obwohl es sich um quantitative Daten handelt, ist die Modellierung mit bedingten Wahrscheinlichkeiten angebrachter als mit bedingten Erwartungswerten  $(E(y_i | x_i) = x_i \beta)$ . Letztere Modellierungen können unter Umständen negative Vorhersagen produzieren, was nicht sinnvoll ist, wenn  $y_i$  nur nichtnegative Werte annehmen kann. Dadurch ist ein lineares Regressionsmodell mit kleinster Quadrate Schätzung ungeeignet für diesen Datentyp. Im folgenden werden die gängigsten Modelle für Zähldaten vorgestellt.

## Poisson Regression

Es ist sinnvoll, eine Poisson Regression durchzuführen, falls angenommen werden kann, dass die abhängige Variable  $Y$  poissonverteilt ist.

Die Zähl-dichte der Poisson-Verteilung ist folgendermaßen definiert:  $P(Y=k) = \frac{\lambda^k}{k!} e^{-\lambda}$ ;  $k=0,1,2,\dots$  und  $\lambda > 0$ . Sie liefert eine Aussage darüber, wie groß die Wahrscheinlichkeit für  $k$  erfolgreiche Ereignisse bei  $n \rightarrow \infty$  Versuchsdurchführungen ist, wenn nur zwei verschiedene Ereignisse, Erfolg und Misserfolg, eintreten können.  $\lambda$  beschreibt dabei die mittlere Anzahl zu erwartender Ereignisse. Eine Poisson Regression kann zum Beispiel auf die Anzahl der Geburten in einer Stadt in einem Monat angewandt werden. Dabei ist  $\lambda$  die mittlere Anzahl von Geburten in einem Monat in dieser Stadt. Um eine Poisson Regression anwenden zu können, muss aber folgende Bedingung erfüllt sein: Der Erwartungswert und die Varianz dieser Verteilung sind jeweils  $\lambda$ .  $E(Y) = \lambda$  und  $Var(Y) = \lambda$ . Für das beschriebene Beispiel bedeutet dies, dass die erwartete Anzahl der Geburten in einem Monat in dieser Stadt der Varianz in den Geburtenanzahlen zwischen den Monaten entspricht. Diese Gleichheit von Varianz und Erwartungswert ist in der Anwendung jedoch öfters verletzt. Wenn diese Gleichheit in der Anwendung nicht vorliegt, so muss eine negative Binominal Regression (siehe nächster Abschnitt) durchgeführt werden.

Somit gilt für die Verteilung von  $Y_i | x_{(i)}$  mit  $i=1, \dots, n$ :  $Y_i | x_{(i)} \sim \text{Poisson}(\lambda_i)$ ;  $E(Y_i | x_{(i)}) = \lambda_i$  wird modelliert, indem angenommen wird, dass  $\lambda_i = h(x_{(i)}; \beta)$ .  $h$  ist hierbei die Responsefunktion, die prinzipiell frei gewählt werden kann. In den meisten Fällen wählt man für  $h$  die Exponentialfunktion und spricht dann von einer **Log-linearen Poisson Regression**. Für diese Regression ergibt sich dann folgende Gleichung:  $\lambda_i = e^{(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_P x_{i,P})}$ . Es ist darauf zu achten, dass es sich hier um ein heteroskedastisches Modell handelt, da die Varianzen sich für die unterschiedlichen Beobachtungen unterscheiden ( $Var(Y_i | x_{(i)}) = \lambda_i$ ).

Die Werte für den Parametervektor  $\beta = (\beta_1, \beta_2, \dots, \beta_P)^T$  lassen sich nicht mehr über die Methode der kleinsten Quadrate schätzen, sondern es muss die Maximum-Likelihood-Methode angewandt werden, um diesen Schätzer zu ermitteln.

