31 Logistische Regression

Henning Best und Christof Wolf

[a] Universität Mannheim
[b] GESIS – Leibniz-Institut für Sozialwissenschaften und Universität Mannheim

Zusammenfassung. Die logistische Regression ist ein multivariates Analyseverfahren zur Analyse von dichotomen abhängigen Variablen, d. h. binären Variablen mit zwei Ausprägungen. Aus einer linearen Modellierung der logarithmierten Odds (Logits) des Auftretens von $x = 1$ ergibt sich eine nichtlineare Modellierung der Wahrscheinlichkeiten. Wir werden sehen, dass diese Nichtlinearität zwar einerseits notwendig und sinnvoll ist, andererseits aber auch zu substanziellen Unterschieden in der Interpretation im Vergleich zu OLS-Regressionsverfahren führt. Im vorliegenden Beitrag wird zunächst eine Einführung in die Logik des Verfahrens gegeben und die Interpretation der Ergebnisse vorgestellt. In einem zweiten Schritt werden grundlegende mathematische Eigenschaften der logistischen Regression dargestellt und fortgeschrittene Erweiterungen diskutiert (Standardisierung, Effekte auf die Wahrscheinlichkeiten, Interaktionen). Die Anwendung der logistischen Regression wird daraufhin am Beispiel der Bildungsvererbung praktisch dargestellt. Im letzten Abschnitt wird auf häufige Fehler, insbesondere in der Interpretation, hingewiesen (Odds-Ratios, Nichtlinearität, Interaktionen).

1 Einführung in das Verfahren

In der sozialwissenschaftlichen Forschung werden häufig Tatbestände untersucht, die in dichotomen Variablen[1] abgebildet werden. Typische Untersuchungsgegenstände sind etwa Entscheidungen oder daraus resultierende Zustände. So untersuchen beispielsweise Hubert & Wolf (2007) mit logistischen Regressionen die Determinanten der Teilnahme an beruflicher Weiterbildung, und Best (2008) analysiert die Entscheidung von Landwirten, ihren Betrieb auf ökologische Landwirtschaft umzustellen. Gängige Untersuchungsgegenstände sind z. B. auch Arbeitslosigkeit oder der (Hoch-)Schulabschluss.

Eine dichotome Variable ist dadurch gekennzeichnet, dass sie nur zwei Zustände annehmen kann. Wie bei den meisten regressionsbasierten Verfahren ist es sinnvoll, wenn die Variable 0/1-codiert ist. Nehmen wir an, die Variable soll bezeichnen, ob eine Person Abitur hat oder nicht, so würde man sinnvollerweise „Abitur“ mit „1“ und „kein Abitur“ mit „0“ codieren (also „Abitur ja/nein“).

[1] Dichotome Variablen sind Variablen mit zwei Ausprägungen. Als alternative Bezeichnung wird häufig auch „binäre Variable“ verwendet.

S. 827–854 in: Christof Wolf & Henning Best, Hg. (2010). Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften

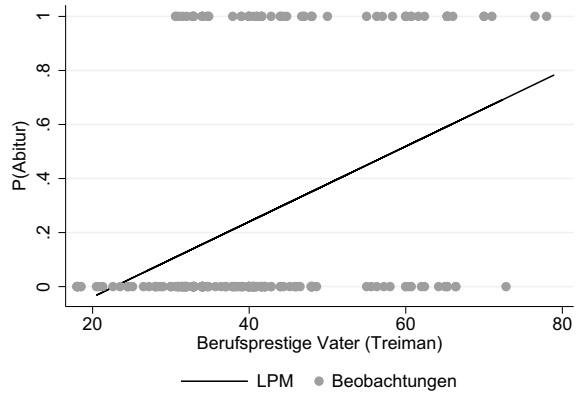


Abb. 1: Lineares Wahrscheinlichkeitsmodell

1.1 Das lineare Wahrscheinlichkeitsmodell

Eine recht einfache Möglichkeit, die Determinanten der Hochschulreife multivariat zu untersuchen, wäre es, eine OLS-Regression mit der Variable „Abitur ja/nein“ zu schätzen. Die geschätzte abhängige Variable $\hat{y}$ ist zwar nicht mehr dichotom, sondern metrisch, kann jedoch mit der Variable „Abitur ja/nein“ in Bezug gesetzt werden, indem man sie als Wahrscheinlichkeit interpretiert, dass der Befragte Abitur hat. Entsprechend wird angenommen, dass die unabhängigen Variablen die Auftrittswahrscheinlichkeit $P(y = 1) = \hat{y}$ linear beeinflussen:

$$P(y = 1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon\,. \tag{1}$$

Aufgrund der Interpretation von $\hat{y}$ als Auftrittswahrscheinlichkeit und der linearen Modellierung wird dieses Verfahren als „linares Wahrscheinlichkeitsmodell“ (LPM, linear probability model) bezeichnet. Schätzt man beispielsweise mit dem ALLBUS[2] die Wahrscheinlichkeit, dass ein Befragter Abitur hat anhand des beruflichen Prestiges des Vaters, ergibt sich die in Abbildung 1 gezeigte Regressionsgerade.[3]

Wie man sieht, steigt nach diesem Modell die Wahrscheinlichkeit, das Abitur zu erwerben mit dem Berufsprestige des Vaters an. Allerdings führt die Anwendung des LPM auch zu einer Reihe von ernstzunehmenden Problemen (siehe z. B. Menard 1995, S. 1–11 für eine ausführlichere Diskussion):

– Für bestimmte Ausprägungen von x können Werte geschätzt werden, die außerhalb des definierten *Wertebereichs* der Wahrscheinlichkeiten liegen ($0 \leq P(y = 1) \leq 1$). Im vorliegenden Beispiel wird etwa eine negative Wahrscheinlichkeit für Treiman-Prestigewerte unter 25 Punkten vorhergesagt.

[2] Kumulation 1980–2006, nur über 21-Jährige, Westdeutschland.
[3] Wir verwenden hier die Skala von Treiman (1977).

- Da die beobachtete abhängige Variable nur Werte von 0 und 1 annehmen kann, ist die Varianz des Fehlerterms bei linearer Modellierung abhängig von der jeweiligen Ausprägung der unabhängigen Variablen (*Heteroskedastizität*). Dies wiederum führt zu einer ineffizienten OLS-Schätzung und verzerrten Standardfehlern.
- Für gegebene Werte der unabhängigen Variablen können die Residuen nur zwei Werte annehmen, so dass die *Normalverteilungsannahme der Residuen* verletzt wird.
- Die funktionale Form des LPM, also die lineare Parametrisierung, ist insbesondere in den Randbereichen nicht angemessen. Es ist anzunehmen, dass sich die Wahrscheinlichkeiten den Extremwerten 0 und 1 nicht linear, sondern allmählich annähern (siehe auch Abbildung 2 auf Seite 831 und die folgenden Ausführungen).

Zwar gibt es eine einfache Erweiterung des LPM, um das Über- und Unterschreiten des Wertebereichs zu verhindern (eine Spline-Funktion mit $P(y = 1) = 0$ für $\hat{y} \leq 0$ und $P(y = 1) = 1$ für $\hat{y} \geq 1$), eine solche Trunkierung der Regressionsgleichung ist jedoch nicht optimal. Dies gilt insbesondere, da die anderen Probleme der linearen Modellierung und OLS-Schätzung nicht behoben werden.

1.2 Die logistische Regression

Wenige Transformationen und ein Wechsel auf Maximum-Likelihood-Schätzung erlauben es, die zwangsläufige Verletzung der statistischen Grundannahmen von OLS zu vermeiden und gleichzeitig eine funktionale Form zu finden, die besser an die Modellierung von Wahrscheinlichkeiten angepasst ist.

Auch in der logistischen Regression wird als abhängige Variable nicht die beobachtete dichotome Variable modelliert, sondern die unbeobachteten Auftrittswahrscheinlichkeiten. Zwei einfache Umformungen führen jedoch zu einer Variable, die im Gegensatz zu Wahrscheinlichkeiten einen Wertebereich von $-\infty$ bis $+\infty$ aufweist. Wir werden sehen, dass die Umformungen gleichzeitig zu einer funktionalen Form führen, die Wahrscheinlichkeitsverläufe sinnvoll darstellen kann.

Odds statt Wahrscheinlichkeiten: Um den Wertebereich der abhängigen Variablen auf $+\infty$ auszudehnen, werden statt Wahrscheinlichkeiten *Odds* betrachtet. Odds werden häufig für Gewinnquoten z. B. bei Pferdewetten verwendet und sind als

$$O = \frac{P}{1 - P} \tag{2}$$

definiert, stehen also für das Verhältnis der Eintrittswahrscheinlichkeit zur Gegenwahrscheinlichkeit. Einer Wahrscheinlichkeit von 10 % entsprechen demnach Odds von $^{10}/_{90} = 0{,}11$, 50 % sind $^{50}/_{50} = 1$ und 90 % ergeben Odds von $^{90}/_{10} = 9$. Odds sind zwischen 0 und $+\infty$ definiert, wobei gilt, dass die Odds sich $+\infty$ annähern, wenn die Wahrscheinlichkeit sich 100 % nähert. Somit ist die Obergrenze des ursprünglichen Wertebereiches beseitigt. Hieraus folgt gleichzeitig, dass die Transformation nicht linear ist (z. B. $P = 99 \rightarrow O = 99; P = 99{,}9 \rightarrow O = 999$).

Logits statt Odds: Um auch die feste Untergrenze zu beseitigen, werden in der logistischen Regression die Odds logarithmiert. Man erhält die so genannten Logits:

$$\text{Logit} = \ln O = \ln \frac{P}{1-P}\,. \tag{3}$$

Die Logarithmusfunktion hat die Eigenschaft, dass ihre Nullstelle bei 1 liegt, so dass Werte kleiner 1 durch Logarithmieren negativ, Werte größer 1 positiv werden.[4] Das heißt, Odds zwischen 0 und 1 – also Wahrscheinlichkeiten unter 0,5 – werden auf den Wertebereich $-\infty$ bis unter 0 der Logit-Skala abgebildet. Odds zwischen 1 und $+\infty$ hingegen – also Wahrscheinlichkeiten zwischen 0,5 und 1 – werden auf den Wertebereich zwischen 0 und $+\infty$ transformiert. Wenn man bedenkt, dass einer Wahrscheinlichkeit von 0,5 Odds von 1 entsprechen, ist diese Eigenschaft durchaus wünschenswert. Damit ist auch die Untergrenze des ursprünglichen Wertebereiches beseitigt.

Die logistische Regression verwendet nun die logarithmierten Odds von $P(y = 1|x)$, also die Logits, als abhängige Variable. Die Logits werden mit einer einfachen linearen Funktion modelliert, so dass die rechte Seite der Regressionsgleichung

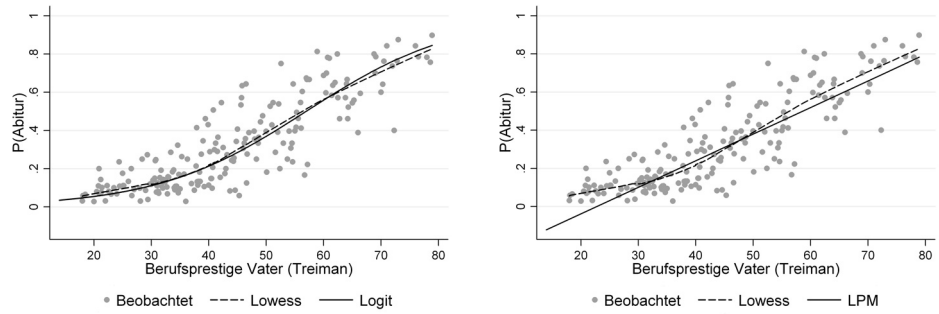
$$\text{Logit} = \ln \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4}$$

auf den ersten Blick identisch mit der des linearen Wahrscheinlichkeitsmodells ist. Der wesentliche Unterschied ist jedoch, dass die lineare Modellierung sich nicht – wie im LPM – auf Wahrscheinlichkeiten, sondern auf Logits bezieht. Möchte man sehen, wie Wahrscheinlichkeiten in der logistischen Regression modelliert werden, muss Gleichung (4) nach P aufgelöst werden (umstellen und entlogarithmieren):

$$P(y = 1) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} = \frac{e^{\text{Logit}}}{1 + e^{\text{Logit}}}\,. \tag{5}$$

Die Basisgleichung der logistischen Regression ist somit linear in Bezug auf die *Logits*, aber nichtlinear in Bezug auf die Wahrscheinlichkeiten. Dies mag zwar zunächst unpraktisch erscheinen, ist aber notwendig um erstens – wie beschrieben – Probleme mit der Ober- und Untergrenze des Wertebereichs zu vermeiden und zweitens eine adäquate Abbildung des Wahrscheinlichkeitsverlaufs zu ermöglichen. Um dies zu demonstrieren, stellt Abbildung 2 die Regressionsgerade des linearen Wahrscheinlichkeitsmodells und die nichtlineare Modellierung der Wahrscheinlichkeit im logistischen Regressionsmodell einander gegenüber. Als Referenz dient jeweils eine nichtparametrische Regressionskurve (Lowess), die an die relative Häufigkeit des Abiturs und das Prestige der Väter angepasst wurde. Hierfür wurde für jeden Prestigewert der Väter die relative Häufigkeit, dass Söhne Abitur haben, im kumulierten ALLBUS berechnet. Man sieht, dass beide Modellierungen (logistisch und linear) relativ gut geeignet sind, um die erwarteten Wahrscheinlichkeiten *in der Mitte des Wertebereichs* von x abzubilden. An den Rändern hingegen (im vorliegenden Beispiel insbesondere am unteren Rand) ist das LPM nicht mehr zu einer angemessenen Abbildung des Wahrscheinlichkeitsverlaufs in der Lage. Die logistische Kurve hingegen approximiert die allmähliche Annährung der Wahrscheinlichkeiten an null bzw. eins sehr gut.

