Ordinale Regression

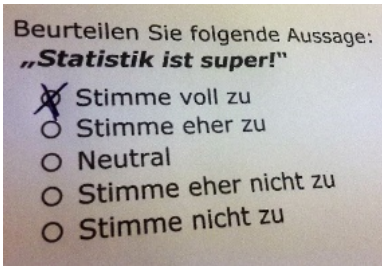
Die ordinale Regression umfasst Modelle, deren Zielvariable ordinal skaliert ist, d.h. es liegt eine kategoriale Variable vor deren Ausprägungen eine Rangordnung vorweisen, z.B. Schulnoten („1“, „2“, „3 “, …,“6“), Ausprägung einer Krankheit („gesund“, „leicht krank“, „mittel krank“, „schwer krank“) oder Zufriedenheit mit einem Produkt (Skala von 0 bis 10). Dabei gilt, dass die Abstände zwischen den Ausprägungen nicht interpretierbar sind, d.h. dass z.B. der Abstand zwischen den Schulnoten „1“ und „2“ nicht dem Abstand der Noten „4“ und „5“ entsprechen muss, weshalb die Nutzung von linearen Regressionsmodellen unangemessen ist. Anstelle von ordinalen Regressionsmodellen könnten Daten mit ordinaler Skalierung auch mit multinominalen Logitmodellen (Logistische Regression (Logit-Modell)) untersucht werden. Jedoch würde in diesem Fall die Information über die Reihenfolge der kategorialen Ausprägungen nicht genutzt werden, was zu ineffizienten Schätzern führen würde. Je nachdem welche Annahme über die Verteilung der Fehlerterme getroffen wird, sprechen wir von geordneten logistischen Regressionen (logistisch verteilt) oder geordneten Probitregressionen (standardnormalverteilt). Wenn eine höhere Kategorie nur erreicht werden kann, wenn eine niedrigere schon erreicht wurde, können sogenannte sequentielle Modelle genutzt werden. Ein Beispiel ist die Variable Dauer der Arbeitslosigkeit unterteilt in verschiedene Kategorien: "1 Jahr", "2 Jahre", "mehr als 2 Jahre". Eine Person kann nur zwei Jahre arbeitslos sein, wenn sie schon ein Jahr arbeitslos war.

Dieser Artikel ist analog zum Artikel Logistische Regression (Logit-Modell) aufgebaut.

Inhaltsverzeichnis

Variablen und deren Zusammenhang

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



Das Ziel der ordinalen Regression ist die Vorhersage von Wahrscheinlichkeiten für das Auftreten der einzelnen Kategorien in Abhängigkeit von Kovariablen.

Ein Beispiel für die Anwendung der ordinalen Regression stellen Likert-Skalen dar. Sie sind ein Spezialfall von Ordinalskalen, das heißt die Werte einer solchen Skala sind verschiedenartig und lassen sich einer eindeutigen Rangfolge zuordnen. Likert-Skalen werden genutzt, um persönliche Einstellungen (Zustimmung/Ablehnung) von Individuen zu messen, weshalb sie häufig als Antwortskalen in Umfragen verwendet werden. Typischerweise haben sie 3, 5, 7 oder 11 Werte.

Typische Likert-Skalen sind:

- überhaupt nicht (1) - wenig (2) - mittel (3) - stark (4) - sehr stark (5)
- trifft zu (1) - teils/teils (2) - trifft nicht zu (3)

Die erste Variable hat 5 Kategorien, die mit den Werten 1, 2, 3, 4 und 5 kodiert werden.

Es mag die Frage aufkommen, ob Likert-Skalen nicht auch mit linearen Regressionsmodellen analysiert werden können. Likert-Skalen stellen keine metrische abhängige Variable dar, daher sollte von der Nutzung von linearen Regressionsmodellen in der Regel abgesehen werden, es gibt jedoch einige Ausnahmen wie z.B. eine relativ große Anzahl an Kategorien (z.B. 11). Für weitere Informationen siehe hier.

Motivation über Schwellenwertmodelle

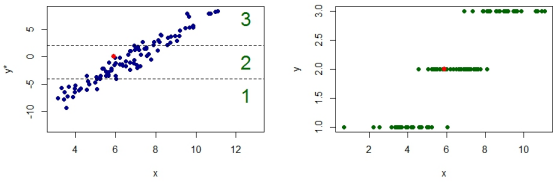
Ordinale Regressionsmodelle werden über Schwellenwertmodelle motiviert. Eine nicht beobachtbare Hintergrundvariable (auch latente Variable genannt) y^* wird angenommen, die metrisch ist. Das Modell lautet: $y^*_i=x'_i\beta+\epsilon_i, \quad i=1,\dots,n$. Hierbei wird für die Verteilung der ϵ_i eine Verteilungsfunktion F mit Erwartungswert 0 angenommen, die symmetrisch um die 0 verteilt ist. Häufig wird für F die Standardnormalverteilung angenommen. In diesem Spezialfall erhält man das Probit-Modell.

