10 Kategorielle Zielgrössen

10.1 Multinomiale Zielgrössen

a In der logistischen Regression war die Zielgrösse zweiwertig. Im Beispiel der Umweltumfrage (8.2.d) hatte die Zielgrösse „Beeinträchtigung" eigentlich vier mögliche Werte, die wir für das dortige Modell zu zwei Werten zusammengefasst haben. Die vier Werte zeigen eine Ordnung von „gar nicht" bis „stark". In der gleichen Umfrage wurde auch eine weitere Frage gestellt: „Wer trägt im Umweltschutz die Hauptverantwortung? – Einzelne, der Staat oder beide?". Diese drei Auswahlantworten haben keine eindeutige Ordnung, denn vielleicht nehmen jene, die mit „beide" antworten, den Umweltschutz besonders ernst, und deshalb liegt diese Antwort nicht unbedingt zwischen den beiden anderen.

Hier soll zunächst ein **Modell für eine ungeordnete, kategorielle Zielgrösse** behandelt werden. Im nächsten Abschnitt wird der Fall einer geordneten Zielgrösse untersucht.

b **Modell.** Für eine einzelne Beobachtung bildet das Modell eine einfache Erweiterung des Falles der zweiwertigen Zielgrösse. Wir müssen festlegen, wie die Wahrscheinlichkeiten $P\langle Y_i = k\rangle$ der möglichen Werte k von den Werten $\underline{x}_i$ der Regressoren abhängen.

Die möglichen Werte der Zielgrösse wollen wir mit 0 beginnend durchnummerieren, damit die zweiwertige Zielgrösse ein Spezialfall der allgemeineren Formulierung wird. Zunächst zeichnen wir eine Kategorie als „**Referenzkategorie**" aus. Wir wollen annehmen, dass es die Kategorie $k=0$ sei.

Eine einfache Erweiterung des logistischen Modells besteht nun darin, dass wir für jedes $k \geq 1$ für das logarithmierte Wettverhältnis gegenüber der Referenzkategorie ein separates lineares Modell ansetzen,

$$\log\left\langle \frac{P\langle Y_i = k\rangle}{P\langle Y_i = 0\rangle}\right\rangle = \log\left\langle \frac{\pi_i^{(k)}}{\pi_i^{(0)}}\right\rangle = \eta_i^{(k)} = \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(j)} \qquad k = 1, 2, ..., k^* .$$

Zunächst scheint es, dass je nach Wahl der Referenzkategorie ein anderes Modell herauskommt. Es zeigt sich aber, dass sich diese Modelle nicht wirklich unterscheiden (ähnlich wie es in der Varianzanalyse keine wesentliche Rolle spielt, welche Kategorie, welches Niveau eines Faktors, im formalen Modell weggelassen wird, um die Lösung eindeutig zu machen).

c* Wählen wir beispielsweise $k=1$ statt $k=0$ als Referenz. Für $k \geq 2$ ergibt sich

$$\begin{aligned}
\log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 1 \mid \underline{x}_i\rangle}\right\rangle &= \log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle}\right\rangle - \log\left\langle \frac{P\langle Y_i = 1 \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle}\right\rangle \\
&= \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(j)} - \beta_0^{(1)} + \sum\nolimits_j \beta_j^{(1)} x_i^{(j)} \\
&= (\beta_0^{(k)} - \beta_0^{(1)}) + \sum\nolimits_j \left(\beta_j^{(k)} - \beta_j^{(1)}\right) x_i^{(j)} .
\end{aligned}$$

Das hat genau die selbe Form wie das Ausgangsmodell, wenn man die Differenzen $(\beta_j^{(k)} - \beta_j^{(1)})$ als neue Koeffizienten $\widetilde{\beta}_j^{(k)}$ einsetzt. Für $k=0$ muss man $\widetilde{\beta}_j^{(0)} = -\beta_j^{(1)}$ setzen.