[4] Der Logarithmus ist nur für Werte größer null definiert.



(a) Lineares Wahrscheinlichkeitsmodell (b) Logistisches Regressionsmodell

Abb. 2: Relative Häufigkeit des Abiturs nach Prestige des Vaters

1.3 Interpretation der Koeffizienten

Aus der in Gleichung (4) dargestellten Formulierung der logistischen Regression ergibt sich, dass sich die Regressionskoeffizienten grundsätzlich auf die *logarithmierten Odds* bzw. *Logits* beziehen. In Bezug auf die Logits verläuft die Interpretation jedoch analog zur OLS-Regression (siehe auch Kapitel 24 in diesem Handbuch): Die Regressionskonstante β_0 ist der y-Achsenabschnitt und gibt an, wie hoch der Logit ist, wenn alle unabhängigen Variablen den Wert null annehmen. Die Steigung der Geraden wird durch die Regressionskoeffizienten β_i dargestellt, so dass sich die logarithmierten Odds um β_i Einheiten verändern, wenn x_i um eine Einheit steigt (unter Konstanthaltung der jeweils anderen unabhängigen Variablen). Entsprechend steht ein negativer β-Koeffizient für einen negativen Zusammenhang zwischen der unabhängigen Variablen und den Logits (fallende Regressionsgerade, je größer x, desto kleiner die Logits), ein positiver Koeffizient für einen positiven Zusammenhang (steigende Regressionsgerade, je größer x, desto größer die Logits). Bei allen augenscheinlichen Parallelen zur OLS-Regression ist jedoch unbedingt zu beachten, dass logarithmierte Odds inhaltlich nicht interpretierbar sind, da eine nichtlineare Verknüpfung zu den Wahrscheinlichkeiten besteht (man erinnere sich: $P = {}^{O}/_{1+O}$). Hieraus folgt, dass auch die β-Koeffizienten der logistischen Regression inhaltlich kaum sinnvoll zu interpretieren sind. Lediglich eine Angabe der Richtung des Zusammenhangs ist möglich, da die Verknüpfung zwischen Logits und Wahrscheinlichkeiten trotz ihrer Nichtlinearität vorzeichenwahrend ist, also ein monotoner Zusammenhang zwischen Logits und Wahrscheinlichkeiten besteht.

Aufgrund dieser Schwierigkeiten wird vielfach vorgeschlagen, statt der β-Koeffizienten die entlogarithmierte Variante e^{β} zu verwenden (sog. „Effektkoeffizienten“ oder „Odds-Ratios“). In der Tat ergibt sich aus einer Entlogarithmierung der Regressionsgleichung (4) eine Modellierung der Odds:

$$O = e^{\text{Logit}} = e^{\ln \frac{P}{1-P}} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\beta_2 x_2} \cdot (\cdots) \cdot e^{\beta_k x_k} \,. \quad (6)$$

Tab. 1: Logits, Odds, Wahrscheinlichkeiten und ihre Veränderung nach Prestige

Prestige	L	ΔL	O	OR	P	ΔP
20	−2,8484		0,0579		0,0548	
30	−2,0776	+0,7708	0,1252	×2,1615	0,1113	+0,0565
40	−1,3068	+0,7708	0,2707	×2,1615	0,2130	+0,1017
50	−0,5360	+0,7708	0,5851	×2,1615	0,3691	+0,1561
60	0,2348	+0,7708	1,2647	×2,1615	0,5584	+0,1893
70	1,0056	+0,7708	2,7335	×2,1615	0,7322	+0,1737

Wie man sieht, wird durch die Entlogarithmierung aus dem ehemals linear-additiven Modell ein multiplikatives Modell. Folglich sind e^{β}-Koeffizienten grundlegend anders zu interpretieren als β-Koeffizienten: Erstens beziehen sie sich nicht mehr auf Veränderungen der Logits, sondern eben auf Odds. Zweitens geben sie eine faktorielle, d. h. multiplikative Veränderung an („factor change“). So bedeutet $e^{\beta_i} = 2$ beispielsweise, dass sich die Odds für $y = 1$ verdoppeln (Multiplikation mit 2), wenn sich x_i um eine Einheit erhöht; entsprechend steht $e^{\beta_i} = 0{,}33$ für eine Verringerung der Odds um $^2/_3$ bei einem Ansteigen von x_i um eine Einheit (Multiplikation mit 0,33). Ein $e^{\beta} < 1$ zeigt demnach einen negativen Zusammenhang an, $e^{\beta} > 1$ steht für eine positive Beziehung der beiden Variablen, und der neutrale Wert ist 1 (kein Zusammenhang). Aufgrund der multiplikativen Verknüpfung geben die Koeffizienten gleichzeitig das Verhältnis an, in dem die Odds vor und nach einer Veränderung von x um eine Einheit zueinander stehen, sie können als „Odds-Ratios“ (OR) interpretiert werden. Die e^{β}-Koeffizienten bieten damit eine *scheinbar* einfache und anschauliche Interpretation, insbesondere bei Verwendung von Dummies als unabhängige Variablen. Beliebt sind hierbei Interpretationen in der Art „die Chancen [manchmal auch: die Odds], Abitur zu haben, sind bei Männern 1,5-mal höher als bei Frauen“. Zwar ist eine solche Interpretation *formal* korrekt, inhaltlich jedoch nur wenig sinnvoll, da die Gefahr besteht, dass diese Aussage falsch interpretiert wird. Odds sind Wahrscheinlichkeitsverhältnisse und Odds-Ratios entsprechend sogar *Verhältnisse von Wahrscheinlichkeitsverhältnissen*. Werden in diesem Beispiel Odds-Ratios als Verhältnis $P_{\text{Männer}}/P_{\text{Frauen}}$, also als sog. relatives Risiko, fehlinterpretiert (und das ist mithin die naheliegendste inhaltliche Interpretation), werden Effekte zwar in ihrer Richtung korrekt beurteilt, in ihrer Stärke aber generell überschätzt. Eine OR von 1,5 bedeutet eben nicht, dass Männer eine 1,5-mal höhere Wahrscheinlichkeit aufweisen, das Abitur zu erwerben als Frauen. Vielmehr ist – abhängig von der Basiswahrscheinlichkeit – *jedes* relative Risiko zwischen 1 und 1,5 möglich. *Wir raten daher von der Verwendung von e^{β}-Koeffizienten bzw. Odds-Ratios ab.* Eine über die Effekt-Richtung hinausgehende Interpretation ist allein auf der Basis des Koffizienten nicht möglich, wird aber häufig impliziert.

Soll der Zusammenhang zwischen zwei Variablen auf Basis einer logistischen Regression über Vorzeichen und Signifikanz hinaus interpretiert werden, ist es anzuraten, vorhergesagte Wahrscheinlichkeiten zu berechnen (siehe Gleichung (5)). Der nichtlineare Wahrscheinlichkeitsverlauf kann dann für ausgewählte Konstellationen der

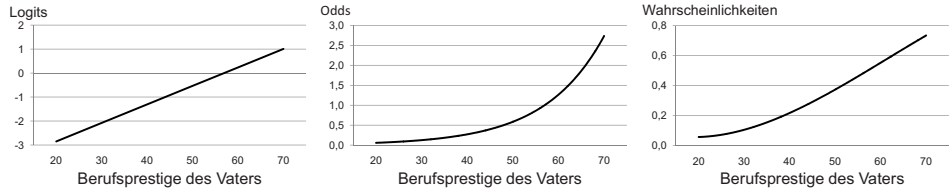


Abb. 3: Logits, Odds und Wahrscheinlichkeiten

unabhängigen Variablen z. B. anhand eines Conditional-Effect-Plot interpretiert werden (siehe hierzu Abschnitt 3, hilfreich ist auch Kapitel 34 in diesem Handbuch).

Wir möchten an dieser Stelle das Verhältnis von Logits, Odds und Wahrscheinlichkeiten an einem einfachen Beispiel veranschaulichen. Dafür gehen wir von dem bereits aus Abbildung 2 bekannten Modell aus, in dem der Erwerb des Abiturs in Abhängigkeit vom Berufsprestige des Vaters untersucht wird. Für dieses Modell wurde auf Basis des kumulierten ALLBUS 1980–2006 die Gleichung

$$\widehat{\text{Logit}}(\textit{Abitur}) = -4{,}3900 + 0{,}07708\ \textit{Prestige}$$

geschätzt. Am Vorzeichen des Steigungskoeffizienten erkennen wir, dass das Prestige des Vaters sich positiv auf die Wahrscheinlichkeit für den Erwerb des Abiturs beim Kind auswirkt. Es ist auch klar, dass die Logits für Abitur mit jeweils 10 Prestigeeinheiten um 0,77 Einheiten ansteigen. Was das für die Wahrscheinlichkeiten heißt, ist zunächst jedoch nicht klar. Dazu müssen die Logits mit Hilfe von Gleichung (5) in Wahrscheinlichkeiten umgerechnet werden. Exemplarisch ist dies in Tabelle 1 geschehen. Eine grafische Darstellung des Zusammenhangs zwischen dem Prestige des Vaters und den Logits, Odds bzw. Wahrscheinlichkeiten der Kinder, das Abitur zu erreichen, findet sich in Abbildung 3.

Demnach kommt man auf Basis des Modells für 20 Prestigepunkte zur Erwartung eines Logit von −2,85 oder einer Wahrscheinlichkeit von 0,05. Bei einem Berufsprestige von 40 Punkten wird ein Logit von −1,31 geschätzt, was einer Wahrscheinlichkeit von 0,21 entspricht. Und bei 70 Prestigepunkten schließlich wird ein Logit von 1 und entsprechend eine Wahrscheinlichkeit von 0,73 vorhergesagt. Die Tabelle enthält auch die Angaben zu den entsprechenden Odds, die gewissermaßen als Zwischenschritte berechnet wurden. Unter Kindern von Vätern mit einem Berufsprestige von 20 Punkten kommen auf 100 Kinder ohne Abitur lediglich sechs Kinder mit Abitur ($O = 0{,}06$).[5] Bei Kindern, deren Väter es auf 70 Prestigepunkte bringen, kommen auf 100 Kinder ohne Abitur bereits 273 Kinder mit Abitur ($O = 2{,}73$). Schließlich wird aus der Tabelle auch deutlich, dass sich die Logits linear verändern. Das bedeutet, dass bei Zunahme der unabhängigen Variablen um eine Einheit sich die Logits jeweils um einen konstanten Betrag verändern – hier +0,77 Einheiten pro 10 Prestigepunkte. Die Odds hingegen verändern sich jeweils um einen konstanten *Faktor*. Im Beispiel steigen die Odds alle 10 Prestigepunkte um mehr als das Doppelte, genauer um den Faktor

[5] Man beachte, dass es nicht heißt „*haben von* 100 Kindern sechs Kinder kein Abitur“. Diese Formulierung würde eine Wahrscheinlichkeit beschreiben.