Anstatt den Zusammenhang zwischen der beobachtbaren ordinalen Variable und den Einflussvariablen zu schätzen, wird der Zusammenhang zwischen der latenten metrischen Variable und den Einflussvariablen geschätzt. Dabei stellen die beobachtbaren Kategorien (mit endlicher Anzahl an Kategorien (m + 1)) das Überschreiten der Schwelle der latenten metrischen Variablen dar:

$$y_i={\begin{cases}0\ ,&{\text{für}}&-\infty<y^*_i\leq\alpha_1\ ,\\1\ ,&{\text {für}}\\&\alpha_1<y^*_i\leq\alpha_2\ ,\\&\vdots\\m\ ,&{\text{für}}&\alpha_m<y^*_i\leq\infty\end{cases}}$$

α_j ($j = 1, 2, ..., m$) stehen für geordnete Schwellenwerte, die neben den β auch geschätzt werden müssen. Weil der Wertebereich der latenten Variable nicht bekannt ist, nimmt man $\alpha_0 = - \infty$ und $\alpha_{m+1} = \infty$ an.

Die folgende Abbildung zeigt den Zusammenhang für eine Likert-Skala mit 3 Kategorien und einer Einflussvariablen x. Der rote Punkt hat im ersten Bild die Koordinaten $x = 5,9$ und $y^* = 0.05$. Da y^* zwischen $\alpha_1 = -4$ und $\alpha_2 = 1$ liegt, nimmt der y - Wert im zweiten Bild einen Wert von 2 an.



Abbildung

Grafischer Zusammenhang zwischen der Hintergrundvariable y^* und einer Likert-Skala mit 3 Kategorien.

Bei der ordinalen Regression werden die Wahrscheinlichkeiten für das Auftreten von Kategorien $j = 1, 2, ..., m$ durch erklärende Variablen $x_1, x_2, ..., x_P$ mit Hilfe von bedingten kumulierten Wahrscheinlichkeiten $P(y \leq j |x_i) = P(Y_i = 1 |x_i) + ... + P(Y_i = j |x_i)$ geschätzt, daher spricht man auch von kumulativen (Logit-/Probit-) Modellen.

Weil die Wahrscheinlichkeit für das Eintreten der höchsten Kategorie oder einer niedrigeren 100% beträgt ($P(Y_i \leq m |x_i) = 1$), werden nur $m - 1$ Kategorien modelliert, damit eine Überparametrisierung verhindert werden kann. Als Referenzkategorie wird im ordinalen Logitmodell entweder die kleinste oder größte Kategorie der Zielvariable ausgewählt.

Aus der Modellannahme über die latente Variable und das Schwellenwertkonzept ergibt sich das kumulative Modell mit Verteilungsfunktion F

$P (Y_i = 1 | x_i) = F(\alpha_1 - x'_i \beta)$

$P (Y_i = j | x_i) = F(\alpha_j - x'_i \beta) - F(\alpha_{j - 1} - x'_i \beta),$ $j = 2, ..., m - 1$

$P (Y_i = m | x_i) = 1 - F(\alpha_{m - 1} - x'_i \beta)$

Die Wahl von F bestimmt, ob ein Logit, Probit oder anderes Modell vorliegt.

Das kumulative Logit-Modell

Die Wahl der logistischen Funktion $F(x) = \frac{e^x}{1+ e^x}$ resultiert im kumulativen Logit-Modell - auch Proportional Odds Modell genannt -

$$P(Y_i \leq j | x_i) = \frac{exp(\alpha_j - x'_i \beta)}{1 + exp(\alpha_j - x'_i \beta)}$$

oder

$$log \left(\frac{P(Y_i \leq j | x_i)}{P(Y_i > j | x_i)} \right) = log \left(\frac{P(Y_i \leq j | x_i)}{1 - P(Y_i \leq j | x_i)} \right) = \alpha_j - x'_i \beta$$

Während die Schwellenparameter α_j kategorienspezifisch sind und monoton in j steigen, gehen die Einflussvariablen x_i global ins Modell ein und führen zu globalen Regressionskoeffizienten β. Die Parameter α_j und β werden durch die Maximum-Likelihood (ML) Methode geschätzt. Aus Identifikationsgründen darf die Designmatrix (Matrix der erklärenden Variablen) keinen Intercept enthalten, sonst könnten die Schwellenwerte davon nicht unterschieden werden und blieben unidentifiziert. Die ML-Schätzer sind konsistent, asymptotisch effizient und asymptotisch normalverteilt.