Version Dezember 2009, © W. Stahel

d* **Gruppierte Daten**. Wie in der logistischen Regression (10.2.n) kann man die Beobachtungen mit gleichen Werten der Eingangsgrössen zusammenfassen und zählen, wie viele von ihnen die verschiedenen Werte k der Zielgrösse zeigen. Es sei wieder m_ℓ die Anzahl der Beobachtungen mit $\underline{x}_i = \widetilde{\underline{x}}_\ell$, und $\widetilde{Y}_\ell^{(k)}$ der *Anteil* dieser Beobachtungen, für die $Y_i = k$ ist. Die *Anzahlen* $m_\ell \cdot \widetilde{Y}_\ell^{(k)}$ folgen dann der multinomialen Verteilung mit den Parametern $\widetilde{\pi}_\ell^{(1)}, ..., \widetilde{\pi}_\ell^{(k^*)}$, die durch das oben angegebene Modell bestimmt sind. Die Wahrscheinlichkeiten sind

$$
\begin{aligned}
P\left\langle \underline{\widetilde{Y}}_\ell = \underline{\widetilde{y}}_\ell \right\rangle &= P\langle m_\ell \widetilde{Y}_0 = m_\ell \widetilde{y}_0,\, m_\ell \widetilde{Y}_1 = m_\ell \widetilde{y}_1,\, \ldots,\, m_\ell \widetilde{Y}_{k^*} = m_\ell \widetilde{y}_{k^*} \rangle \\
&= \frac{m_\ell!}{(m_\ell\, \widetilde{y}_\ell^{(0)})! \cdot \ldots \cdot (m_\ell\, \widetilde{y}_\ell^{(k^*)})!} \,(\widetilde{\pi}_\ell^{(0)})^{m_\ell\, \widetilde{y}_\ell^{(0)}} (\widetilde{\pi}_\ell^{(2)})^{m_\ell\, \widetilde{y}_\ell^{(2)}} \cdot \ldots \cdot (\widetilde{\pi}_\ell^{(k^*)})^{m_\ell\, \widetilde{y}_\ell^{(k^*)}} .
\end{aligned}
$$

Die multinomiale Verteilung bildet eine multivariate Exponentialfamilie. Mit einer geeigneten Link-Funktion versehen, legt die multinomiale Verteilung ein multivariates verallgemeinertes lineares Modell fest. Die kanonische Link-Funktion ist diejenige, die durch das angegebene Modell beschrieben wird.

e Die Tatsache, dass für zusammengefasste Beobachtungen eine multinomiale Verteilung entsteht, erklärt den Namen **multinomiales Logit-Modell** für das oben formulierte Modell. Es ist recht flexibel, denn es erlaubt für jeden möglichen Wert k der Zielgrösse eine eigene Form der Abhängigkeit ihrer Wahrscheinlichkeit von den Regressoren. Ein positiver Koeffizient $\beta_j^{(k)} > 0$ bedeutet für zunehmendes $x^{(j)}$ eine steigende Neigung zur Kategorie k im Verhältnis zur Neigung zur Referenzkategorie 0.
Die Flexibilität bedingt, dass recht viele Parameter zu schätzen sind; die Anzahl ist das Produkt aus k^* und der Anzahl Prädiktoren (plus 1 für die Achsenabschnitte $\beta_0^{(k)}$). Mit kleinen Datensätzen sind diese Parameter schlecht bestimmt.

f **S-Funktionen**. Im Statistik-System R steht im package `nnet` die Funktion `multinom` zur Verfügung, um solche Modelle anzupassen. Für das **Beispiel der Umweltumfrage** zeigt Tabelle 10.1.f ein `summary` des Modells, das die Frage nach der Hauptverantwortung in Abhängigkeit vom Alter und Geschlecht der Befragten beschreibt. Man kann die geschätzten Koeffizienten $\widehat{\beta}_{j\ell}$ und ihre Standardfehler ablesen.

Die Referenzkategorie ist „Einzelne". Der Koeffizient von j = Alter für k = Staat ist $\widehat{\beta}_j^{(k)} = -0.00270$. In 50 Jahren nehmen also die log odds von „Staat":„Einzelne" um $0.0027 \cdot 50 = 0.135$ ab; als odds ratio ergibt sich $\exp\langle -0.135\rangle = 0.874$. Allerdings ist der Koeffizient nicht signifikant, da $\widehat{\beta}_j^{(k)}/\text{standard error}_j^{(k)} = -0.0027/0.0034 = 0.79$ einen klar nicht signifikanten z-Wert ergibt. Zwischen den Geschlechtern besteht ein signifikantes Doppelverhältnis von $\exp\langle -0.244\rangle = 0.78$. Frauen weisen die Verantwortung stärker den Einzelnen anstelle des Staates zu als Männer.

g Ob eine **Eingangsgrösse** einen **Einfluss** auf die Zielgrösse hat, sollte man nicht an den einzelnen Koeffizienten festmachen, da ja k^* Koeffizienten null sein müssen, wenn kein Einfluss da ist. Es muss also ein grösseres mit einem kleineren Modell verglichen werden, und das geschieht wie üblich mit den log-likelihoods oder den Devianzen.