$e^{10*0,077} = 2{,}16$. Die Veränderung der Wahrscheinlichkeiten schließlich folgt keiner einfachen Regel. Zunächst nimmt die Wahrscheinlichkeit beschleunigt zu, dann geht die Zunahme jedoch wieder zurück (+0,06; +0,10; +0,16; +0,19; +0,17).

2 Mathematisch-statistische Grundlagen

In diesem Abschnitt demonstrieren wir zunächst die formale Herleitung der logistischen Regression unter der Annahme latenter Variablen und zeigen Bezüge zu Probit-Modellen. Sodann wird kurz die (Maximum-Likelihood)-Schätzung eines Logitmodells vorgestellt und Eigenschaften der Logitkoeffizienten werden diskutiert (Standardisierung, Interaktionen). Zum Abschluss werden Gütemaße und statistische Inferenz diskutiert.

2.1 Herleitung als nichtlineares Modell mit latenter abhängiger Variable

Alternativ zu der oben geschilderten Herleitung über logarithmierte Odds kann die logistischen Regression auch über die Annahme einer nicht beobachteten, also latenten abhängigen Variable hergeleitet werden. Das Latente-Variable-Modell hat den Vorteil, einfache Bezüge zu anderen Verfahren zu ermöglichen (insbesondere der Probit-Regression) und die Grundannahmen der logistischen Regression systematischer aufzuzeigen.

Logit-Regression

Grundlegend für die logistische Regression ist die Annahme, dass eine latente, d. h. unbeobachtete, Variable y^* existiert, die dazu führt, dass Personen Zustände annehmen oder Entscheidungen treffen, deren Auftreten empirisch beobachtet werden kann (als dichotome Variable y).[6] Die beobachtete Variable y wird eins, wenn y^* einen bestimmten Schwellenwert τ überschreitet, der in der logistischen Regression arbiträr als $\tau = 0$ festgelegt wird. Die latente Variable y^* kann wiederum linear modelliert werden:[7]

$$y^* = \mathbf{x'}\boldsymbol{\beta} + \varepsilon\,. \tag{7}$$

Ist die Verteilung der Fehler bekannt, kann die Wahrscheinlichkeit von $y = 1$ berechnet werden (vgl. z. B. Long 1997, S. 44 f. oder Wooldridge 2002, S. 457). Es gilt, dass

$$P(y = 1|\mathbf{x}) = P(y^* > \tau) = P(y^* > 0)\,. \tag{8}$$

Setzt man entsprechend Gleichung (7) $y^* = \mathbf{x'}\boldsymbol{\beta} + \varepsilon$ ein, ergibt sich

$$P(y = 1|\mathbf{x}) = P(\mathbf{x'}\boldsymbol{\beta} + \varepsilon > 0)\,, \tag{9}$$

[6] Bei Entscheidungen könnte eine solche latente Variable beispielsweise der subjektiv erwartete Nutzen der Handlungsalternative sein.

[7] Aus Gründen der Übersichtlichkeit wird in diesem Abschnitt Vektornotation verwendet. $\mathbf{x'}\boldsymbol{\beta}$ ist hierbei gleichbedeutend mit $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

also nichts anderes als die formalisierte Annahme, dass die Wahrscheinlichkeit von $y = 1$ der Wahrscheinlichkeit entspricht, dass die Schätzwerte der Regressionsgleichung größer als der Schwellenwert $\tau = 0$ sind. Ein einfaches Umstellen der Ungleichung auf der rechten Seite führt zu

$$P(y = 1|\mathbf{x}) = P(\varepsilon > -\mathbf{x}'\boldsymbol{\beta})\,. \tag{10}$$

Die rechte Seite der Gleichung steht nun für die Wahrscheinlichkeit, dass die Fehler ε größer als ein bestimmter Wert sind (hier: $-\mathbf{x}'\boldsymbol{\beta}$). Oben wurde darauf verwiesen, dass die Verteilung der Fehler bekannt sein muss. Nimmt man an, dass die Fehler ε eine stetige, um den Nullpunkt symmetrische Verteilung aufweisen, gilt, dass $P(\varepsilon > -a) = P(\varepsilon \leq +a)$. D. h., die Fläche unter der Wahrscheinlichkeitsdichtefunktion rechts eines negativen Werts a ist identisch mit der Fläche links des entsprechenden positiven Werts (eine einfache Folge der Symmetrie). Damit kann auf der rechten Seite der Gleichung das Vorzeichen von $-\mathbf{x}'\boldsymbol{\beta}$ geändert werden, so dass

$$P(y = 1|\mathbf{x}) = P(\varepsilon \leq \mathbf{x}'\boldsymbol{\beta})\,. \tag{11}$$

Der rechte Teil der Gleichung, $P(\varepsilon \leq \mathbf{x}'\boldsymbol{\beta})$, beschreibt nun die Wahrscheinlichkeit, dass ε kleiner oder gleich ein bestimmter Wert ist. Exakt diese Wahrscheinlichkeit wird von einer kumulativen Verteilungsfunktion (CDF) beschrieben. Bezeichnet man diese Verteilungsfunktion als $G(\cdot)$ kann man die Gleichung als

$$P(\varepsilon \leq \mathbf{x}'\boldsymbol{\beta}) = G(\mathbf{x}'\boldsymbol{\beta}) \tag{12}$$

schreiben. Dies wiederum bedeutet, dass

$$P(y = 1|\mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta})\,. \tag{13}$$

Es muss also die genaue Form der Funktion $G(\cdot)$ bekannt sein, und diese ist – wie in Gleichung (12) beschrieben – nichts anderes als die Verteilung der Fehler ε. Im Gegensatz zum OLS-Modell können diese jedoch empirisch nicht beobachtet werden (bei unbeobachtetem y^* ist auch der Fehlerterm $\varepsilon = y^* - \widehat{y}^*$ nicht beobachtbar). Daher sind Annahmen über Verteilung, Standardabweichung und Erwartungswert der Schätzfehler notwendig. Im Rahmen der logistischen Regression wird angenommen, dass a) die Fehler (Residuen) einer logistischen Verteilung folgen, sie b) eine Standardabweichung von $\sigma_{\varepsilon|\mathbf{x}} = \pi/\sqrt{3}$ haben und c) ihr bedingter Erwartungswert $E(\varepsilon|\mathbf{x}) = 0$ ist. Aus diesen drei Annahmen folgt, dass

$$G(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}}\,. \tag{14}$$

Der etwas ungewöhnlich erscheinende Wert der Standardabweichung wurde hierbei so gewählt, dass Gleichung (14) eine möglichst einfache Form annimmt. Setzt man nun Gleichung (14) in Gleichung (13) ein, ergibt sich die schon aus Abschnitt 1.2 bekannte Basisgleichung der logistischen Regression:

$$P(y = 1|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}}\,. \tag{15}$$

Zur Identifikation der logistischen Regression sind damit eine Reihe von Annahmen und Festlegungen notwendig: Erstens über den Schwellenwert τ, bei dessen Überschreiten $y = 1$ beobachtet wird, zweitens über die Verteilung der Residuen, drittens die Residualvarianz und viertens ihren Erwartungswert. Werden die Annahmen wie oben beschrieben getroffen, führt dies zum Modell der logistischen Regression.

Probit-Regression

Trifft man andere Annahmen über die Verteilung der Residuen, ergibt sich gleichzeitig auch eine andere Modellierung der Wahrscheinlichkeit $P(y = 1)$. Wird angenommen, dass die Residuen standardnormalverteilt sind (d. h. $\sigma_{\varepsilon|\mathbf{x}} = 1$ und $E(\varepsilon|\mathbf{x}) = 0$) ergibt sich analog zu der oben dargestellten Herleitung das *Probit-Modell*. Aufgrund der Verteilungsannahme gilt hier, dass

$$G(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{z^2}{2}} \mathrm{dz}\,. \tag{16}$$

Da die Verteilungsfunktion der Standardnormalverteilung sich in ihrem Verlauf nur geringfügig von der CDF einer logistischen Verteilung mit $\sigma = 1$ unterscheidet, führen Logit- und Probit-Modelle zu nahezu identischen Schätzergebnissen in Bezug auf die Wahrscheinlichkeit von $y = 1$. Die Koeffizienten des linearen Logit- oder Probit-Modells sind jedoch skalenabhängig und werden daher von der Standardabweichung der jeweiligen Verteilung mit bestimmt. Die Koeffizienten können jedoch näherungsweise ineinander umgerechnet werden. Da die Standardabweichung der Residuen in der logistischen Regression mit $\sigma_\varepsilon = \pi/\sqrt{3}$ festgesetzt wurde, gilt, dass $\beta_{\text{Logit}} \approx \frac{\pi}{\sqrt{3}}\beta_{\text{Probit}} = 1{,}81\beta_{\text{Probit}}$. Long (1997, S. 48) bezieht die Unterschiede im funktionalen Verlauf der beiden Verteilungsfunktionen mit in die Berechnung ein und schätzt ein Verhältnis der Koeffizienten von ungefähr 1,7.

2.2 Schätzung

Wie in allen Regressionsverfahren wird auch in der logistischen Regression angestrebt, Werte für die Regressionskoeffizienten $\boldsymbol{\beta}$ zu finden, mit denen sich die beobachteten Daten möglichst gut reproduzieren lassen. Aufgrund von Eigenschaften der logistischen Regression (insbesondere Heteroskedastizität und Nicht-Linarität in Bezug auf die Wahrscheinlichkeiten) führt die Methode der kleinsten Quadrate (OLS) jedoch zu ineffizienten Schätzungen. Es wird daher auf eine *Maximum-Likelihood*-Schätzung zurückgegriffen (vgl. Kapitel 10 in diesem Handbuch für eine genauere Darstellung).

Die Anpassung der Regressionsgeraden an die empirischen Daten ist in der logistischen Regression dann besonders gut, wenn für Fälle, bei denen empirisch $y_i = 1$ beobachtet wird, eine möglichst hohe Wahrscheinlichkeit P(y=1|**x**) vorhergesagt wird (nahe eins). Bei Fällen, die empirisch $y_i = 0$ aufweisen, sollte die vorhergesagte Wahrscheinlichkeit $P(y = 1|\mathbf{x})$ hingegen möglichst gering sein (nahe null). Die letzte Bedingung ist äquivalent zu der Forderung, dass $1-P(y = 1|\mathbf{x})$ möglichst hoch sein soll. Da die vorhergesagte Wahrscheinlichkeit als $P(y = 1|\mathbf{x}) = {e^{\mathbf{x}'\boldsymbol{\beta}}}/{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$ berechnet wird,

können die beiden genannten Bedingungen für einen einzelnen Fall in der folgenden Gleichung kombiniert werden, so dass sich

$$f(y_i) = P(y_i = 1)^{y_i}(1 - P(y_i = 1))^{1-y_i} \tag{17}$$

bzw.

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = \left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i} \tag{18}$$

ergibt. Da y nur die Werte null und eins annehmen kann, ist aufgrund der Potenzierung mit y_i bzw. $1-y_i$ jeweils nur ein Teil der Formel interessant: bei $y_i = 1$ die erste Hälfte des rechten Terms, bei $y_i = 0$ die zweite Hälfte. Die Funktion kann Werte zwischen null und eins annehmen, wobei höhere Werte auf eine bessere Anpassung verweisen. Im Maximum-Likelihood-Verfahren wird angenommen, dass $\ell(\boldsymbol{\beta}|\mathbf{x_i};y_i) = f(y_i|\mathbf{x_i};\boldsymbol{\beta})$. Folglich berechnet sich die Likelihood eines einzelnen Falls analog zu Gleichung (18) als

$$\ell(\boldsymbol{\beta}|\mathbf{x}_i;y_i) = \left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{1-y_i}. \tag{19}$$

Selbstverständlich kann jedoch nicht nur ein einzelner Fall der Stichprobe betrachtet werden, sondern die Schätzung der Regressionskoeffizienten muss auf Basis aller Fälle erfolgen. Hierfür wird das Produkt der sog. „Likelihoods“ über alle Fälle betrachtet, so dass

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{y};\mathbf{X}) = \prod_{i=1}^{n}\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{\mathbf{y}}\prod_{i=1}^{n}\left(1-\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right)^{1-\mathbf{y}} \tag{20}$$

die zu maximierende Likelihood-Funktion darstellt. Auch diese Funktion folgt der oben geschilderten Logik. Um die Maximierung zu vereinfachen, wird üblicherweise nicht die Likelihood-Funktion als solches verwendet, sondern ihre logarithmierte Variante, die Log-Likelihood:

$$\ln\mathcal{L}(\boldsymbol{\beta}|\mathbf{y};\mathbf{X}) = \sum_{i=1}^{n}\mathbf{y}\ln\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right) + \sum_{i=1}^{n}(1-\mathbf{y})\ln\left(1-\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right). \tag{21}$$

Durch das Logarithmieren werden z. B. die Produkte in Summen transformiert, die Stelle, an der die Funktion ihr Maximum erreicht, jedoch nicht beeinflusst. Die eigentliche Maximierung erfolgt in einem iterativen Verfahren, bei dem für $\boldsymbol{\beta}$ generische Startwerte verwendet werden – beispielsweise die (ineffizienten) Ergebnisse einer OLS-Schätzung.