Geordnete Probitregression

Die geordnete Probitregression folgt den gleichen Überlegungen wie die geordnete logistische Regression. Der Unterschied liegt in der Annahme über die Verteilung der Fehlerterme, welche bei der Probitregression als standard normal verteilt angenommen wird. Weitere Linkfunktionen sind z.B. log-log, kumulative inverse Cauchy-Verteilung, etc.

Sequentielle Modelle

Neben kumulativen Modellen stellen sequentielle Modelle eine wichtige Modellklasse der ordinalen Regression dar. Sequentielle Modelle werden genutzt, wenn eine höhere Kategorie nur erreicht werden kann, wenn eine niedrigere schon erreicht wurde, d.h. wenn die Kategorien nur sukzessiv/schrittweise erreicht werden können. Ein Beispiel ist die abhängige Variable Dauer der Arbeitslosigkeit unterteilt in verschiedene Kategorien: "1 Jahr", "2 Jahre", "mehr als 2 Jahre". Eine Person kann nur mehr als zwei Jahre arbeitslos sein, wenn sie schon zwei Jahre arbeitslos war.

Der schrittweise Übergang zwischen den Kategorien wird dabei durch dichotome Zusammenhänge mit Hilfe von binären Regressionsmodellen modelliert. Der Prozess, der nur zur Modellierung dient, startet mit $Y_i=1$ und modelliert den Übergang zu $Y_i=1$ durch ein binäres Modell:

$$P(Y_i=1 |x_i) = F(\alpha_1 + x'_i\beta)$$

Wenn die Zielvariable Y in der ersten Kategorie verbleibt, endet der Prozess, ansonsten geht es sukzessive weiter mit dem nächsten Übergang von $Y_i =2$ zu $Y_i >2$:

$$P(Y_i=2 |Y_i \geq 2, x_i) = F(\alpha_2 + x'_i\beta)$$

Der entsprechend j-te Übergang sieht wie folgt aus:

$$P(Y_i=j |Y_i \geq j, x_i) = F(\alpha_j + x'_i\beta), \qquad \forall j=1,...,m$$

Wenn ein Übergang nicht stattfindet, endet der schrittweise Mechanismus, d.h. eine Kategorie j ($j = 1, 2, ..., m$) ist erreicht.

Wie bei den kumulativen Modellen bestimmt die Wahl der Verteilungsfunktion F das tatsächliche Modell. Bei der logistischen Verteilungsfunktion, spricht man vom sequentiellen Logit-Modell, während ein sequentielles Probit-Modell vorliegt, wenn die Standardnormalverteilung als Verteilungsfunktion F gewählt wird. Das Modell kann erweitert werden, indem Koeffizienten β genutzt werden, die kategorienspezifische Effekte beschreiben.

Im folgenden Artikel werden sequentielle, sowie Probitmodelle, nicht im Detail ausgeführt. Der Fokus liegt auf kumulativen Logit-Modellen, die in der praktischen Anwendung am häufigsten genutzt werden. Einen Überblick über die Unterschiede von kumulativen Logit-Modellen und sequentiellen Modellen gibt der verlinkte Artikel.

Einführung in das Beispiel

Einführung in das Beispiel: Schulabschluss

Die abhängige Variable Schulabschluss ("SCHULABSCHLUSS") ist ordinal skaliert und hat die Ausprägungen "HAUPT" für Volks-/ Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse, "MITTEL" für Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse und "ABI" für Fachhochschulreife (Abschluss an einer Fachoberschule etc.)/Abitur bzw. Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife). In diesem Beispiel sind die erklärenden Variablen eine Dummy-Variable für das Geschlecht ("GESCHL", 1 = männlich, 2 = weiblich), eine metrisch skalierte Variable "NETTO", die das Nettoeinkommen einer Person misst und eine Faktorvariable Schulabschluss des Vaters "SCHULABSCHLUSS_V" mit den gleichen Ausprägungen wie die der abhängigen Variablen.

Die Variablen stammen aus einem vereinfachten Datensatz der ALLBUS-Umfrage (siehe Beispieldatensätze und Beispielprogramme).

Um einen ersten Überblick über die verwendeten Variablen zu gewinnen, werfen wir einen Blick auf einfache deskriptive Statistiken der genannten Variablen. Die folgenden Tabellen beschreiben die Häufigkeiten der Faktorvariablen und einige Eckdaten über die metrische Variable "NETTO".

SCHULABSCHLUSS

HAUPT	MITTEL	ABI
603	780	897

GESCHL

MAENNLICH	WEIBLICH
1188	1092

SCHULABSCHLUSS_V

HAUPT	MITTEL	ABI
1428	415	437

NETTO

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
37	800	1300	1597	2000	60000

Interpretation der Parameter und anderer Kenngrößen

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.