S-Funktionen. Im R-System sieht die Funtion `drop1` für multinomiale Modelle leider keinen Test vor. Man muss mit der Funktion `anova` die einzelnen Modelle vergleichen (oder `drop1` entsprechend ergänzen). Tabelle 10.1.g zeigt die Resultate einer erweiterten Funktion `drop1`, die den Test durchführt, für ein ausführlicheres Modell.

Erstaunlicherweise haben weder die politische Partei, noch das Alter oder die Wohnlage einen signifikanten Einfluss auf die Zuweisung der Hauptverantwortung. Das liegt nicht an einem starken Zusammenhang der Eingangs-Variablen analog zum Kollinearitätspro-

```
Call:
multinom(formula = Hauptv ~ Alter + Schulbildung + Beeintr + Geschlecht,
    data = t.d)

Coefficients:
     (Intercept)    Alter Sch.Lehre Sch.ohne.Abi Sch.Abitur Sch.Studium
Staat      0.599 -0.00270    -0.518       -0.500      -0.66      -0.366
beide     -1.421  0.00262    -0.562       -0.257       0.34       0.220
      Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat       -0.722          -0.719      -0.685      -0.244
beide        0.135           0.106       0.716      -0.179

Std. Errors:
      (Intercept)   Alter Sch.Lehre Sch.ohne.Abi Sch.Abitur Sch.Studium
Staat       0.228 0.00340     0.149        0.174      0.221       0.231
beide       0.349 0.00495     0.234        0.257      0.284       0.307
      Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat        0.123           0.163       0.243       0.107
beide        0.179           0.224       0.271       0.154

Residual Deviance: 3385
AIC: 3425
```

Tabelle 10.1.f: Ergebnisse einer multinomialen Logit-Regression im Beispiel der Umweltumfrage

blem, das in der linearen Regression besprochen wurde, denn auch bei einer schrittweisen Elimination bleiben diese drei Variablen nicht-signifikant.

	Df	AIC	Chisq	p.value
<none>	58	3436	NA	NA
Alter	56	3433	1.35	0.508
Schulbildung	50	3454	34.00	0.000
Beeintr	52	3488	64.34	0.000
Geschlecht	56	3437	5.56	0.062
Ortsgroesse	46	3455	43.10	0.000
Wohnlage	46	3422	9.82	0.632
Partei	44	3418	10.56	0.720

Tabelle 10.1.g: Signifikanzen von einzelnen Termen im Beispiel der Umweltumfrage

h* Wenn man kein geeignetes Programm zur Verfügung hat, kann man die $\beta_j^{(k)}$ für die verschiedenen k getrennt schätzen, indem man k^* logistische Regressionen rechnet, jeweils mit den Daten der Kategorie k und der Referenzkategorie. Das gibt zwar leicht andere Resultate, aber die Unterschiede sind nicht allzu gross, wenn die Referenzkategorie einen genügenden Anteil der Beobachtungen umfasst.

Eine Möglichkeit, die genauen Schätzungen zu erhalten, führt über eine andere Anordnung der Daten, die in 11.2.l besprochen wird.

i Die **Residuen-Devianz** ist wie in der logistischen Regression (8.3.i) sinnvoll bei Daten, die zu Anzahlen zusammengefasst werden können (mit $m_\ell > 3$ oder so). Hier wird die maximale Likelihood erreicht für $\widehat{\pi}_\ell^{(k)} = \widetilde{y}_\ell^{(k)}$ und man erhält

$$D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle = 2(\ell\ell^{(M)} - \ell\ell\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle) = 2\sum\nolimits_{\ell,k} m_\ell \widetilde{y}_\ell^{(k)} \log\left\langle \frac{\widetilde{y}_\ell^{(k)}}{\widehat{\pi}_\ell^{(k)}} \right\rangle .$$

Dies gilt für alle möglichen Links zwischen den Wahrscheinlichkeiten $\underline{\pi}$ und den Koeffizienten $\beta_j^{(k)}$ der linearen Prädiktoren.

j Eine weitere Anwendung des multinomialen Logitmodells ist die **Diskriminanzanalyse mit mehr als 2 Kategorien**. Ähnlich wie beim binären logistischen Modell schätzt man einen Score aus der Modellgleichung für jede Kategorie. Dann ordnet man die Beobachtung derjenigen Kategorie zu, für die der lineare Prädiktor maximal ist.

k* Ein noch allgemeineres Modell erlaubt es, die Eingangs-Variablen von den möglichen Werten der Zielgrösse abhängig zu machen.

$$\log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle} \right\rangle = \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(jk)} .$$

Es werden also jeweils 2 Wahlmöglichkeiten miteinander verglichen. Man erlaubt für jedes Verhältnis eine andere Wirkung der Eingangsgrössen.