## Negative Binomial Regression

Liegen Zähl-daten vor, bei denen die Gleichheit von Erwartungswert und Varianz verletzt ist, kann dies im Falle von Überdispersion (Varianz ist größer als Erwartungswert) mithilfe der negativen Binomial Regression modelliert werden. Ausgehend von einem Poissonmodell mit  $Y_i | x_{(i)} \sim \text{Poisson}(\lambda_i)$ , wobei  $\lambda_i = \exp(x_{(i)}; \beta)$ , könnte das Modell misspezifiziert sein, da nicht alle relevanten Informationen beobachtet werden können. Es wäre also denkbar, dass  $\lambda_i$  außer durch die erklärenden Variablen noch durch eine Störgröße beeinflusst wird, die nicht gemessen werden kann:

$$\lambda_i = \exp(x_{(i)}; \beta + \epsilon_i) = \lambda_i \cdot v_i; \quad v_i = \exp(\epsilon_i)$$

Der Parameter  $v_i$  beinhaltet dabei unbeobachtete Heterogenität, welche dazu führt, dass die Annahme der Equidispersion der Poisson-Verteilung verletzt ist (in diesem Fall  $Var(Y_i | x_{(i)}) > E(Y_i | x_{(i)})$ ). Deshalb wird im allgemeinen stattdessen angenommen:

$$Y_i | x_{(i)}, v_i \sim \text{Poisson}(\lambda_i); \quad v_i \sim (1, \sigma^2)$$

Da nur  $Y_i | x_{(i)}$  beobachtet werden können,  $v_i$  aber nicht, ist nur die auf  $x_{(i)}$  bedingte Verteilung  $Y_i | x_{(i)} \sim (\lambda_i, \lambda_i + \sigma^2)$  von Interesse. Deshalb integriert man den unbeobachtbaren Parameter aus der Wahrscheinlichkeitsverteilung heraus. In der Praxis nimmt man für  $v_i$  oft eine Gamma-Verteilung unabhängig von  $x_{(i)}$  an, da die Integration dann eine geschlossene Lösung hat. Hieraus erhält man die Wahrscheinlichkeitsverteilung der negativen Binomialverteilung, welche als Verteilung der Anzahl von Erfolgen in einer Folge von Bernoulli Versuchen mit gleicher Erfolgswahrscheinlichkeit, bevor eine bestimmte Anzahl von Misserfolgen aufgetreten ist, angesehen werden kann.

$$f(y|x) = \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)} \left(\frac{\lambda}{\lambda+\alpha}\right)^y \left(\frac{\alpha}{\lambda+\alpha}\right)^\alpha; \quad E(y) = \lambda; \quad Var(y) = \lambda + \frac{\lambda^2}{\alpha}$$

Mit geeigneter Wahl für  $\alpha$  erhält man schließlich verschiedene negative Binomial Modelle. Verwiesen sei hier auf die zwei geläufigsten Modelle **Negbin I** und **Negbin II**:

- **Negbin I:**  $\alpha$  kann zwischen Beobachtungen variieren:  $\alpha^{-1} = \frac{\sigma^2}{\lambda}$ , was zu  $Var(Y_i | x_{(i)}) = \lambda + \sigma^2 \lambda = (1 + \sigma^2) \lambda \exp(x_{(i)}; \beta)$  führt
- **Negbin II:**  $\alpha$  ist konstant:  $\alpha^{-1} = \sigma^2$ , was zu  $Var(Y_i | x_{(i)}) = \lambda + \sigma^2 \lambda^2 = \exp(x_{(i)}; \beta) + \sigma^2 (\exp(x_{(i)}; \beta))^2$  führt

Diese Modelle werden wie die Poissonregression mit der Maximum Likelihood Methode geschätzt. Für  $\sigma^2 \rightarrow 0$  konvergieren die negativen Binomialmodelle gegen das Poissonmodell, was die Grundlage für den Likelihood-Ratio oder den Wald Test bildet, um das richtige Modell für Zähl-daten auszuwählen.

Der Vorteil einer negativen Binomial Regression gegenüber einem Poissonmodell ist die Effizienz der Schätzung. Sind die getroffenen Annahmen jedoch verletzt führt das im Allgemeinen zu inkonsistenten Schätzergebnissen.

## Kategoriale abhängige Variable

Bei kategoriale skalierten abhängigen Variablen  $(y_i)$  kommen meist generalisierte lineare Modelle zur Anwendung. Eine wichtige Annahme des [linearen Regressionsmodells](#), Normalverteilungsannahme der Störterme, ist in Modellen mit diskreten erklärten Variablen nicht immer gerechtfertigt. Bei Modellen der generalisierten linearen Klasse kann die Verteilung der Fehlerterme zu anderen Verteilungen der Exponentialfamilie gehören. Das bedeutet, sie können unter anderem normal-, binomial-, bernoulli-, oder poissonverteilt sein.