Es lassen sich jedoch auch wie bei einem linearen Regressionsmodell Wahrscheinlichkeiten vorhersagen, indem man Werte für alle unabhängigen Variablen einsetzt. Hier ein Beispiel:

Wahrscheinlichkeit, mit der laut dem geschätzten Modell, eine Frau, die 1200 Euro im Monat verdient und dessen Vater Abitur hat, höchstens einen mittleren Schulabschluss hat:

$$P(Y_i \leq j \mid x_i) = \frac{\exp(\alpha_j - x'_i \beta)}{1 + \exp(\alpha_j - x'_i \beta)}$$

Die entsprechenden Werte wurden dem R-Output entnommen.

$$\hat{p}=\frac{\exp(2.0906 - (1 \times 0.244 + 1 \times 2.327 + 1200 \times 0.0005))}{1 + \exp(2.0906 - (1 \times 0.244 + 1 \times 2.327 + 1200 \times 0.0005))} \approx 0.253$$

Die besagte Frau hat also mit einer vorhergesagten Wahrscheinlichkeit von 25,3 % höchstens einen mittleren Schulabschluss.

Da die marginalen Effekte keiner so direkten Interpretation wie im linearen Modell zugänglich sind und die vorhergesagten Wahrscheinlichkeiten auch nur spezielle Aussagen ermöglichen, werden oft die sogenannten Odds, Log-Odds (Logits) oder die Oddsratio betrachtet.

Der Quotient $\frac{P(Y_i \leq j \mid x_i)}{P(Y_i > j \mid x_i)}$ wird oft als "Chance" oder englisch "odds" bezeichnet. Odds beschreiben das Verhältnis von Wahrscheinlichkeit zur Gegenwahrscheinlichkeit.

Da die Odds exponentiell sind, bietet es sich an, sie zu logarithmieren, um Zusammenhänge zu linearisieren. So entstehen die **Log-Odds**, auch **Logits** genannt:

$$ log \left(\frac{P(Y_i \leq j | x_i)}{P(Y_i > j | x_i)} \right) = log \left(\frac{P(Y_i \leq j | x_i)}{1 - P(Y_i \leq j | x_i)} \right) = \alpha_j - x'_i \beta $$

Die erklärenden Variablen stehen nun in einem linearen Zusammenhang mit den logarithmierten Odds.

Die **Oddsratio (OR)** setzt die Odds in Relation. Wenn eine Einflussvariable x_{i1} um eine Einheit auf $x_{i1} + 1$ erhöht wird, gilt

$$ \frac{\frac{P(Y_i \leq j| x_{i1} + 1)}{P(Y_i > j| x_{i1} + 1)}}{\frac{P(Y_i \leq j| x_{i1})}{P(Y_i > j| x_{i1})}} = exp(\beta_1)$$

Die Odds-Ratio hängt nicht von einer einzelnen Kategorie j ab, sondern nur von den Differenzen in den Kovariaten. Eine wichtige Annahme des kumulativen Logit-Modells ist, dass die Beziehung zwischen jeder möglichen Kombination an Stufenpaaren der Zielvariablen gleich ist. Deswegen kann der Effekt einer erklärenden Variablen durch einen globalen β-Koeffizienten dargestellt werden. Dieser Effekt gilt dann für jeden Stufenwechsel zwischen den Ausprägungen der Zielvariable und ändert sich auch nicht, wenn Kategorien zusammengefasst werden. Daher wird das Modell auch "Proportional Odds"-Modell genannt. Die Annahme ist bekannt unter „proportional Odds“- oder „equal slopes"-Annahme.

Bei einer Odds-Ratio (also $exp(\beta)$), die größer als eins ist, hat die Einflussvariable einen positiven Effekt auf die abhängige Variable, da die Chance $\frac{P(Y_i \leq j | x_i)}{P(Y_i > j | x_i)}$ größer wird, wenn die Einflussvariable um eine Einheit erhöht wird. Im Gegenzug hat die Variablen einen negativen Einfluss bei einer Odds-Ratio kleiner als eins. Wenn die Odds-Ratio gleich eins ist, hat die Variable keinen Einfluss, weil die Chance/Odds $\frac{P(Y_i \leq j | x_i)}{P(Y_i > j | x_i)}$ gleich bleibt.

Oddsratio beispielhaft für die Variable Nettoeinkommen:

OR: $\frac{odds(NETTO_i + 1)}{odds(NETTO_i)} = exp(\beta_{NETTO}) = exp(0.000506) = 1.000506$.

Die Chance, einen höheren Schulabschluss zu haben, steigt um den Faktor 1.000506, wenn man 1 € pro Monat mehr verdient.