Es sollte unbedingt beachtet werden, dass für ML-Schätzungen größere Stichproben benötigt werden als für OLS-Schätzungen. Zwar gibt es keine belastbaren Mindestgrößen, Long (1997, S. 54) empfiehlt jedoch ein Minimum von 100 Fällen und 10 Fällen pro Parameter.

2.3 Vergleich von Koeffizienten und unbeobachtete Heterogenität

Die grundlegende Interpretation der Regressionskoeffizienten wurde bereits in Abschnitt 1.3 diskutiert. Dort wurde argumentiert, dass die β-Koeffizienten im logistischen

Regressionsmodell lediglich in Hinblick auf ihr Vorzeichen interpretiert werden sollten. Für eine detaillierte Interpretation wurde auf grafische Darstellungen vorhergesagter Wahrscheinlichkeiten verwiesen. Es kann allerdings durchaus von Interesse sein, zu prüfen, wie sich Regressionskoeffizienten beim schrittweisen Aufbau von Modellen verändern; in der OLS-Regression können hierdurch Schlussfolgerungen auf Drittvariableneffekte gezogen werden.

In der logistischen Regression wird diese Interpretation durch die Tatsache erschwert, dass erstens die Varianz von y^* unbekannt ist, und zweitens die Fehlervarianz als konstant angenommen wird ($\sigma_\varepsilon^2 = \pi^2/3$). Die unbeobachtete Varianz wird aus dem Regressionsmodell geschätzt und ist somit von der Erklärungskraft des Modells bzw. vom Ausmaß unbeobachteter Heterogenität abhängig. Die Varianz erhöht sich, wenn das Modell besser an die Daten angepasst ist. Hieraus folgt, dass sich die (geschätzte) Varianz der latenten abhängigen Variablen y^* verändert, wenn zusätzliche erklärende Variablen in das Regressionsmodell aufgenommen werden (oder sich, in anderen Worten, die unbeobachtete Heterogenität verringert). Folglich beziehen sich auch die β-Koeffizienten zweier genesteter Modelle[8] auf unterschiedlich skalierte abhängige Variablen und können in ihrer Größe nicht sinnvoll verglichen werden. Es ist unklar, ob eine eventuelle Veränderung der Koeffizienten auf Veränderungen von y^* oder auf Interkorrelationen der unabhängigen Variablen zurückgehen. In OLS-Regressionen wirkt sich unbeobachtete Heterogenität (also letztlich eine Fehlspezifikation des Modells) nur dann auf die Regressionskoeffizienten aus, wenn die unbeobachteten Variablen mit den im Modell enthaltenen Unabhängigen korreliert sind. Folglich ist das Problem in der logistischen Regression deutlich ausgeprägter, da unbeobachtete Heterogenität die Koeffizienten verzerren kann, selbst wenn die unabhängigen Variablen unkorreliert sind.

Um dennoch Koeffizienten zu erhalten, die zwischen zwei (genesteten) Modellen verglichen werden können, stehen zwei recht einfache Möglichkeiten zur Verfügung: Erstens können vollstandardisierte bzw. y^*-standardisierte Koeffizienten berechnet werden, bei denen die artifizielle Veränderung der Varianz von y^* durch Standardisierung weitgehend ausgeglichen wird. Zweitens ist es möglich, den durchschnittlichen marginalen Effekt (average marginal effect, *AME*) der unabhängigen Variablen zu berechnen. *AME* gibt einen durchschnittlichen Effekt auf die Wahrscheinlichkeiten an und ist nicht von (unkorrelierter) unbeobachteter Heterogenität betroffen.

Standardisierte Koeffizienten

Der vollstandardisierte Koeffizient β^s ist definiert als

$$\beta_j^s = \beta_j \frac{\sigma_{x_j}}{\sigma_{y^*}}\,, \tag{22}$$

wobei σ_{y^*} aus den Daten geschätzt werden muss. Dies ist jedoch einfach möglich, da

[8] Zwei Modelle sind genested (verschachtelt), wenn Modell 1 eine Untermenge von Modell 2 ist. Mit anderen Worten baut Modell 2 auf Modell 1 auf und erweitert es um zusätzliche Parameter.

$$\hat{\sigma}^2_{y^*} = \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\sigma}}^2_x \hat{\boldsymbol{\beta}} + \sigma^2_\varepsilon \,. \tag{23}$$

Bei Verwendung von nominalskalierten unabhängigen Variablen kann es sinnvoll sein, statt der Vollstandardisierung lediglich eine Teilstandardisierung an y^* vorzunehmen. ($\beta^{s_{y^*}} = {}^{\beta}/_{\sigma_{y^*}}$) Stellt das Statistikpaket keine Routine zur Berechnung der standardisierten Koeffizienten bzw. der latenten Varianz zur Verfügung, kann letztere leicht aus der Summe der Varianz der vorhergesagten Werte plus ${}^{\pi^2}/_3$ errechnet werden. Prinzipiell können – wie in der OLS-Regression – vollstandardisierte Koeffizienten als relative Einflussstärke der unabhängigen Variable interpretiert werden.[9] Hinzu kommt, dass der Koeffizient durch unbeobachtete Heterogenität nur geringfügig verzerrt wird, also besser zwischen Modellen verglichen werden kann als unstandardisierte Koeffizienten. Die Standardisierung gleicht die beschriebene Verzerrung jedoch nicht vollständig aus.[10]

Durchschnittliche marginale Effekte

Eine Alternative zu standardisierten Koeffizienten ist (im Hinblick auf den Vergleich zwischen Modellen) die Berechnung von Effekten auf die Wahrscheinlichkeit, so genannten marginalen Effekten. Der average marginal effect (*AME*) versucht hierbei, den durchschnittlichen Einfluss der unabhängigen Variable auf die Wahrscheinlichkeit des Auftretens $P(y = 1|\mathbf{x})$ in einer einzigen Kennziffer auszudrücken.

Die Darstellung von Effekten auf die Wahrscheinlichkeit $P(y = 1|\mathbf{x})$ in einer einzelnen Kennziffer ist in der logistischen Regression jedoch problematisch, da es sich um ein nichtlineares Modell handelt, in dem der Effekt einer Variablen – d. h. die Steigung der Wahrscheinlichkeitskurve – nicht konstant ist. Dies wird deutlich wenn man die Basisgleichung der logistischen Regression (Gleichung (15) auf Seite 835) partiell ableitet. Man erhält

$$\frac{\partial P(y = 1|\mathbf{x})}{\partial x_j} = g(\mathbf{x}'\boldsymbol{\beta})\beta_j \,, \tag{24}$$

wobei $g(\mathbf{x}'\boldsymbol{\beta})$ die Dichtefunktion der logistischen Verteilung ist. Somit ist der Effekt auf die Wahrscheinlichkeiten nicht nur abhängig vom Regressionskoeffizienten β_j, sondern zusätzlich von der Ausprägung aller Variablen und ihrem Effekt ($\mathbf{x}'\boldsymbol{\beta}$ in Gleichung (24)). Mit anderen Worten variiert der marginale Effekt von x_j erstens mit der Ausprägung von x_j selbst und zweitens mit den Ausprägungen der anderen unabhängigen Variablen.

Der *AME* gibt nun den Effekt von x_j auf einem durchschnittlichen Niveau an; der Durchschnitt kann jedoch auf zwei Arten spezifiziert werden: Entweder man berechnet

[9] Wie wir in Kapitel 24 zeigen, ist diese Interpretation jedoch in beiden Verfahren, OLS wie logistischer Regression, nicht unproblematisch. Bei der logistischen Regression muss zudem beachtet werden, dass sich der Einfluss auf die latente abhängige Variable bzw. die logarithmierten Odds bezieht.

[10] In Monte-Carlo-Simulationen der Autoren ergaben sich Veränderungen der standardisierten Koeffizienten um weniger als 20 %, verglichen einer Variation in den unstandardisierten Koeffizienten von 175 %.

den Durchschnittseffekt als Mittelwert der marginalen Effekte über alle Beobachtungen oder als marginalen Effekt am Mittelwert aller Variablen. Die zweite Variante wird auch als „marginal effect at the mean“ (*MEM*) bezeichnet und ist *nicht* identisch mit dem average marginal effect:

$$MEM_j = g(\bar{\mathbf{x}}'\boldsymbol{\beta})\beta_j\,, \tag{25}$$

während

$$AME_j = \frac{\sum_{i=1}^{N} g(\mathbf{x}_i'\boldsymbol{\beta})}{N}\beta_j\,. \tag{26}$$

Es lässt sich zeigen (z. B. Wooldridge 2002, S. 470 ff.), dass der durchschnittliche marginale Effekt (*AME*) nicht von unkorrelierter unbeobachteter Heterogenität verzerrt wird. Insofern ist der *AME* geeignet, um Koeffizienten schrittweise aufgebauter Modelle miteinander zu vergleichen. Es ist zu beachten, dass der *MEM* diese Eigenschaft nicht besitzt, vielmehr verändert er sich, wenn in ein Logitmodell weitere unkorrelierte Prädiktoren aufgenommen werden. Insofern ist *MEM* für den Vergleich zwischen Modellen nicht geeignet.[11]

Neben der Robustheit gegenüber unbeobachteter Heterogenität haben average marginal effects den Vorteil, eine intuitive Interpretation zu ermöglichen: eben als durchschnittlicher Effekt auf die Wahrscheinlichkeit. Insofern steigt die Wahrscheinlichkeit von $y = 1$ durchschnittlich um *AME* Punkte, wenn x_j um eine Einheit steigt. Selbstverständlich ist dies nur ein Durchschnittseffekt, der den nichtlinearen Verlauf der Wahrscheinlichkeitskurve ignoriert. Dennoch sind *AME* den in der Sozialwissenschaft zu Unrecht sehr beliebten Odds Ratios in mehrfacher Hinsicht überlegen (Robustheit, Interpretierbarkeit, Additivität).

2.4 Interaktionseffekte

Da die logistische Regression, wie in Abschnitt 1.3 ausführlich diskutiert, in Bezug auf die Wahrscheinlichkeiten nicht linear und nicht additiv ist, ergeben sich in der Anwendung und Interpretation von Interaktionseffekten eine Reihe von wichtigen Unterschieden zur OLS-Regression. So folgt aus der Nicht-Additivität, dass der Effekt einer unabhängigen Variable auf die Wahrscheinlichkeit $P(y = 1|\mathbf{x})$ vom Niveau der anderen Variablen abhängen kann. Genau dies ist jedoch die Aussage von Interaktionshypothesen: Der Effekt einer Variable x_1 auf y hängt vom Niveau einer zweiten Variable x_2 ab. Mit anderen Worten werden in der logistischen Regression bis zu einem gewissen Ausmaß implizit modellinhärente Interaktionseffekte (bzw. bedingte Effekte) auf die Wahrscheinlichkeit modelliert, selbst wenn sie nicht explizit spezifiziert werden. Diese modellinhärenten Interaktionseffekte führen prinzipiell zu einer geringeren Sensibilität der Logit-Modelle für explizit spezifizierte (variablenspezifische) Interaktionseffekte. Man sollte daher unbedingt visuell auf Basis der vorhergesagten Wahrscheinlichkeiten

[11] Wird auf Angaben zu marginalen Effekten zurückgegriffen, die von einem Statistikpaket automatisch berechnet werden, ist daher unbedingt zu prüfen, ob *AME* oder *MEM* ausgegeben werden (in Stata berechnen beispielsweise sowohl `mfx` als auch `prchange` den *MEM*; *AME* wird lediglich durch das ado `margeff` zur Verfügung gestellt).

prüfen, wie die Wahrscheinlichkeiten sich in Abhängigkeit von x_1 und x_2 verändern (siehe Abschnitt 3 oder ausführlicher Kapitel 34 in diesem Handbuch). Aufgrund der Nicht-Linearität der Logit-Modelle gilt das Gesagte ebenso für quadratische Terme und Polynome höherer Ordnung.