## Dichotome oder multinomiale abhängige Variable

Wenn die abhängige Variable, die untersucht werden soll, kategoriale skaliert ist, jedoch keine aufsteigende Reihenfolge der Kategorien gebildet werden kann (z.B. Geschlecht, Präferenz einer Automarke), spricht man von nominalem Skalenniveau. Die gängigsten Methoden im Umgang mit solchen Variablen finden sich im folgenden Abschnitt.

## Logistische Regression (Logit-Modell)

Bei der logistischen Regression können die unabhängige/n Variable/n Variablen jedes beliebige Skalenniveau annehmen und müssen auch nicht innerhalb der einzelnen unabhängigen Variablen  $(x_1, \dots, x_P)$  einheitlich sein. Die abhängige Variable nimmt allerdings nur diskrete Werte an. Meist liegt die abhängige Variable binomial  $(\left(Y_i | x_{(i)}\right) \sim \text{mathcal{Ber}}(p_i))$  vor, d.h. es treten nur zwei unterschiedliche Ausprägungen, "0" und "1", auf. Die Fehlerterme werden bei diesem Modell als logistisch verteilt angenommen. Falls allerdings die abhängige Variable multinomial  $(\left(Y_i | x_{(i)}\right) \sim \text{operatorname{Categorical}}(p_{i,1}, \dots, p_{i,m}))$  vorliegt (es treten mehr als zwei unterschiedliche Ausprägungen auf), kann eine verallgemeinerte Version, das multinomiale logistische Regressionsmodell, verwendet werden.

abhängige Variable $(y)$	dichotom (binomial), multinomial
unabhängig e/n Variable/n $(x)$	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $(x_1, \dots, x_P)$ dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Eine Fragestellung, bei der sich eine logistische Regression anbieten würde, wäre beispielsweise, welche Faktoren die Wahrscheinlichkeit beeinflussen, dass eine Person eine Arbeitsstelle hat. In diesem Fall würde man als abhängige Variable eine binomiale 0-1 kodierte Variable verwenden, wobei 1 für Erwerbstätigkeit und 0 für Arbeitslosigkeit steht.

Das Ziel der logistischen Regression ist die Vorhersage der Wahrscheinlichkeit, mit der ein bestimmtes Ereignis (unter Verwendung von Einflussfaktoren) eintritt.

Das (binomiale) logistische Regressionsmodell ist durch folgende Gleichung gegeben:

$$P(y_i=1 | X=x_{(i)}) = G(x'_{(i)}\beta) = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_P x_{i,P})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_P x_{i,P})} = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_P x_{i,P})}{1 + \exp(-\beta_0 - \beta_1 x_{i,1} - \dots - \beta_P x_{i,P})}, \text{ für } i \in \{1, \dots, n\}$$

Die Parameter  $(\beta_p)$  werden mit der Maximum-Likelihood-Methode geschätzt, da eine direkte Berechnung mittels kleinster Quadrate (siehe [lineare Regression](#)) nicht möglich ist. Die Schätzwerte werden anhand iterativer Verfahren, wie dem Newton-Raphson Algorithmus, ermittelt. Da die log-Likelihood Funktion des logistischen Regressionsmodells überall konkav ist, existiert ein eindeutiger Maximum-Likelihood Schätzer für die zu bestimmenden Parameter.

Die Interpretation der marginalen Effekte dieser Modellklasse unterscheidet sich deutlich vom linearen Regressionsmodell. Die marginalen Effekte der Logitregression entsprechen dem Produkt aus geschätztem Parameter und Wahrscheinlichkeitsdichte des Modells:

$$\frac{\partial P(y_i=1 | X=x_{(i)})}{\partial x_j} = g(x'_{(i)}\beta) \beta_j$$

wobei  $g(z) = \frac{\partial G(z)}{\partial z}$ . Die marginalen Effekte sind also immer von den Ausprägungen aller unabhängigen Variablen abhängig. Da Wahrscheinlichkeitsdichten immer positiv sind, gibt das Vorzeichen des geschätzten Parameters die Richtung des Effekts auf die bedingte Wahrscheinlichkeit an.