Logistische Regression (Logit-Modell)).

binären Logit-Modell

Modellselektion.

AIC (Akaike-Information-Criterion)

BIC (Bayesian-Information-Criterion)

Berechnung der Pseudo R² im Beispiel:

$\text{McFadden } R^2 = 1 - \frac{-2164.66}{-2475.44} = 0.126$

$\text{Cox&Snell } R^2 = 1 - \left(\frac{L_{null}}{L_{voll}}\right)^{\frac{2}{2280}} = 0.238$
$\text{Nagelkerkes } R^2 = \frac{1 - \left(\frac{L_{null}}{L_{voll}}\right)^{\frac{2}{2280}}}{1 - (L_{null})^{\frac{2}{2280}}} = 0.269$

$R^2 > 0.2$: Modellanpassung ist akzeptabel

$R^2 > 0.4$: Modellanpassung ist gut

$R^2 > 0.5$: Modellanpassung ist sehr gut

Likelihood Ratio Test

(Chi-Quadrat-Verteilung)

Maximum-Likelihood Estimator

Ein interaktives Tool für den Maximum-Likelihood Estimator wurde vom Institut für Meteorologie an der Freien Universität Berlin entwickelt.

F-Test der multiplen linearen Regression

p-Wert Signifikanzniveau

Berechnung des p-Wertes im Beispiel:

\(-2[(-2475.44) - (-2164.6)] = 621.68 \)

\(\chi^2_4(621.68) = 1\)

\(p = 1-1 = 0 \)

Zu den gängigen Signifikanzniveaus kann die Nullhypothese, dass das volle Modell keine Erklärungskraft besitzt, abgelehnt werden.

(Link, S. 21)

Standardwerkzeugen

Komponenten und Begriffe

Die Güte des Modells

1. Gesamtzahl an Beobachtungen:

Die gesamte Anzahl an Beobachtungen im Datensatz entspricht der Anzahl an Zeilen. Diese wird häufig mit **n** gekennzeichnet. In diesem Datensatz gibt es insgesamt **2280 Beobachtungen**.

2. Gelöschte Beobachtungen:

Bei fehlenden Werten in Variablen können Beobachtungen für die Modellanalyse nicht berücksichtigt werden. Im Beispiel sind dies **0 Beobachtungen**.

3. Zahl der Beobachtungen:

Hiermit ist die Zahl der Beobachtungen gemeint, die zur Anpassung des Modells genutzt wird. Das bedeutet, dass diese Anzahl sich aus der Differenz der Gesamtzahl an Beobachtungen und den gelöschten Beobachtungen auf Grund von fehlenden Werten in den gewünschten Variablen ergibt. In dem Modell wurden **2280 Beobachtungen** genutzt.

6. Pseudo R^2

Das geschätzte Modell im Beispiel hat ein McFadden R^2 von **0.126**. Diese Zahl ist nicht direkt interpretierbar (siehe Faustregel oben).

8. Standardfehler des Schätzers:

Da das Logit Modell nicht analytisch lösbar ist, wird der Schätzer numerisch mittels der Maximum-Likelihood Methode ermittelt. Über diese Art von Schätzern können nur asymptotische Aussagen getroffen werden. So entspricht auch der Standardfehler asymptotisch dem Inversen der Fisher-Information.

Schätzergebnisse

9. Abhängige oder endogene Variable:

Im Beispiel ist der **Schulabschluss (SCHULABSCHLUSS)** die abhängige Variable.

10. Erklärende oder exogene Variable:

Im Beispiel sind das **Geschlecht (GESCHL)**, der **Schulabschluss des Vaters (SCHULABSCHLUSS_V)** und das **Nettoeinkommen (NETTO)** die erklärenden Variablen.

11. Geschätzte Parameter:

Bei der ordinalen Regression werden die Schwellenwertparameter (**11a**), sowie die Regressionskoeffizienten **(11b**) geschätzt. Die Interpretation ist schwieriger als im linearen Regressionsmodell. In der praktischen Anwendung ist es nicht üblich, die Schwellenwerte zur Interpretation des Modells zu nutzen.

Schätzung im Beispiel:

$log\left(\frac{P(Y_i \leq j \mid x_i)}{1 - P(Y_i \leq j \mid x_i)}\right) = \alpha_j - x'_i \beta$

Interpretation der Parameter:

Der Schwellenwertparameter (α_j) zwischen HAUPT und MITTEL entspricht 0.2999, der zwischen MITTEL und ABI 2.0906.