Sieht man von diesem Problem zunächst einmal ab, werden Interaktionseffekte (und auch explizite Polynome) genau wie in der OLS-Regression spezifiziert. Metrische Variablen werden zentriert, um die Multikollinearität zu verringern, und es wird ein multiplikativer Term x_1x_2 gebildet. Dieser Term wird in ein hierarchisch wohldefiniertes Modell aufgenommen (d. h. ein Modell, das neben dem multiplikativen Term auch die entsprechenden Haupteffekte enthält):

$$y^* = a + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon\,. \tag{27}$$

Einfaches Ausklammern von x_2 zeigt, dass der Effekt von x_2 nunmehr von x_1 abhängig ist: Er verändert sich um β_{12} Einheiten, wenn x_2 um eine Einheit steigt. β_{12} ist der Interaktionseffekt. β_2 hingegen ist ein konditionaler Effekt, der die Steigung der Geraden unter der Bedingung $x_1 = 0$ beschreibt:

$$y^* = a + \beta_1 x_1 + (\beta_2 + \beta_{12} x_1) x_2 + \varepsilon\,. \tag{28}$$

Selbstverständlich lässt sich die Umformung auch für x_1 vornehmen, die oben skizzierte Interpretation gilt entsprechend. Vergleicht man ein Modell mit und ohne multiplikativen Termen, lässt sich anhand der Veränderung verschiedener Anpassungsmaße beurteilen, ob sich das Modell verbessert hat (z. B. *AIC*, siehe Abschnitt 2.6). Ein Wald- bzw. Likelihood-Ratio-Test gibt Auskunft über die statistische Signifikanz (Abschnitt 2.5). Es ist jedoch wichtig, sich bewusst zu machen, dass die oben geschilderten Interpretationsmöglichkeiten und auch der Signifikanztest *ausschließlich in Bezug auf die latente abhängige Variable y^* bzw. die logarithmierten Odds gültig* sind. Aufgrund von Nicht-Additivität und Nicht-Linearität in der logistischen Regression kann eine Interaktion in Bezug auf die Wahrscheinlichkeiten unter bestimmten Voraussetzungen statistische Signifikanz verlieren und sogar ihr Vorzeichen wechseln. Sozialwissenschaftliche Hypothesen beziehen sich jedoch in aller Regel auf die Wahrscheinlichkeiten und nicht auf logarithmierte Odds oder latente Variablen (für eine formalisierte Darstellung des Problems siehe Ai & Norton 2003 oder Huang & Shields 2000).

Was tun? Wir schlagen vor, die Einwände ernst zu nehmen, aber pragmatisch damit umzugehen. Erstens sollte, wenn eine Interaktion vermutet wird, immer anhand eines LR-Tests und eines Vergleichs von *AIC* oder *BIC* geprüft werden, ob und inwieweit sich der Modellfit durch Aufnahme eines multipilkativen Terms verbessert (siehe Abschnitt 2.6). Ist dies nicht der Fall, sollte ein Conditional-Effect-Plot erzeugt werden (vorhergesagte Wahrscheinlichkeiten des Modells ohne multiplikativen Term), um zu prüfen, ob die Regressionskurven gleichartig verlaufen. Verbessert sich hingegen der Modellfit, ist von einer substanziellen Interaktion auszugehen. Hier sollte die Interaktion mit *mehreren* Conditional-Effect-Plots geprüft werden, bei dem die Kovariaten mindestens auf einem niedrigen, einem mittlerem und einem hohen Niveau konstant gehalten werden (z. B. $\bar{x}_i$ und $\bar{x}_i \pm 1$ Standardabweichung). Ein „niedriges Niveau“ ergibt sich hierbei, wenn bei Variablen mit positivem Effekt ein niedriger Wert eingesetzt wird

(z. B. $\bar{x}_i - \sigma$), bei Variablen mit negativem Effekt dagegen ein hoher Wert (z. B. $\bar{x}_i + \sigma$). Ceteris paribus sollte natürlich für das „hohe Niveau" verfahren werden. Hierdurch kann abgeschätzt werden, wie variabel bzw. wie stabil die Interaktion ist und welche Richtung sie annimmt.

2.5 Statistische Inferenz

Der Signifikanztest für Koeffizienten, die aus einer Stichprobe geschätzt wurden ($\hat{\beta}_j$), folgt dem aus der OLS-Regression bekannten Muster (siehe Kapitel 24 in diesem Handbuch). Die Stichprobenkoeffizienten können als Realisierungen einer Zufallsvariable aufgefasst werden, die asymptotisch normalverteilt ist mit $E(\hat{\beta}_j) = \beta_j$. Bezeichnet man den Standardfehler als $se(\hat{\beta}_j)$[12], ergibt sich die Prüfgröße z als

$$z_j = \frac{\hat{\beta}_j - \beta_{H_0}}{se(\hat{\beta}_j)} \,. \tag{29}$$

Der Test eignet sich zur Prüfung von Hypothesen der Form H_0: $\beta_j = \beta_{H_0}$ und folglich auch für einen klassischen Signifikanztest mit H_0: $\beta_j = 0$. Der z-Wert kann in diesem Fall einfach als Koeffizient dividiert durch Standardfehler berechnet werden.

Sollen komplexere Hypothesen überprüft werden, ist der oben dargestellte Signifikanztest einzelner Koeffizienten mitunter nicht ausreichend. Dies ist beispielsweise der Fall, wenn ein Konstrukt durch mehrere Variablen erfasst wird, etwa bei multinominalen unabhängigen Variablen (z. B. Familienstand). Eine andere Anwendung wäre ein Test auf Gleichheit zweier Regressionskoefizienten. Hierfür stehen in der logistischen Regression drei Verfahren zur Verfügung: *Likelihood-Ratio-Test*, *Wald-Test* und *Lagrange Multiplier Test* (vgl. Kapitel 10 in diesem Handbuch).

Beispielhaft wird an dieser Stelle der Likelihood-Ratio-Test (LR-Test) dargestellt. Dieser Test ist geeignet, um zwei genestete Modelle miteinander zu vergleichen, also Modelle, in denen die geschätzten Parameter des einen Modells eine echte Teilmenge der Parameter des anderen Modells sind. Der LR-Test beantwortet die Frage, ob die Hinzunahme bestimmter Parameter in ein Modell zu einer statistisch signifikanten Verbesserung des Modells beiträgt. Konkret prüft der Test, ob die Likelihood des unrestringierten Modells mit mehr Parametern deutlich über der Likelihood des restriktiven Modells liegt (das restriktivste Modell ist ein Nullmodell). Der LR-Test zum Vergleich zweier Modelle ist als

$$LR = -2 \ln \frac{\mathcal{L}_\mathrm{R}}{\mathcal{L}_\mathrm{U}} = -2(\ln \mathcal{L}_\mathrm{R} - \ln \mathcal{L}_\mathrm{U}) \tag{30}$$

definiert. Dabei steht $\mathcal{L}_\mathrm{U}$ für den Wert der Likelihood-Funktion des Modells ohne Restriktionen, also mehr Parametern. Entsprechend bezeichnet $\mathcal{L}_\mathrm{R}$ den Wert der Likelihood-Funktion des restringierten Modells ohne bzw. mit weniger Parametern.

[12] Die Formel zur Schätzung der asymptotischen Standardfehler ist recht kompliziert und eine Darstellung erscheint an dieser Stelle nur wenig hilfreich. Der interessierte Leser sei auf Wooldridge (2002, S. 460 f.) verwiesen.

Unter der Nullhypothese, dass sich die beiden Modelle nicht unterscheiden, folgt LR einer χ^2-Verteilung mit $df_{\mathrm{LR}} = df_{\mathrm{U}} - df_{\mathrm{R}}$. Die Freiheitsgrade entsprechen also der Anzahl der Parameter, die das weniger restriktive Modell gegenüber dem stärker restriktiven Modell mehr enthält. Der Test ist, wie bereits erwähnt, nur gültig, wenn die verglichenen Modelle genestet sind und wenn die Analyse auf Basis derselben Stichprobe bzw. denselben Einheiten durchgeführt wird.

2.6 Goodness of fit und Modellvergleich

Zwar kann man mit den oben diskutierten Tests überprüfen, ob eine Modellverbesserung statistisch signifikant ist (also auch für die Grundgesamtheit als gültig angenommen werden kann). In den meisten Fällen ist jedoch auch eine Maßzahl zur Beschreibung der Anpassungsgüte des Modells erwünscht. In OLS-Regressionen dient hierzu R^2, das als das Verhältnis der erklärten Varianz zur Gesamtvarianz der abhängigen Variable definiert ist. In der logistischen Regression ist die abhängige Variable latent und lediglich über einen Schwellenwert mit der beobachteten dichotomen Variable verknüpft, so dass kein einfaches Maß der erklärten Varianz existiert. Basierend auf der (Log-)Likelihood der Modelle wurden jedoch eine Reihe von Maßzahlen vorgeschlagen, die eine Interpretation analog zu R^2 anstreben. Diese so genannten *Pseudo-*R^2-Koeffizienten variieren idealerweise zwischen Null und Eins, wobei ein Wert von Null ein Regressionsmodell beschreibt, das keine Erklärungskraft besitzt.

Wie bereits in Abschnitt 2.2 beschrieben, ist die Likelihood-Funktion eines Regressionsmodells die Basis der ML-Schätzung (siehe Gleichung (18)). Die logarithmierte Likelihood (LL oder $\ln \mathcal{L}$, siehe Gleichung (21)) ist immer negativ und umso kleiner im Betrag, je besser das Modell an die Daten angepasst ist. Insofern kann bereits die von vielen Statistikprogrammen angegebene $-2LL$ als ein Indikator für den Modellfit angesehen werden (je kleiner, desto besser), der allerdings – wie auch der F-Wert in der OLS-Regression – fallzahlabhängig ist. McFadden (1973) schlägt vor, die Log-Likelihood eines Nullmodells (ohne erklärende Variablen) analog zur Gesamtstreuung in R^2 und die Log-Likelihood des spezifizierten Modells analog zur erklärten Streuung zu interpretieren. Entsprechend berechnet sich McFaddens Pseudo-R^2 als

$$R^2_{\mathrm{MF}} = 1 - \frac{\ln \mathcal{L}(M_{\mathrm{spez}})}{\ln \mathcal{L}(M_0)}\,, \tag{31}$$

wobei M_0 das Basis-Modell ohne erklärende Variablen bezeichnet, M_{spez} das spezifizierte Modell. R^2_{MF} ist recht weit verbreitet (und wird z. B. in Stata standardmäßig berechnet), hat allerdings den Nachteil, nie den Wert 1 erreichen zu können. Eine Alternative stellt das von Cox & Snell (1989) vorgeschlagene Pseudo-R^2 dar, das anhand der Fallzahl N eine Korrektur vornimmt:

$$R^2_{\mathrm{CS}} = 1 - \left(\frac{\mathcal{L}(M_0)}{\mathcal{L}(M_{\mathrm{spez}})} \right)^{\frac{2}{N}}\,. \tag{32}$$

Allerdings kann auch R^2_{CS} den Wert 1 nicht erreichen. Nach einem Vorschlag von Cragg & Uhler (1970) wird die Maßzahl so normiert, dass sie auch den Wert 1 erreichen kann. In den meisten Anwendungen wird die hieraus resultierende Pseudo-R^2-Variante

$$R^2_{\text{NK}} = \frac{R^2_{CS}}{\max(R^2_{\text{CS}})} = \frac{R^2_{CS}}{1 - \mathcal{L}(M_0)^{\frac{2}{N}}} \tag{33}$$

als Nagelkerke-R^2 bezeichnet. Das Nagelkerke-R^2 wird beispielsweise von SPSS ausgegeben. R^2_{NK} liefert immer größere Werte als andere Pseudo-R^2-Varianten. Generell sollte Pseudo-R^2 mit Vorsicht interpretiert werden, da sich die Likelihood-basierten Maßzahlen erstens nicht auf die erklärte Varianz beziehen (wie in der OLS-Regression) und sich zweitens kein verbindliches Standardmaß herauskristallisiert hat.