Da die marginalen Effekte nicht konstant und deshalb keiner so direkten Interpretation wie im linearen Modell zugänglich sind, werden oft die sogenannten Odds oder die Oddsratio betrachtet. Dabei werden die Odds (für ein kleines Modell mit zwei zu schätzenden Parametern) als  $\text{odds} = \exp(\beta_0 + \beta_1 x)$  und die Oddsratio als

$$\text{OR} = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{\frac{G(x+1)}{1-G(x+1)}}{\frac{G(x)}{1-G(x)}} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1)$$

dargestellt. Äquivalent dazu ist die Gleichung

$$\ln(\text{OR}) = \beta_1$$

Wird  $x_1$  ceteris paribus um eine Einheit erhöht (alle anderen erklärenden Variablen verbleiben auf dem alten Wert), verändern sich die Odds um  $(\beta_1 \cdot 100\%)$ .

Hat die abhängige Variable mehr als zwei Ausprägungen ( $J + 1$ ), ist also multinomial skaliert, wird das multinomiale Logitmodell verwendet. Wenn die Fehlerterme unabhängig und identisch nach der Gumbel Verteilung verteilt sind, ergibt sich als Modellgleichung für die Wahrscheinlichkeit, dass  $y_i$  die Ausprägung  $j$  annimmt:

$$P(y_i = j | X = x_i) = p_{ij} = \frac{\exp(x_i' \beta_j)}{1 + \sum_{h=1}^J \exp(x_i' \beta_h)}, \text{ for all } j \in \{1, \dots, J\}$$

Hierbei ist zu beachten, dass zur Parameteridentifikation eine Basiskategorie derart angenommen werden muss, dass beispielsweise gilt  $(\beta_0 = 0)$ . Sonst können die Parameter nicht eindeutig geschätzt werden. Anders ausgedrückt reicht es,  $J$  Wahrscheinlichkeiten zu berechnen, um  $J + 1$  Wahrscheinlichkeiten zu bestimmen, da sie sich insgesamt zu eins addieren müssen. Im Fall von  $J + 1 = 2$  landet man wieder beim Standard logistischen Modell (siehe oben).

## Probitregression

abhängige Variable ( $y$ )	dichotom (binomial), multinomial
unabhängig e/n Variable/n ( $x$ )	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $x_1, \dots, x_P$ ) dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Wie bei der logistischen Regression, geht man von einer dichotomen  $(Y_i | x_i) \sim \text{Ber}(p_i)$  oder auch multinomialen  $(Y_i | x_i) \sim \text{Categorical}(p_{i,1}, \dots, p_{i,m})$  abhängigen Variable aus. Der Unterschied zwischen den beiden Modellen liegt in der Annahme über die Verteilung der Fehlerterme  $(\epsilon_i)$ , denn im Probitmodell werden standardnormalverteilte Residuen angenommen. Im allgemeinen motiviert man ein Probitmodell über die Annahme einer latenten Zufallsvariable  $(Y^*_i)$  mit  $(Y^*_i = x_i' \beta + \epsilon_i)$  mit  $(\epsilon_i | x_i) \sim \text{N}(0, 1)$ . Dann kann die beobachtete dichotome abhängige Variable als Indikator dafür betrachtet werden, ob  $(Y^*_i)$  größer als null ist. Aus dieser Herangehensweise ergibt sich:

$$P(y_i = 1 | X = x_i) = p_i = \Phi(x_i' \beta), \text{ for all } i \in \{1, \dots, n\}$$

wobei  $(\Phi(\cdot))$  für die kumulierte Verteilungsfunktion der Standardnormalverteilung steht. Das sorgt dafür, dass die Wahrscheinlichkeit nur zwischen 0 und 1 liegen kann.