Der Steigungsparameter für "NETTO" enspricht bspw. 0.000506. Um diesen sinnvoll zu interpretieren, betrachtet man die Odds Ratio: $exp(\beta_{NETTO}) = exp(0.000506) = 1.000506$. D.h. die Chance, einen höheren Schulabschluss zu haben, steigt um den Faktor 1.000506, wenn man 1 € pro Monat mehr verdient.

12. Standardabweichung der Schätzung (Standardfehler, $\widehat{SF}_{\beta_p}$):

Da die Parameter basierend auf einer Zufallsstichprobe geschätzt werden, unterliegen diese Schätzungen einer gewissen Ungenauigkeit, die durch die Standardabweichung der Schätzung quantifiziert wird. Standardfehler werden genutzt, um statistische Signifikanz zu überprüfen und um Konfidenzintervalle zu bilden.

13. Z-Statistik (empirischer Z-Wert).

Mit Hilfe eines Wald- oder Likelihood-Ratio Tests lässt sich prüfen, ob die Nullhypothese, dass ein Koeffizient gleich 0 ist, abgelehnt werden kann. Wenn dies nicht der Fall sein sollte, ist davon auszugehen, dass die zugehörige Kovariate keinen signifikaten Einfluss auf die abhängige Variable ausübt, d.h. die erklärende Variable ist nicht sinnvoll, um die Eigenschaften der abhängigen Variablen zu erklären.

Hypothese: $H : \beta_p = 0$ gegen $A : \beta_p \neq 0$

Teststatistik: $T = \frac{\hat{\beta_p}}{\widehat{SF}_{\beta_p}}$

Überprüfen, ob z.B. das Geschlecht einen Einfluss auf den Schulabschluss hat, anhand der Z-Statistik:

Die Teststatistik vom Parameter für das Einkommen ist $T_{GESCHL} = \frac{-0.244}{0.087} \approx -2.81.$ Diese Teststatistik wird mit dem kritischen Wert verglichen:

$|T_{GESCHL}| = 2.81 > 1.96 = z_{1-\frac{\alpha}{2}}$.

Schon anhand der Teststatistik kann man erkennen, dass die Nullhypothese $\beta_{GESCHL} = 0$ hier abgelehnt werden kann, d.h. dass das Geschlecht einen signifikanten Einfluss auf den Schulabschluss hat.

14. p-Wert zur Z-Statistik:

Zusätzlich zur Z-Statisik wird meistens ein p-Wert ausgegeben. Aus einer methodisch-praktisch orientierten Perspektive gibt der p-Wert das kleinste Signifikanzniveau an, zu dem die Nullhypothese $\beta_p=0$ gerade noch abgelehnt werden kann. Ist also das tatsächliche Signifkanzniveau α, welches vor dem Test gewählt wird, geringer als der p-Wert, so kann die Nullhypothese nicht abgelehnt werden.

Überprüfen, ob z.B. das Geschlecht einen Einfluss auf den Schulabschluss hat, anhand des p-Wertes:

Im Beispiel liegt der p-Wert zur Nullhypothese $\beta_{GESCHL} = 0$ bei 0.005. Daraus kann man schließen, dass das Geschlecht einen signifikanten Einfluss auf den Schulabschluss hat, und zwar schon bei einem Signifikanzniveau von 0.001.

Der p-Wert gibt die Wahrscheinlichkeit an, dass wir unter der Nullhypothese einen Wert der Teststatistik beobachten, der noch stärker in Richtung Ablehnung der Nullhypothese geht. Das heißt er macht eine Aussage über die Wahrscheinlichkeit der Beobachtung der Stichprobe, nicht aber direkt über die Wahrscheinlichkeit der Nullhypothese selbst.

Zum p-Wert gibt es viele Missverständnisse, selbst in veröffentlichter Literatur. Aussagen wie z.b. dass "der p-Wert den Fehler 1. Art wieder gibt" bzw. "die Wahrscheinlichkeit ist, dass unsere Hypothese wahr ist, gegeben, dass der Test abgelehnt wird", sind falsch und sollten in Arbeiten vermieden werden.

Eine gute Quelle für die den richigen Umgang und ein tieferes Verständnis vom p-Wert gibt es beispielsweise hier.

15. 95%-Konfidenzintervall:

Konfidenzintervalle sind im Allgemeinen eine Möglichkeit, die Genauigkeit der Schätzung zu überprüfen. Ein 95%-Konfidenzintervall ist der Bereich, der im Durchschnitt in 95 von 100 Fällen den tatsächlichen Wert des Parameters einschließt.