Diese Form wird auch „Discrete Choice Models" genannt, da sie bei Studien des Wahlverhaltens von Konsumenten verwendet wird.

Literatur: Agresti (2002), Kap. 9, Fahrmeir and Tutz (2001), Kap. 3.2.

l **Residuen-Analyse.** Was Residuen sein sollen, ist im Zusammenhang mit der multinomialen Regression nicht klar. Zunächst gibt es für jede der logistischen Regressionen, auf denen sie beruht, die entsprechenden Residuen, und diese hängen von von der Referenzkategorie ab. Man könnte also für jedes Paar von Werten der Zielgrösse für jede Beobachtung ein Residuum definieren. Wie diese in geeigneter Form gemeinsam dargestellt werden können, ist dem Autor zurzeit noch zu wenig klar. Hinweise werden gerne entgegen genommen.

10.2 Geordnete Zielgrössen

a Wie früher erwähnt (7.1.a), haben Variable oft einen geordneten Wertebereich. Wie kann man diesen Aspekt ausnützen, wenn eine solche Grösse die Zielgrösse einer Regression ist?

Im **Beispiel der Umweltumfrage** (7.1.c) interessierte uns die Frage nach der Beeinträchtigung mit ihren geordneten Antwortmöglichkeiten von „überhaupt nicht" bis „sehr". Bei der Auswertung mit Kreuztabellen wurde diese Ordnung nicht berücksichtigt. Nun soll sie als Zielgrösse betrachtet und ihr Zusammenhang mit Eingangsgrössen wie Schulbildung, Geschlecht und Alter untersucht werden.

b **Modell.** Zur Beschreibung eines Modells hilft, wie für die binäre Zielgrösse (8.2.j), die Annahme einer **latenten Variablen** Z, aus der sich die Kategorien der Zielgrösse durch Klassieren ergeben. Das frühere Modell wird erweitert, indem man mehrere **Schwellenwerte** α_k festlegt. Die Zielgrösse Y ist =0, wenn Z kleiner ist als die kleinste Schwelle α_1, sie ist =1, wenn Z zwischen α_1 und α_2 liegt, usw. Bei k^* Schwellenwerten nimmt Y die k^*+1 Werte $0, 1, ..., k^*$ an.

In Formeln:

$$\begin{array}{lcll} Y=0 & \Longleftrightarrow & Z<\alpha_1 & \\ Y=k & \Longleftrightarrow & \alpha_k \le Z < \alpha_{k+1} & \quad k=1,...,k^*-1 \\ Y=k^* & \Longleftrightarrow & \alpha_{k^*} \le Z \;. & \end{array}$$

Das bedeutet, dass

$$P\langle Y \ge k\rangle = P\langle Z \ge \alpha_k\rangle \qquad k=1,\ldots,k^* \;.$$

Für die latente Variable Z soll der Einfluss der Eingangsgrössen durch eine multiple lineare Regression gegeben sein, also

$$Z_i = \beta_0 + \sum\nolimits_j x_i^{(j)}\beta_j + E_i \;.$$

Der Fehlerterm in dieser Regression hat einen bestimmten Verteilungstyp F, z. B. eine logistische oder eine Normalverteilung.

Abbildung 10.2.b veranschaulicht diese Vorstellung für eine einzige Eingangs-Variable. Bei mehreren Eingangsgrössen wäre auf der horizontalen Achse, wie üblich, der lineare Prädiktor $\eta_i = \underline{x}_i^T\underline{\beta}$ zu verwenden.