Auch das Probitmodell wird über die Maximum-Likelihood-Methode berechnet, wobei es keine analytische Lösung der Gleichungen gibt und sie beispielsweise mit dem Newton-Raphson Verfahren näherungsweise gelöst werden. Die log-Likelihood Funktion des Probitmodells ist überall konkav, was die Existenz eines eindeutigen ML-Schätzers garantiert. Da sich das Probit- und das Logitmodell vor allem über die Wahl der Linkfunktion unterscheiden, können auch beim Probitmodell die geschätzten Parameter  $\beta$  nicht als marginale Effekte interpretiert werden. Die Effekte hängen ebenso von der Wahrscheinlichkeitsdichte (hier Dichte der Standardnormalverteilung) ab. Das Vorzeichen des geschätzten Parameters gibt aber die Richtung des Effekts an.

### **i** Praktischer Hinweis für Logit- und Probitmodelle

Im Unterschied zum linearen Modell, muss mit heteroskedastischen Fehlern anders umgegangen werden. Das Vorliegen von Heteroskedastie im Modell (darauf sollte getestet werden) muss beim Aufstellen der log-Likelihood beachtet werden. Wird ein Modell unter Nichtbeachtung gerechnet, sind die Parameterschätzer inkonsistent. Es reicht daher nicht aus, nach der Schätzung mit beispielsweise der Methode der Eicker-Huber-White Standardfehler eine konsistente Kovarianzmatrix zu erzeugen.)

Im allgemeineren Fall mit mehr als zwei möglichen Ausprägungen ( $J + 1$ ) der abhängigen Variable kann das multinomiale Probitmodell angewendet werden. Im Unterschied zum multinomialen Logitmodell werden die Fehlerterme  $(\epsilon_i = (\epsilon_{i,0}, \dots, \epsilon_{i,J}))' \sim \mathcal{N}(0, \Sigma)$  als gemeinsam normalverteilt (möglicherweise paarweise korreliert angenommen). Handelt es sich bei der Varianz-Kovarianzmatrix um die Einheitsmatrix, spricht man vom unabhängigen Probitmodell.

## Ordinale abhängige Variable

Liegt eine kategoriale abhängige Variable vor, deren Ausprägungen in eine Reihenfolge gebracht werden können (z.B. Zufriedenheit mit einem Produkt, Schulnoten), handelt es sich um eine ordinal skalierte Variable, deren Beziehung zu erklärenden Variablen häufig mit geordneten Modellen beschrieben wird. Theoretisch ist es möglich, anstatt von geordneten Regressionsmodellen, ein multinomiales Modell aus Teil 2.1 zu verwenden. Dadurch erhält man aber ineffiziente Schätzer, da die zusätzliche Information durch die bestehende Reihenfolge nicht beachtet wird. Die Analyse mit einem [linearen Regressionsmodell](#) ist unangemessen, da der Unterschied zwischen "1" und "2" nicht gleich dem Unterschied zwischen "8" und "9" ist. Die Abstände der Ausprägungen sind also nicht interpretierbar, wie es im linearen Regressionsmodell mit metrischer abhängiger Variable der Fall ist.

## Geordnete Probitregression

abhängige Variable ( $y$ )	ordinal (Reihenfolge in Ausprägungen liegt vor)
unabhängige Variable/n ( $x$ )	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $(x_1, \dots, x_P)$ dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Ist man beispielsweise daran interessiert, die Medaillenvergabe an Sportler bei den olympischen Spielen vorherzusagen, liegt eine ordinal skalierte abhängige Variable vor (kein Podiumsplatz, Bronze, Silber, Gold). Diese Vorhersage kann mithilfe unabhängiger Variablen wie Größe, Muskelkraft, Lungenvolumen, Essverhalten, usw. mit einer geordneten Probitregression getroffen werden.

Ausgehend von einer metrischen, nicht beobachtbaren (latenten) Variable  $(y^*)$  mit Modell:  $(y^*_i = x'_i \beta + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, 1) \text{ i.i.d.}; i=1, \dots, n)$  nimmt man folgende Beziehung zur vorliegenden abhängigen Variable (mit endlichen Anzahl an Kategorien ( $J + 1$ )) an:

$$y_i = \begin{cases} 0, & \text{für } -\infty < y^*_i \leq \mu_1 \\ 1, & \text{für } \mu_1 < y^*_i \leq \mu_2 \\ \vdots \\ J, & \text{für } \mu_{J-1} < y^*_i \leq \infty \end{cases}$$