Beispielhaftes Konfidenzintervall für den Regressionskoeffizienten der Variable Geschlecht:

[-0.244 - 1.96 * 0.087; 0.244 + 1.96 * 0.087] = [-0.41452 ; -0.07348]

Outputs in den verschiedenen Statistikprogrammen

Beispieldaten

Hier werden die Outputs aus den verschiedenen Statistikprogrammen vorgestellt. Die Outputs einer ordinalen Regression unterscheiden sich teils in den verschiedenen Statistikprogrammen. Sowohl sind die Werte unterschiedlich angeordet, als auch werden teils nicht die gleichen Werte ausgegeben. Einen hilfreichen Überblick über ausführlichere Code-Beispiele für die verschiedenen Statistikprogramme liefert folgende Website (Unterpunkt 'Ordinal Logistic Regression').

Im Folgenden werden die Werte 1-15, wenn vorhanden, an den Output der verschiedenen Statistikprogramme geschrieben, damit die Werte im Output gefunden werden können.

Output in R

```
formula: SCHULABSCHLUSS ~ GESCHL + SCHULABSCHLUSS_V + NETTO
data:    allbus
 1
 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  2280 -2164.66 4341.33 5(0)  6.09e-08 4.9e+07

Coefficients:                 11b        12        13       14
 10                    Estimate Std. Error z value Pr(>|z|)
GESCHLWEIBLICH         2.443e-01  8.705e-02   2.807    0.005 **
SCHULABSCHLUSS_VMITTEL 1.266e+00  1.082e-01  11.698   <2e-16 ***
SCHULABSCHLUSS_VABI    2.327e+00  1.270e-01  18.321   <2e-16 ***
NETTO                  5.060e-04  4.813e-05  10.512   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:      12        13
      11a   Estimate Std. Error z value
HAUPT|MITTEL   0.2999     0.1044   2.873
MITTEL|ABI     2.0906     0.1146  18.246
> # Parameterinterpretation (Odds-Ratios)
> coef <- coef(fit)
> exp(coef[3:6])
         GESCHLWEIBLICH SCHULABSCHLUSS_VMITTEL    SCHULABSCHLUSS_VABI                NETTO
               1.276760               3.547321              10.242412             1.000506
```

Output in Stata

```
. ologit SCHULABSCHLUSS i.GESCHL i.SCHULABSCHLUSS_V NETTO

Iteration 0:   log likelihood = -2475.4436
Iteration 1:   log likelihood = -2194.8342
Iteration 2:   log likelihood = -2164.7945
Iteration 3:   log likelihood = -2164.6627
Iteration 4:   log likelihood = -2164.6627

Ordered logistic regression                     3 Number of obs   =       2280
                                                  LR chi2(4)      =     621.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -2164.6627                     6 Pseudo R2       =     0.1255

------------------------------------------------------------------------------
 9 SCHULABSCHLUSS |    11bCoef.  12 Std. Err.     13 z   14 P>|z|    15 [95% Conf. Interval]
-----------------+------------------------------------------------------------
10       2.GESCHL |    .2443253   .0870492     2.81   0.005     .0737121    .4149385
                 |
SCHULABSCHLUSS_V |
              2  |    1.266193   .1082392    11.70   0.000     1.054048    1.478338
              3  |    2.326537   .1269858    18.32   0.000      2.07765    2.575425
                 |
           NETTO |     .000506   .0000481    10.51   0.000     .0004116    .0006003
-----------------+------------------------------------------------------------
           /cut1 |11a .2999434   .1044076                      .0953082    .5045785
           /cut2 |    2.090593   .1145786                      1.866023    2.315163
------------------------------------------------------------------------------
```

Parameterinterpretation (Odds-Ratios)

```
. ologit SCHULABSCHLUSS i.GESCHL i.SCHULABSCHLUSS_V NETTO, or

Iteration 0:   log likelihood = -2475.4436
Iteration 1:   log likelihood = -2194.8342
Iteration 2:   log likelihood = -2164.7945
Iteration 3:   log likelihood = -2164.6627
Iteration 4:   log likelihood = -2164.6627

Ordered logistic regression                     Number of obs   =       2280
                                                LR chi2(4)      =     621.56
                                                Prob > chi2     =     0.0000
Log likelihood = -2164.6627                     Pseudo R2       =     0.1255

------------------------------------------------------------------------------
  SCHULABSCHLUSS | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
       2.GESCHL |    1.27676   .1111408     2.81   0.005     1.076497    1.514278
                |
SCHULABSCHLUSS_V |
              2  |   3.547321   .3839592    11.70   0.000     2.869242    4.385649
              3  |   10.24241   1.300641    18.32   0.000     7.985677    13.1369
                 |
           NETTO |   1.000506   .0000482    10.51   0.000     1.000412      1.0006
-----------------+------------------------------------------------------------
           /cut1 |   .2999434   .1044076                      .0953082    .5045785
           /cut2 |   2.090593   .1145786                      1.866023    2.315163
------------------------------------------------------------------------------
```

Output in SPSS

Das kumulative Logit-Modell wird in SPSS über den Pfad *Analysieren Regression Ordinal...* durchgeführ t. Für die Parameterinterpretation müssen die Odds Ratios händisch ausgerechnet werden (\(exp(\beta) \)).