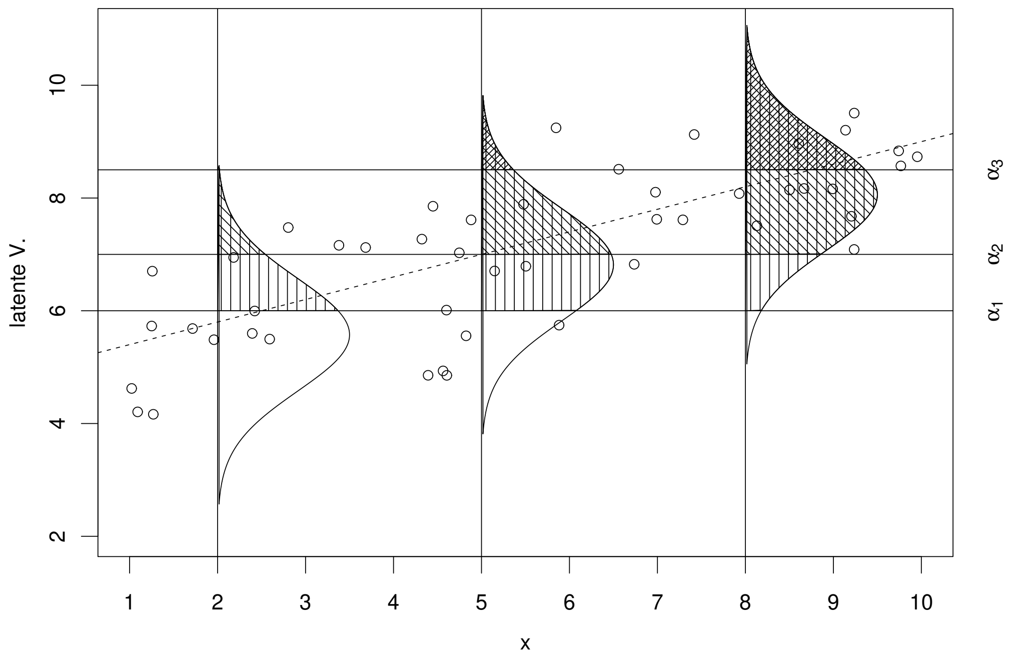

Abbildung 10.2.b: Zum Modell der latenten Variablen

c Wir betrachten die Ereignisse $\{Y_i \geq k\} = \{Z_i \geq \alpha_k\}$ und erhalten für ihre Wahrscheinlichkeiten

$$\begin{aligned}\gamma_k \langle \underline{x}_i \rangle := P\langle Y_i \geq k \rangle &= P\langle Z_i > \alpha_k \rangle = P\left\langle E_i > \alpha_k - \beta_0 - \sum\nolimits_j \beta_j x_i^{(j)} \right\rangle \\ &= 1 - F\left\langle \alpha_k - \left(\beta_0 + \sum\nolimits_j \beta_j x_i^{(j)}\right)\right\rangle ,\end{aligned}$$

wobei F die kumulative Verteilungsfunktion der Zufallsabweichungen E_i bezeichnet.

d Man sieht leicht, dass β_0 unbestimmt ist, da wir zu jedem Schwellenwert α_k eine Konstante hinzuzählen und diese von β_0 abzählen können, ohne dass sich die Y_i ändern. Wir setzen daher $\beta_0 = 0$. – Die Streuung der latenten Variablen ist ebenfalls nicht bestimmt. Wir können Z und alle Schwellenwerte mit einer Konstanten multiplizieren, ohne Y_i zu ändern. Für die kumulative Verteilungsfunktion F der Zufallsfehler kann man daher eine feste Verteilung, ohne den in der multiplen Regression üblichen Streuungsparameter σ, annehmen.

Wenn wir jetzt, wie bei der Regression mit binärer Zielgrösse, $1 - F\langle -\eta \rangle =: g^{-1}\langle \eta \rangle$ setzen, wird

$$g\langle \gamma_k \langle \underline{x}_i \rangle \rangle = \sum\nolimits_j \beta_j x_i^{(j)} - \alpha_k$$

Für jeden Schwellenwert α_k ergibt sich also ein Regressions-Modell mit der binären Zielgrösse, die 1 ist, wenn $Y \geq k$ ist. Diese Modelle sind miteinander verknüpft, da für alle die gleichen Koeffizienten β_j der Regressoren vorausgesetzt werden.