Die diskutierten Anpassungsmaße leiden zudem unter dem Problem, dass sie prinzipiell größere Werte annehmen, je mehr erklärende Variablen im Modell enthalten sind. Letztlich ist es jedoch wünschenswert, ein Modell zu verwenden, das nicht nur möglichst gut an die Daten angepasst, sondern gleichzeitig auch möglichst sparsam spezifiziert ist. In OLS-Regressionen wird teilweise das korrigierte R^2 verwendet, um dieser Tatsache gerecht zu werden. Für logistische Modelle (und andere ML-basierte Verfahren) steht das „Akaike Informationskriterium“ (*AIC*) und das „Bayessche Informationskriterium“ (*BIC*) zur Verfügung. *AIC* und *BIC* erlauben es auch, Modelle miteinander zu vergleichen, die nicht genestet sind.

AIC (siehe Akaike 1973) ist eine Likelihood-basierte Maßzahl, bei der eine zusätzliche Parametrisierung des Modells bestraft wird. Der Koeffizient wird berechnet als

$$AIC = -2 \ln \mathcal{L}(M_{\text{spez}}) + 2(k+1)\,, \tag{34}$$

wobei k die Zahl der unabhängigen Variablen bezeichnet. Der *AIC* kann theoretisch von 0 bis $+\infty$ variieren, wobei niedrigere Werte auf ein geeigneteres Modell hindeuten.

Auch *BIC* (Raftery 1995) wird über die Likelihood der Modelle berechnet, baut logisch jedoch auf einem bayesianischen Modellvergleich auf. Bei dieser Maßzahl wird eine zusätzliche Parametrisierung tendenziell stärker bestraft als bei *AIC*. *BIC* kann beispielsweise als

$$BIC = -2 \ln \mathcal{L}(M_{\text{spez}}) + \ln N(k+1) \tag{35}$$

berechnet werden. Auch hier bezeichnet N die Fallzahl, k die Zahl der unabhängigen Variablen und M_{spez} das spezifizierte Modell. Kleinere *BIC*-Werte deuten auf einen auf einen besseren Fit hin.

3 Ein Beispiel

Im Folgenden wollen wir die Verwendung der logistischen Regression am Beispiel einer Analyse der *Bildungsvererbung* in Westdeutschland darstellen. Hierfür wird untersucht, welche Herkunftsfaktoren die Wahrscheinlichkeit bestimmen, dass eine Person das (Fach-)Abitur erwirbt. Als Datengrundlage dient der kumulierte ALLBUS (1980–2006), betrachtet werden Personen im Alter von mindestens 21 Jahren mit Wohnort in Westdeutschland. Als abhängige Variable wird eine dichotome Variable mit 1=„Abitur“ und 0=„kein Abitur“ verwendet, so dass eine logistische Regression das angemessene Analyseverfahren ist.

Die bisherige Forschung zeigt, dass es in Deutschland starke Effekte der sozialen Herkunft auf die Bildungschancen gibt (siehe z. B. Becker & Lauterbach 2008 für

eine Übersicht). Zu beachten ist außerdem, dass die Bildungschancen mit dem Geschlecht variieren, der Geschlechtseffekt sich in den letzten Jahren und Jahrzehnten allerdings deutlich abgeschwächt hat. Schließlich ist zu beachten, dass sich im Zuge gesellschaftlicher Modernisierungsprozesse, insbesondere der „Bildungsexpansion", die Bildungschancen generell verbessert haben. Es ist außerdem zu prüfen, inwiefern die Bildungsexpansion Herkunftseffekte verringert hat. Als unabhängige Variablen verwenden wir daher das Alter, Geburtsjahr und Geschlecht der Befragten sowie den Schulabschluss beider Eltern und das Berufsprestige des Vaters (Treiman-Prestige). Die gleichzeitige Verwendung von Alter und Geburtsjahr ist möglich, da im kumulierten ALLBUS wiederholte Befragungen über mehr als zwei Jahrzehnte vorliegen. Damit ist es möglich, Kohorten- von Alterseffekten zu trennen (wenn auch vermengt mit Periodeneffekten).

Zur Analyse der Determinanten des Abiturerwerbs schätzen wir eine Reihe von logistischen Regressionen (siehe Tabelle 2). Für alle Modelle sind standardisierte und unstandardisierte Logit-Koeffizienten sowie der Standardfehler angegeben, Modell 1 weist beispielhaft zusätzlich durchschnittliche marginale Effekte aus (AME).

Modell 1 enthält lediglich Angaben zum Befragten. Das Pseudo-R^2 der Modells liegt bei 0,06 (McFadden) bzw. 0,09 (Nagelkerke). Die Variablen „Geburtsjahr" und „Alter" prüfen Alters- und Kohorteneffekte auf die Bildungschancen. Der signifikant positive Koeffizient des Geburtsjahres zeigt an, dass die Wahrscheinlichkeit der Hochschulreife in späteren Geburtskohorten höher ist als in früheren. Mit einem standardisierten Koeffizienten von 0,26 ist der Effekt moderat stark. Am durchschnittlichen marginalen Effekt kann man erkennen, dass sich die Wahrscheinlichkeit eines durchschnittlichen Befragten, das Abitur zu erwerben, im Mittel pro Dekade um 5 Prozentpunkte erhöht. Der Alterseffekt ist negativ und nur auf dem 5 %-Niveau signifikant, was bei der hier verwendeten Stichprobe von über 20. 000 nicht viel bedeutet. Mit einem um zehn Jahre höheren Alter verringert sich die Wahrscheinlichkeit des Abiturerwerbs im Durchschnitt um einen Prozentpunkt (zu erkennen am AME). Der sehr niedrige standardisierte Koeffizient zeigt, dass der Effekt substanziell nicht bedeutsam ist (ein Alterseffekt wäre theoretisch auch nicht zu erwarten gewesen).

Wir möchten nun den Einfluss des Geschlechts auf die Bildungschancen genauer diskutieren. Man sieht, dass der Dummy „Mann" einen positiven β-Koeffizienten aufweist, der statistisch signifikant ist. Dies bedeutet, dass Männer im Mittel mit einer höheren Wahrscheinlichkeit das Abitur erwerben als Frauen. Die absolute Größe ($\beta_j = 0{,}44$) hat jedoch keine inhaltliche Bedeutung, sondern bezieht sich auf die Veränderung der latenten abhängigen Variable y^*, deren Skala nicht bekannt ist.[13] Eine

[13] Es ist in dieser Situation zwar verführerisch, aber kontraproduktiv, einen entlogarithmierten „Effekt"-Koeffizienten anzugeben und zu interpretieren, die „Chancen (Odds) eines Mannes, das Abitur zu erwerben, seien 1,55-mal höher als die einer Frau". Odds – und insbesondere Odds-Ratios – sind weniger intuitiv als es erscheint und bieten nach unserer Ansicht keinerlei Vorteile gegenüber den Logits. Ihre Verwendung erhöht jedoch das Risiko, dass der Koeffizient vom Leser als Effekt auf die Wahrscheinlichkeiten fehlinterpretiert wird. Die Wahrscheinlichkeit von Männern, Abitur zu erwerben, steht jedoch gerade *nicht* im Verhältnis 1,55 zu der von Frauen. Vielmehr lässt sich aus vorhergesagten Wahrscheinlichkeiten des Modells 1 berechnen, dass Männer, die 1950 geboren wurden, eine 1,40-mal

intuitive Interpretation erlaubt jedoch der durchschnittliche marginale Effekt, der mit einem Wert von 0,08 anzeigt, dass Männer im Durchschnitt unserer Stichprobe eine um 8 Prozentpunkte höhere Wahrscheinlichkeit des Abiturerwerbs haben als Frauen. Es ist jedoch zu beachten, das Logit-Modelle in Bezug auf die Wahrscheinlichkeiten nichtlinear sind und der Effekt einer Variable von den Ausprägungen der anderen Unabhängigen abhängig ist. Der *AME* umgeht diese Eigenschaft logistischer Modelle, indem er einen Durchschnittseffekt angibt. Hiermit ist selbstverständlich ein Informationsverlust verbunden.

Detaillierte Interpretationen der Ergebnisse machen es erforderlich, vorhergesagte Wahrscheinlichkeiten zu berechnen und den Wahrscheinlichkeitsverlauf in *Conditional-Effect-Plots* darzustellen (zu graphischen Darstellungen von Regressionsergebnissen vgl. auch Kapitel 34 in diesem Handbuch). Abbildung 4 a zeigt daher den in Modell 1 berechneten Effekt der Geburtskohorte auf die Wahrscheinlichkeit, das Abitur zu erwerben getrennt für Männer und Frauen. Man kann erkennen, dass erstens der Verlauf der Bildungsexpansion nicht linear war, sondern sich die Bildungschancen für spätere Kohorten stärker verbessert haben als für frühere.[14] Die Wahrscheinlichkeitskurve der Männer liegt über der der Frauen, was auf die erwarteten Diskriminierungsprozesse verweist. Man sieht zudem, dass die Kurven für Männer und Frauen nicht den gleichen Verlauf nehmen, obwohl in Modell 1 keine explizite Interaktion der Variablen spezifiziert wurde. Es ist jedoch sehr überraschend, dass die Wahrscheinlichkeit bei Männern stärker ansteigt als bei Frauen, Modell 1 also anzeigt, dass sich die Diskriminierung von Frauen verstärkt hätte – ein Ergebnis in klarem Widerspruch zu allen bisherigen Befunden. Ausgehend von publizierten Forschungsergebnissen ist vielmehr anzunehmen, dass Frauen überproportional von der Bildungsexpansion profitiert haben, die früher sehr deutliche Bildungsdiskriminierung sich also abgeschwächt hat bzw. mittlerweile überwunden wurde.

Modell 2 beinhaltet daher einen multiplikativen Term zwischen den beiden Variablen. Die Interaktion ist negativ und hochsignifikant, was darauf hindeutet, dass die geringen Bildungschancen von Frauen – wie erwartet – sich im Zeitverlauf verbessert haben. Zwar hat sich das Pseudo-R^2 des Modells nicht (bzw. nur auf der dritten Nachkommastelle) verändert, der im Vergleich zu Modell 1 niedrige *AIC* zeigt jedoch, dass das Modell besser an die Daten angepasst ist und die verbesserte Anpassung eine komplexere Parametrisierung rechtfertigt. Die vorhergesagten Wahrscheinlichkeiten sind in Abbildung 4 b dargestellt. Man sieht, dass die explizite Aufnahme des Interaktionseffektes wichtig war und sich die Wahrscheinlichkeitskurven deutlich von der Prognose aus Modell 1 unterscheiden. Modell 2 zeigt, dass zwar beide Geschlechter von der Bildungsexpansion profitiert haben, die Bildungschancen von Frauen jedoch viel stärker als die von Männern gestiegen sind. Waren in früheren Generationen die

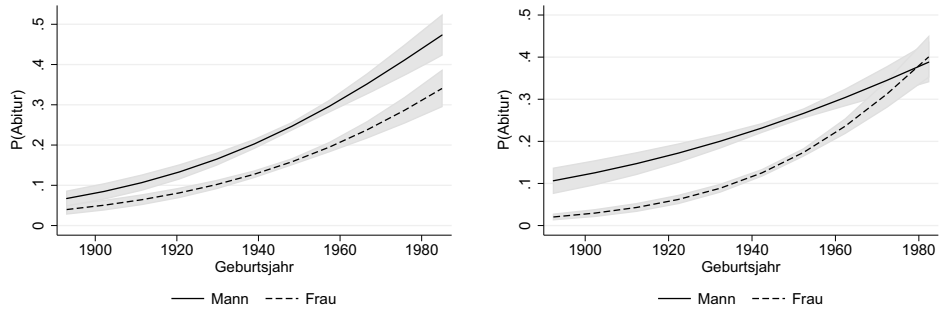
höhere Wahrscheinlichkeit aufweisen, und bei Geburtsjahr 1970 wird ein Verhältnis von 1,33 geschätzt. Der *AME* zeigt, dass diese Wahrscheinlichkeitsverhältnisse im Durchschnitt einem Unterschied von 8 Prozentpunkten entsprechen.