Achtung: SPSS hat im Gegensatz zu den anderen Statistikprogrammen bei den Faktorvariablen GESCHL und SCHULABSCHLUSS_V eine andere Referenzkategorie gewählt. Entweder muss dies bei der Interpretation beachtet werden oder die Variablen müssen vorher transformiert werden. Hier wurden GESCHL (3 - GESCHL) und SCHULABSCHLUSS_V (4 - SCHULABSCHLUSS_V) transformiert, um die Ergebnisse analog interpretieren zu können.

Sie erhalten unter anderem diesen Output:

allbus.dta

allbus.RData

allbus.sav

Code

Zusammenfassung der Fallverarbeitung

Gültig 3	2280	100,0%
Fehlend 2	0	
Gesamt 1	2280	

Information zur Modellanpassung

Modell	-2 Log-Likelihood	Chi-Quadrat	Freiheitsgrade	Sig.
Nur konstanter Term	2622,092			
Final	2000,530	621,562	4	,000

Verknüpfungsfunktion: Logit.

Anpassungsgüte

	Chi-Quadrat	Freiheitsgrade	Sig.
Pearson	1305,823	1272	,249
Abweichung	1343,248	1272	,081

Verknüpfungsfunktion: Logit.

6 Pseudo R-Quadrat

Cox und Snell	,239
Nagelkerke	,269
McFadden	,126

Verknüpfungsfunktion: Logit.

Parameterschätzer

			Schätzer	Standard Fehler 12	Wald 13	Freiheitsgrade	Sig. 14	15 Konfidenzintervall 95% Untergrenze	Obergrenze
Schwelle	[SCHULABSCHLUSS = 1]	11a	,300	,104	8,297	1	,004	,096	,504
	[SCHULABSCHLUSS = 2]		2,091	,114	335,493	1	,000	1,867	2,314
Lage 10	NETTO	11b	,001	4,792E-5	111,497	1	,000	,000	,001
	[GESCHL=1]		,244	,087	7,913	1	,005	,074	,415
	[GESCHL=2]		0[a]	.	.	0	.	.	.
	[SCHULABSCHLUSS_V=1]		2,327	,129	326,788	1	,000	2,074	2,579
	[SCHULABSCHLUSS_V=2]		1,266	,110	133,590	1	,000	1,051	1,481
	[SCHULABSCHLUSS_V=3]		0[a]	.	.	0	.	.	.

Verknüpfungsfunktion: Logit.

a. Dieser Parameter wird auf Null gesetzt, weil er redundant ist.

Output in SAS

Sie erhalten unter anderem diesen Output:

Anzahl gelesener Beobachtun	2280	1
Anzahl verwendeter Beobacht	2280	3

Modellanpassungstatistiken		
Kriterium	Nur Konstante	Konstanten und Kovariate
AIC	4954.887	4341.325
SC	4966.351	4375.717
-2 LOG L	4950.887	4329.325

Test - Globale Nullhypothese: BETA=0			
Test	Chi-Quadrat	DF	Pr > ChiSq
Likelihood-Ratio	621.5619	4	<.0001
Score	488.1671	4	<.0001
Wald	481.8122	4	<.0001

Code für R.R

Code für Stata.do

Code für SAS.sas

Analyse Maximum-Likelihood-Schätzer

Parameter		DF	Schätzwert (11a / 11b)	Standard-fehler (12)	Waldsches Chi-Quadrat (13)	Pr > ChiSq (14)
Intercept	3	1	-2.3348	0.1785	171.1317	<.0001
Intercept	2	1	-0.5442	0.1711	10.1216	0.0015
GESCHL (10)		1	0.2443	0.0869	7.9109	0.0049
SCHULABSCHLUSS_V	2	1	1.2662	0.1095	133.5931	<.0001
SCHULABSCHLUSS_V	3	1	2.3266	0.1287	326.7921	<.0001
NETTO		1	0.000506	0.000048	111.4804	<.0001

Odds-Ratio-Schätzer

Effekt	Punktschätzer	95% Wald Konfidenzgrenzen (15)	
GESCHL	1.277	1.077	1.514
SCHULABSCHLUSS_V 2 vs 1	3.547	2.862	4.397
SCHULABSCHLUSS_V 3 vs 1	10.243	7.959	13.181
NETTO	1.001	1.000	1.001

Anmerkung: SAS gibt statt den Schwellenwerten Intercepts aus, diese sind wie die Schwellenwerte nur mit negativem Vorzeichen.