Die üblichste Wahl der Link-Funktion ist wieder die Logit-Funktion. Man spricht dann vom Modell der **kumulativen Logits**. Die inverse Link-Funktion g^{-1} ist dann die logistische Funktion, und die Verteilung der $-E_i$ ist damit die logistische Verteilung.

e Die **Schwellenwerte** α_k müssen nicht etwa gleich-abständig sein. Sie sind unbekannt, und man wird versuchen, sie gleichzeitig mit den Haupt-Parametern β_j zu schätzen. In der Regel sind sie Hilfsparameter, die nicht weiter interessieren.

f Der Name **kumulatives Modell** bezeichnet die Tatsache, dass das Modell die Wahrscheinlichkeiten $P\langle Y \geq k \rangle$, also für die „von oben her kumulierten" Wahrscheinlichkeiten der möglichen Werte k von Y, festlegt.

In Büchern und Programmen wird üblicherweise umgekehrt ein Modell für die „von unten her kumulierten" Wahrscheinlichkeiten formuliert. Das hat den Nachteil, dass diese Wahrscheinlichkeiten mit zunehmendem $\underline{x}^T \underline{\beta}$ abnehmen, so dass positive Koeffizienten β_j einen negativen Zusammenhang der betreffenden Eingangs-Variablen mit der Zielgrösse bedeuten. Wenn so vorgegangen wird, wie wir es hier getan haben, dann bedeutet dagegen ein positiver Koeffizient β_j, dass eine Zunahme von $x^{(j)}$ zu einer Zunahme von Y (oder der latenten Variablen Z) führt. Zudem wird der Fall der Regression mit einer binären Zielgrösse, insbesondere die logistische Regression, ein Spezialfall des neuen Modells, nämlich der Fall von $k^* = 1$.

g Die Wahrscheinlichkeiten für die einzelnen Kategorien erhält man aus sukzessiven Differenzen,

$$P\langle Y_i = k \rangle = \gamma_k \langle \underline{x}_i \rangle - \gamma_{k+1} \langle \underline{x}_i \rangle$$

h Bei einer logistischen Verteilung hat man den Vorteil, dass das Ergebnis mit Hilfe der **Wettverhältnisse** (odds) interpretiert werden kann. Dazu wird jeweils das Wettverhältnis bezüglich eines Schwellenwerts gebildet („cumulative odds“): Wahrscheinlichkeit für niedrigere Kategorien vs. Wahrscheinlichkeit für höhere Kategorien

$$\text{odds}\langle Y_i \geq k \mid \underline{x}_i\rangle = \frac{P\langle Y_i \geq k\rangle}{P\langle Y_i < k\rangle} = \frac{\gamma_k}{1-\gamma_k} = \exp\langle -\alpha_k\rangle \cdot \exp\langle \beta_1\rangle^{x^{(1)}} \cdot \ldots \cdot \exp\langle \beta_m\rangle^{x^{(m)}} .$$

Die Eingangsgrössen wirken auf alle Unterteilungen $Y_i < k$ vs. $Y_i \geq k$ gleich. Die einzelnen Regressoren wirken multiplikativ auf die Wettverhältnisse. Ein solches Modell heisst deshalb Modell der proportionalen Verhältnisse, **proportional-odds model**.

Die Formel vereinfacht sich noch, wenn man die logarithmierten **Doppelverhältnisse** (log odds ratios) für verschiedene Werte $\underline{x}_i$ der Regressoren betrachtet,

$$\log\left\langle \frac{\text{odds}\langle Y_1 \geq k \mid \underline{x}_1\rangle}{\text{odds}\langle Y_2 \geq k \mid \underline{x}_2\rangle}\right\rangle = \beta_1 \cdot (x_1^{(1)} - x_2^{(1)}) + \cdots + \beta_m \cdot (x_1^{(m)} - x_2^{(m)}) .$$

In dieser Gleichung kommt α_k nicht vor. Die Doppelverhältnisse sind also für alle Kategorien k der Zielgrösse gleich!