[14] Nochmal: Diese Nicht-Linearität ist nicht in einer einzigen Zahl darzustellen und zeigt, dass sorgfältige Interpretationen ohne die Betrachtung der vorhergesagten Wahrscheinlichkeiten nicht möglich sind. Da die Kurven jedoch streng monoton sind, erlauben die Koeffizienten zumindest eine Aussage über die Richtung der Effekte.

Tab. 2: Ergebnisse der logistischen Regression zum Abiturerwerb

	Modell 1			Modell 2		Modell 3	
	β (se)	β^s	AME	β (se)	β^s	β (se)	β^s
Befragter							
Mann	0,44	$0{,}11^{***}$	0,08	0,54	$0{,}14^{***}$	0,67	$0{,}15^{***}$
	(0,03)			(0,03)		(0,04)	
Geburtsjahr (in 10 J.)	0,28	$0{,}26^{***}$	0,05	0,37	$0{,}38^{***}$	0,39	$0{,}30^{***}$
	(0,02)			(0,03)		(0,03)	
Alter (in 10 J.)	−0,06	$-0{,}05^{*}$	−0,01	−0,06	$-0{,}05^{*}$	−0,01	−0,01
	(0,02)			(0,02)		(0,03)	
Mann*Geburtsjahr				−0,15	$-0{,}09^{***}$	−0,20	$-0{,}11^{***}$
				(0,02)		(0,02)	
Vater							
Treiman-Prestige						0,04	$0{,}20^{***}$
						(0,00)	
Kein/anderer Abschl.						−0,98	$-0{,}08^{***}$
						(0,13)	
Hauptschule						−1,32	$-0{,}26^{***}$
						(0,06)	
Mittlere Reife						−0,42	$-0{,}06^{***}$
						(0,07)	
Mutter							
Kein/anderer Abschl.						−1,65	$-0{,}17^{***}$
						(0,14)	
Hauptschule						−1,24	$-0{,}23^{***}$
						(0,09)	
Mittlere Reife						−0,42	$-0{,}06^{***}$
						(0,09)	
Konstante	−1,48			−1,54		0,36	
	(0,02)			(0,03)		(0,09)	
Pseudo-R^2 (NK)	0,09			0,09		0,35	
Pseudo-R^2 (MF)	0,06			0,06		0,24	
$-2LL$	24746,34			24687,50		19927,99	
AIC	1,05			1,04		0,84	
N	23641			23641		23641	

Referenzkategorien: Frau, Vater Abitur, Mutter Abitur
Alter, Prestige und Geburtsjahr zentriert
†: p ≤ 0,1; *: p ≤ 0,05; **: p ≤ 0,01; ***: p ≤ 0,001



(a) ohne multiplikativen Term (Mod. 1) (b) mit multiplikativem Term (Mod. 2)

Abb. 4: Vorhergesagte Wahrscheinlichkeit, Abitur zu erwerben (nach Geburtskohorte und Geschlecht)

geschlechtstypischen Bildungschancen noch sehr ungleich, weisen Männer und Frauen, die seit Mitte der 1970er Jahre geboren wurden, nun im Wesentlichen die gleiche Wahrscheinlichkeit auf, das Abitur zu erwerben. Dieses Beispiel macht deutlich, dass es auch in der logistischen Regression notwendig sein kann, Interaktionsterme zu verwenden, um Effekte korrekt zu modellieren und eine Fehlspezifikation zu vermeiden. Andererseits bedeutet ein nicht-signifikanter Interaktionseffekt in nichtlinearen Modellen nicht notwendig, dass die Effekte auf die Wahrscheinlichkeit unabhängig von der Ausprägung der anderen unabhängigen Variablen sind. Im Gegenteil hängen die Effekte *immer* von anderen Variablen ab, wie man durch eine partielle Ableitung der Regressionsgleichung nachprüfen kann (siehe Gleichung (24) auf Seite 839). Daher sollte der Verlauf der Kurven mit Conditional-Effect-Plots analysiert werden.[15]

Um Herkunftseffekte zu untersuchen, also zu prüfen, wie und in welchem Ausmaß die Bildungschancen in Deutschland von der sozialen Lage der Eltern beeinflusst sind, werden in Modell 3 Variablen für das Berufsprestige des Vaters und den Schulabschluss beider Elternteile aufgenommen.[16] Die Modellanpassung verbessert sich im Vergleich zu Modell 2 deutlich, und das Pseudo-R^2 steigt auf 0,24 (McFadden) bzw. 0,35 (Nagelkerke). Da die Schulbildung mit Dummy-Variablen erfasst wurde,

[15] Die Berechnung von durchschnittlichen marginalen Effekten ist in Modellen, die Interaktionseffekte beinhalten, nur bedingt sinnvoll, da der Interaktionseffekt auf die Wahrscheinlichkeiten nicht nur durch den Koeffizienten des multiplikativen Terms bestimmt wird (siehe hierzu auch die Ausführungen in Abschnitt 2.4 sowie Ai & Norton 2003).

[16] In einer stärker inhaltlich motivierten Analyse wäre es sinnvoll, die Bildungsabschlüsse von Vater und Mutter getrennt voneinander in das Modell aufzunehmen, um ihre relative Stärke beurteilen zu können. Aus Platzgründen können die entsprechenden Regressionsmodelle an dieser Stelle nicht dargestellt werden. Entsprechende Berechnungen zeigen, dass die Effekte ähnlich stark sind: Unter Kontrolle des Berufsprestiges des Vates zeigen sich keine Unterschiede, ohne Kontrolle ist die Bildung des Vaters etwas wichtiger als die der Mutter. Die Schulbildung wiederum hat eine höhere Erklärungskraft als das Berufprestige.

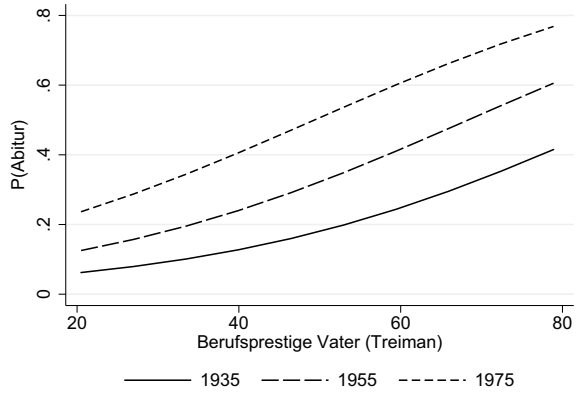


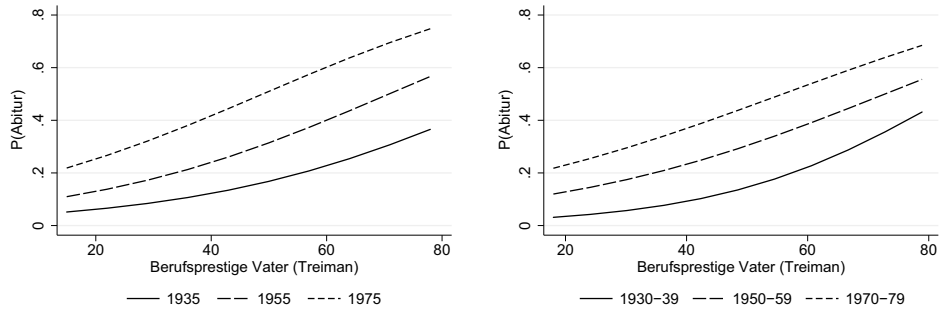
Abb. 5: Prestige nach Geburtskohorte I

beziehen sich die Koeffizenten (und ihre Signifikanz) auf den Unterschied zur jeweiligen Referenzkategorie. Der negative Koeffizient für Mittlere Reife des Vaters bedeutet, dass die Wahrscheinlichkeit, Abitur zu erwerben, bei Kindern von Vätern mit Mittlerer Reife geringer ist, als wenn der Vater selbst Abitur hat. Gleiches gilt für einen Hauptschulabschluss des Vaters; der betragsmäßig größere Koeffizient zeigt jedoch, dass der Unterschied größer ist.[17] Die Effekte der Bildung der Mutter sind – wie die des Vaters – alle signifikant und weisen ebenso ein negatives Vorzeichen auf. Somit sind die Bildungschancen der Kinder geringer, wenn die Mutter eine niedrige Schulbildung hat (auf eine detailliertere Interpretation wird an dieser Stelle aus Platzgründen verzichtet). Auch unter Kontrolle der elterlichen Schulbildung ist noch ein deutlicher Einfluss des Berufsprestiges des Vaters zu erkennen. Der Koeffizient ist statistisch hochsignifikant und positiv, so dass die Bildungschancen der Kinder mit dem Prestige des Vaters ansteigen. Der standardisierte Koeffizient von 0,20 verweist auf einen substanziellen Zusammenhang.

Wir wollen den Effekt der Herkunft im Folgenden etwas detaillierter betrachten und insbesondere untersuchen, wie sich der Einfluss des Elternhauses auf die Schulbildung der Kinder im Laufe der Zeit gewandelt hat. Hierfür betrachten wir zunächst vorhergesagte Wahrscheinlichkeiten des Abiturerwerbs in Abhängigkeit vom Berufsprestige des Vaters getrennt für drei Geburtskohorten (siehe Abbildung 5).

Zunächst sieht man auch in dieser Abbildung an den vertikal verschobenen Kurven, dass sich die Bildungschancen in Deutschland mit der Zeit verbessert haben. Zudem ist deutlich zu erkennen, wie die Wahrscheinlichkeit, Abitur zu erwerben, mit dem Berufsprestige des Vaters ansteigt. Der Kurvenverlauf ist jedoch für die drei Geburtskohorten unterschiedlich. In der ältesten Kohorte erhöht sich die Steigung der Kurve

[17] Diese Codierung erlaubt jedoch keine Aussage darüber, ob der Unterschied zwischen Mittlerer Reife des Vaters und Hauptschulabschluss des Vaters statistisch signifikant ist. Soll eine solche Aussage getroffen werden, muss Hauptschule oder Mittlere Reife als Referenzkategorie gewählt werden.



(a) mit multiplikativem Term (b) getrennte Modelle

Abb. 6: Prestige nach Geburtskohorte II

mit zunehmendem Prestige; die Kurve steigt zunächst flach an und wird im weiteren Verlauf steiler. In der jüngsten Kohorte hingegen ist die Steigung zunächst stärker, die Kurve nähert sich jedoch einem linearen Verlauf an. Um die Unterschiede zwischen den Kohorten genauer beurteilen zu können, ist es sinnvoll, entweder Interaktionsterme in das Modell aufzunehmen oder getrennte Modelle zu schätzen. Getrennte Modelle haben den Vorteil, übersichtlicher zu sein, geben allerdings keine Auskunft über die statistische Signifikanz eventueller Unterschiede. Tabelle 3 stellt daher getrennte Modelle für die Geburtskohorten 1930–1939, 1950–1969 und 1970–1979 sowie ein Gesamtmodell mit Interaktionseffekten dar.[18]

Beim Vergleich der Modelle für die drei Geburtskohorten ist es besonders interessant, sich die Anpassungsgüte anzuschauen (Pseudo-R^2).[19] Man sieht, dass die Erklärungskraft des Modells für die Kohorte der in den 1930ern Geborenen sehr hoch ist, und in den beiden jüngeren Kohorten um etwa 10 Punkte darunter liegt. Zwei wichtige Punkte sind jedoch zu bemerken: Erstens kam es zwischen den beiden jüngeren Kohorten nicht zu einer weiteren Abschwächung des Herkunftseffektes, und zweitens ist auch in der jüngsten Geburtskohorte die Erklärungskraft des Modells mit einem Pseudo-R^2 von 0,18 (MF) bzw. 0,28 (NK) besorgniserregend hoch. Das deutsche Schulsystem scheint seit den 1960er-Jahren nicht viel durchlässiger geworden zu sein.[20] Das Ge-

[18] Da es sich nicht um genestete Modelle handelt, kann der LR-Test hier nicht zum Vergleich der Modelle herangezogen werden. Stattdessen berichten wir AIC, dessen Interpretation hier zu denselben Schlussfolgerungen wie die Interpretation der Pseudo-R^2 führt.

[19] Wie in Abschnitt 2.3 ausgeführt, können die Regressionskoeffizienten über Modelle hinweg nicht sinnvoll verglichen werden.