Dabei stehen  $(\mu_j)$  für geordnete Schwellenwerte, die neben den  $(\beta)$  die zu schätzenden Parameter des Modells darstellen. Nimmt man nun  $(\mu_0 = -\infty)$  und  $(\mu_{J+1} = \infty)$  an, kann die die Wahrscheinlichkeit des Eintretens der jeweiligen Kategorie abhängig von den erklärenden Variablen dargestellt werden mit:

$$\begin{aligned} p_{ij} &= P(y_i = j | x_{(i)}) = P(\mu_{j-1} < y^*_i \leq \mu_j | x_{(i)}) = P(\mu_{j-1} < x'_i \beta + \epsilon_i \leq \mu_j | x_{(i)}) \\ &= P(\mu_{j-1} - x'_i \beta < \epsilon_i \leq \mu_j - x'_i \beta | x_{(i)}) = \Phi(\frac{\mu_j - x'_i \beta - \Phi(\mu_{j-1} - x'_i \beta)}{\sigma}) - \Phi(\frac{\mu_{j-1} - x'_i \beta - \Phi(\mu_{j-1} - x'_i \beta)}{\sigma}) \end{aligned}$$

Dabei steht  $(\Phi(\cdot))$  für die kumulierte Verteilungsfunktion der Standardnormalverteilung. Aus Identifikationsgründen darf in der Designmatrix (Matrix der erklärenden Variablen) keine Konstante enthalten sein. Wäre das der Fall, könnten die Schwellenwerte davon nicht unterschieden werden und sie blieben dadurch unidentifiziert. Die Parameter werden durch die Maximum-Likelihood Methode geschätzt, wobei die log-Likelihood überall konkav ist. Dadurch ist ein eindeutiger ML-Schätzer bestimmt. Die ML-Schätzer sind hierbei konsistent, asymptotisch effizient und asymptotisch normalverteilt. Die Interpretation der marginalen Wahrscheinlichkeitseffekte gestaltet sich etwas schwieriger als in einem multinomialen Modell. Der Effekt hängt sowohl vom geschätzten Parameter, als auch von der Differenz von Wahrscheinlichkeitsdichten ab. Auch das Vorzeichen des Effekts lässt sich nur im Falle der ersten oder letzten Kategorie eindeutig über das Vorzeichen des jeweiligen Schätzers bestimmen. Im allgemeinen Fall gilt:  $(\frac{\partial p_{ij}}{\partial x_{ik}}) = \beta_k (\phi(\mu_j - x'_i \beta) - \phi(\mu_{j-1} - x'_i \beta))$ , wobei  $(\phi(\cdot))$  für die Wahrscheinlichkeitsdichte der Standardnormalverteilung steht.

## Geordnete logistische Regression

abhängige Variable ( $y$ )	ordinal (Reihenfolge in Ausprägungen liegt vor)
unabhängige Variable/n ( $x$ )	beliebiges Skalenniveau (die Skalenniveaus der einzelnen $x_1, \dots, x_P$ dürfen sich auch unterscheiden, liegt eine multinomiale Variable vor, so muss eine Zerlegung in Dummy-Variablen stattfinden)

Die geordnete logistische Regression folgt den gleichen Überlegungen wie die geordnete Probitregression. Der Unterschied liegt in der Annahme über die Verteilung der Fehlerterme, denn sie werden wie im binomialen oder multinomialen Fall (siehe oben) als logistisch verteilt angenommen. Daraus ergibt sich für die bedingten Wahrscheinlichkeiten der jeweiligen Kategorien:

$$p_{ij} = \frac{\Lambda(\mu_{j+1} - x'_{ij}\beta) - \Lambda(\mu_j - x'_{ij}\beta)}{\Lambda(\mu_{j+1} - x'_{ij}\beta) - \Lambda(\mu_j - x'_{ij}\beta)}$$

wobei  $\Lambda(\cdot)$  für die kumulierte Verteilungsfunktion der logistischen Verteilung steht. Auch hier gilt, dass keine Konstante im Modell enthalten sein darf, da die Schwellenwerte  $\mu_j$  sonst nicht zu identifizieren sind. Die Schätzung und Interpretation der Parameter verläuft analog zur Probitregression.