[20] Um zu prüfen, in wie weit die Veränderungen des R^2 am Geschlechts- oder Herkunftseffekt liegen, haben wir zusätzlich Modelle geschätzt, in die keine Befragten-Variablen (Geschlecht, Alter) eingehen. Die Modelle haben Anpassungsgüten nach Mc-Fadden von 0,26 (30er), 0,16 (50er) und 0,18 (70er). Folglich ist der Einfluss des Elternhauses auf die Bildungschancen zwar zunächst gesunken (Vergleich der Kohorte 1930er/1950er), daraufhin jedoch wieder leicht angestiegen (Vergleich der Kohorten 1950er/1970er).

Tab. 3: Getrennte Modelle nach Geburtskohorten

	1930–39 β (se)	1950–59 β (se)	1970–79 β (se)	Gesamt β (se)
Befragter				
Mann	1,01***	0,66***	−0,08	0,69***
	(0,13)	(0,07)	(0,11)	(0,04)
Alter (in 10 J.)	0,04	−0,11*	−0,11	−0,02
	(0,08)	(0,05)	(0,14)	(0,03)
Vater				
Treiman-Prestige	0,05***	0,03***	0,03***	0,04***
	(0,01)	(0,00)	(0,01)	(0,00)
Kein/anderer Abschl.	−0,75†	−1,09***	−1,15**	−0,98***
	(0,41)	(0,28)	(0,39)	(0,13)
Hauptschule	−1,80***	−1,22***	−1,10***	−1,30***
	(0,21)	(0,14)	(0,20)	(0,06)
Mittlere Reife	−0,25	−0,25†	−0,65**	−0,41***
	(0,21)	(0,15)	(0,21)	(0,07)
Mutter				
Kein/anderer Abschl.	−1,87***	−2,09***	−1,84***	−1,67***
	(0,45)	(0,29)	(0,39)	(0,14)
Hauptschule	−1,18***	−1,81***	−1,42***	−1,25***
	(0,29)	(0,21)	(0,25)	(0,09)
Mittlere Reife	−0,59*	−1,01***	−0,75**	−0,44***
	(0,29)	(0,22)	(0,26)	(0,09)
Interaktionen etc.				
Geburtsjahr (in 10 J.)				0,44***
				(0,09)
Mann*Geburtsjahr				−0,22***
				(0,02)
Prestige*Geburtsjahr				−0,01***
				(0,00)
Konstante	−0,25	1,13***	1,26***	0,32***
	(0,29)	(0,22)	(0,38)	(0,09)
Pseudo-R^2 (NK)	0,37	0,27	0,28	0,35
Pseudo-R^2 (MF)	0,28	0,18	0,18	0,24
AIC	0,59	0,97	1,11	0,84
N	3226	4887	1666	23641

Referenzkategorien: Frau, Vater Abitur, Mutter Abitur
Alter, Prestige und Geburtsjahr zentriert
†: p ≤ 0,1; *: p ≤ 0,05; **: p ≤ 0,01; ***: p ≤ 0,001

samtmodell weist einen statistisch signifikanten, negativen Interaktionseffekt zwischen Kohorte und Prestige auf. Aufgrund der Nicht-Linearität des Modells ist es jedoch auch hier sinnvoll, vorhergesagte Wahrscheinlichkeiten zu berechnen. Abbildung 6 a zeigt daher die Vorhersagen aus dem Interaktionsmodell, und 6 b stellt die Ergebnisse der getrennten Modellschätzungen zusammen. Vergleicht man Abbildungen 5, 6 a und 6 b, sieht man, dass sich der Verlauf der Kurven zwischen den Abbildungen nur geringfügig unterscheidet. In allen Abbildungen wird der Wahrscheinlichkeitsverlauf in der ältesten Kohorte zunehmend steiler; in der jüngsten Kohorte hingegen ist die Kurve zwar auch deutlich steigend, verläuft aber nahezu linear. Insofern ist in diesem Fall die substanzielle Interpretation identisch mit den Ergebnissen des einfacheren Modells 3 (siehe Tab. 2 und Abbildung 5). Dies spiegelt die Tatsache wider, dass die logistische Regression als nichtlineares Modell interdependente Effekte in einem gewissen Umfang auch ohne explizite Parametrisierung modelliert.

4 Häufige Fehler

Prinzipiell ist die logistische Regression ein recht einfach anzuwendendes Analyseverfahren. Probleme in der Anwendung bzw. der Interpretation der Ergebnisse resultieren meist daraus, dass die Gemeinsamkeiten mit der linearen (OLS)-Regression überschätzt werden.

Erstens muss immer berücksichtigt werden, dass die logistische Regression lediglich in Bezug auf die Logits linear-additiv parametrisiert ist. In Bezug auf die Wahrscheinlichkeiten beschreiben Logit-Modelle *nichtlineare Effekte, die nicht in einem einzelnen Koeffizienten ausgedrückt werden können.* Daher besteht das Risiko einer Fehlinterpretation, wenn lineare Beziehungen zwischen den unabhängigen Variablen und der Wahrscheinlichkeit $P(y = 1)$ angenommen und die β-Koeffizienten blauäugig wie in der OLS-Regression interpretiert werden. Gleichzeitig besteht das Risiko, dass die Ergebnisse nur zu oberflächlich interpretiert werden, wenn lediglich die Richtung des Zusammenhangs angegeben wird. Um beides zu vermeiden, plädieren wir für die routinemäßige Berechnung von Conditional-Effect-Plots, bei denen die vorhergesagten Wahrscheinlichkeiten für bestimmte Ausprägungen einer unabhängigen Variablen gegen eine andere unabhängige Variable geplottet werden. Eine Alternative hierzu ist die Berechnung von average marginal effects, die den durchschnittlichen additiven Effekt auf die Wahrscheinlichkeit von $y = 1$ angeben. Allerdings geht mit der Verwendung von AMEs ein Informationsverlust einher, da sie die Nichtlinearität der Beziehung nicht wiedergeben können. Die Verwendung von entlogarithmierten Koeffizienten (Odds-Ratios) halten wir für keine geeignete Alternative zu β-Koeffizienten, da auch Odds-Ratios in der Interpretation extrem komplex sind. Es ist zu beachten, dass a) Odds nichtlinear mit Wahrscheinlichkeiten verknüpft sind, und daher b) ein gegebenes Odds-Ratio bei unterschiedlichen Basiswahrscheinichkeiten für völlig unterschiedliche Wahrscheinlichkeitsverhältnisse stehen kann. Da zudem praktisch niemand ein Alltagsverständnis von Odds hat (geschweige denn von Odds-Ratios), ist zu befürchten, dass sie beim Lesen implizit als Wahrscheinlichkeitsverhältnisse (bzw. als „so etwas ähnliches“) aufgefasst werden. Teilweise geschieht dies sogar explizit.

Beispielsweise werden in einem Aufsatz, der 2008 in der KZfSS erschienen ist, die Ergebnisse der Logitmodelle wie folgt interpretiert: „So sinken mit jeder Zunahme auf der in fünf Stufen erhobenen Bildungsskala die *Teilnahmewahrscheinlichkeit*, die odds der Spielteilnahme um das 0,8-fache, d. h. jeweils um durchschnittlich 20 Prozent“ [Hervorhebung HB/CW, Fehler im Original]. Selbstverständlich ist die Interpretation nicht korrekt. Fehlinterpretationen dieser Art können jedoch leicht vermieden werden, indem auf die Verwendung von Odds-Ratios verzichtet wird.

Ein zweites mögliches Problem stellen Interaktionseffekte dar. Zwar ist die Vorgehensweise (und auch die Interpretation) in Bezug auf die Logits analog zur linearen Regression durchzuführen, aber sozialwissenschaftliche Hypothesen beziehen sich in aller Regel auf Wahrscheinlichkeiten, nicht auf Logits. In Bezug auf die Wahrscheinlichkeiten kann der Effekt einer Variable x_1 jedoch zu einem gewissen Ausmaß auch ohne Spezifikation eines multiplikativen Terms vom Niveau einer anderen Variable x_2 abhängen. Logistische Regressionen sind erstens weniger sensibel bei der Identifikation von Interaktionseffekten, und zweitens komplexer in der Interpretation als dies bei OLS-Regressionen der Fall ist. Es wurde vorgeschlagen, auch hier auf Conditional-Effect-Plots zurückzugreifen.

Eng verbunden mit diesen beiden Punkten ist, drittens, eine Besonderheit beim Vergleich von Koeffizienten zwischen Modellen. Da die Varianz der (latenten) abhängigen Variablen unbekannt ist und die Residualvarianz in der logistischen Regression als konstant angenommen wird, verändert sich mit der Aufnahme weiterer Variablen in ein Modell nicht nur die erklärte Varianz, sondern auch die Gesamtvarianz der latenten abhängigen Variablen. Es verändert sich also die Skala der abhängigen Variablen, so dass eine Veränderung der β-Koeffizienten nicht notwendigerweise auf die Kontrolle eines Drittvariableneffektes zurückgeführt werden kann. Sollen Koeffizienten beim schrittweisen Modellaufbau verglichen werden, empfiehlt sich daher die Verwendung der robusteren standardisierten Logitkoeffizienten oder der durchschnittlichen marginalen Effekte (AME).

Viertens ist zu beachten, dass die in der logistischen Regression verwendeten Pseudo-R^2-Koeffizienten auf Veränderungen der Likelihood eines Modells basieren und sich nicht – wie in der OLS-Regression – als Maß der erklärten Varianz interpretieren lassen. Zudem gibt es eine Vielzahl verschiedener Varianten von Pseudo-R^2, die in ihrer Größe nicht miteinander vergleichbar sind.

5 Literaturempfehlungen

Die nach Ansicht der Autoren beste Monographie zur logistischen Regression ist das englischsprachige Lehrbuch von Long (1997). Eine stärker anwendungsorientierte Einführung (für Stata) findet sich bei Long & Freese (2006). Auch Wooldridge (2002) ist eine hervorragende Darstellung, die jedoch stark formalisiert ist und daher nicht für den Einstieg empfohlen werden kann. Hierfür eignet sich die Einführung von Menard (1995). Auf dem deutschsprachigen Markt ist das Angebot an geeigneten Darstellungen deutlich geringer. Gut verständlich sind die Ausführungen von Andreß et al. (1997), tiefer gehende Diskussionen finden sich bei Tutz (2000).

Literaturverzeichnis

Ai, C. & Norton, E. C. (2003). Interaction Terms in Logit and Probit Models. *Economics Letters*, 80, 123–129.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov & B. F. Csaki (Hg.), *Second International Symposium on Information Theory* (S. 267–281). Budapest: Academiai Kiado.

Andreß, H.-J., Hagenaars, J. A., & Kühnel, S. (1997). *Analyse von Tabellen und kategorialen Daten. Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Berlin: Springer.

Becker, R. & Lauterbach, W., Hg. (2008). *Bildung als Privileg*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Best, H. (2008). Die Umstellung auf ökologische Landwirtschaft. Empirische Analysen zur Low-Cost-Hypothese des Umweltverhaltens. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 60, 314–338.

Cox, D. R. & Snell, E. J. (1989). *The Analysis of Binary Data*. London: Chapman & Hall.

Cragg, J. G. & Uhler, R. (1970). The Demand for Automobiles. *Canadian Journal of Economics*, 3, 386–406.

Huang, C. & Shields, T. G. (2000). Interpretation of Interaction Effects in Logit and Probit Analyses. *American Politics Research*, 28, 80–95.

Hubert, T. & Wolf, C. (2007). Determinanten der beruflichen Weiterbildung Erwerbstätiger. Empirische Analysen auf der Basis des Mikrozensus 2003. *Zeitschrift für Soziologie*, 36, 473–493.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Long, J. S. & Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata*. College Station: Stata Press.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behaviour. In P. Zarembka (Hg.), *Frontiers in Econometrics* (S. 105–142). New York: Academic Press.

Menard, S. (1995). *Applied Logistic Regression*, Band 07-106 von *Quantitative Applications in the Social Sciences*. Thousand Oaks: Sage.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodolgy*, 25, 111–163.

Treiman, D. J. (1977). *Occupational Prestige in Comparative Perspective*. New York: Academic Press.

Tutz, G. (2000). *Die Analyse kategorialer Daten - eine anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression*. München: Oldenbourg Verlag.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.