Wenn $x^{(j)}$ nur eine Indikatorvariable ist, die Behandlung B_0 von Behandlung B_1 unterscheidet, so ist der Koeffizient $\beta^{(j)}$ ein Mass für den Behandlungs-Effekt („unit risk“), der gemäss dem Modell für alle Schwellenwerte gleich ist.

i* Für die logistische Regression wurden neben der Verwendung der logit-Funktion als **Link** noch zwei weitere vorgestellt. Zunächst wurde erwähnt, dass die Annahme einer Normalverteilung für die latente Variable zur Probit-Funktion führt, dass aber die Unterschiede höchstens in riesigen Datensätzen spürbar werden könnten; die beiden Verteilungen unterscheiden sich nur in den Schwänzen, und diese werden mit den hier betrachteten Beobachtungen nur ungenau erfasst. Die Verwendung der Probit-Funktion hat den Nachteil, dass die Interpretation der Koeffizienten über ihre Veränderung der log odds nicht mehr (genau) gilt.

j* Die dritte gebräuchliche Link-Funktion war die „**komplementäre Log-Log-Funktion**“

$$g\langle\mu\rangle = \log\Big\langle -\log\langle 1-\mu\rangle\Big\rangle , \quad 0 < \mu < 1$$

Die entsprechende inverse Link-Funktion ist $g^{-1}(\eta) = 1 - \exp\langle -\exp\langle\eta\rangle\rangle$, und das ist die Verteilungsfunktion der Gumbel-Verteilung.

Für Überlebens- oder Ausfallzeiten bewährt sich die Weibull-Verteilung. Logarithmiert man solche Variable, dann erhält man die Gumbel-Verteilung. Hinter einer Gumbel-verteilten Zielgrösse mit additiven Wirkungen der Regressoren steht oft die Vorstellung einer Weibull-verteilten Grösse und multiplikativen Wirkungen.

k* In der Literatur gibt es neben dem kumulativen Logit-Modell für geordnete Zielgrössen auch das Modell, das für aufeinanderfolgende Kategorien proportionale Wettverhältnisse postuliert. Clogg and Shihadeh (1994) zeigt, dass die Normalverteilung der latenten Variablen dieses Modell der **adjacent classes logits** näherungsweise rechtfertigt.

l **S-Funktionen**. Im R findet man die Funktion `polr`, was für „Proportional Odds Logistic Regression“ steht. Das `summary` (Tabelle 10.2.l (i)) liefert, wie üblich, die Tabelle der Koeffizienten mit Werten der t-Statistik für die Tests $\beta_j = 0$, die für Faktoren mit mehr als 2 Werten wenig Sinn machen. (Die P-Werte werden nicht mitgeliefert; man muss sie selbst ausrechnen.)

Wie in früheren Modellen zeigt die Funktion `drop1(t.r,test="Chisq")` die Signifikanz der Faktoren (Tabelle 10.2.l (ii)).

```
Call: polr(formula = Beeintr ~ Alter + Schule + Geschlecht
           + Ortsgroesse, data = t.d)
Coefficients:
                           Value Std. Error t value p.value
Alter                   -0.00268    0.00299 -0.8992   0.369
SchuleLehre              0.08594    0.13937  0.6166   0.538
Schuleohne.Abi           0.63084    0.15546  4.0578   0.000
SchuleAbitur             0.81874    0.18502  4.4251   0.000
SchuleStudium            1.07522    0.19596  5.4869   0.000
Geschlechtw              0.00699    0.09110  0.0768   0.939
Ortsgroesse2000-4999     0.57879    0.27104  2.1354   0.033
Ortsgroesse5000-19999    0.58225    0.23455  2.4825   0.013
Ortsgroesse20000-49999   0.85579    0.27155  3.1515   0.002
Ortsgroesse50000-99999   0.60140    0.29400  2.0456   0.041
Ortsgroesse100000-499999 0.87548    0.23167  3.7790   0.000
Ortsgroesse>500000       1.10828    0.21568  5.1386   0.000

Intercepts:
               Value  Std. Error t value
nicht|etwas     0.995  0.273      3.644
etwas|ziemlich  2.503  0.278      9.007
ziemlich|sehr   3.936  0.290     13.592

Residual Deviance: 4114.67
AIC: 4144.67
```

Tabelle 10.2.l (i): Resultate für die Regression der geordneten Zielgrösse Beeinträchtigung auf mehrere Eingangsgrössen im Beispiel der Umweltumfrage

```
Model:
Beeintr ~ Alter + Schule + Geschlecht + Ortsgroesse
            Df  AIC    LRT Pr(Chi)
<none>         4145
Alter        1 4143      1    0.369
Schule       4 4196     59    0.000 ***
Geschlecht   1 4143 0.0059    0.939
Ortsgroesse  6 4174     42    0.000 ***
```

Tabelle 10.2.l (ii): Signifikanz der einzelnen Terme im Beispiel

Achtung! Eine kleine Simulationsstudie mit 500 Beobachtungen und 2-3 Variablen (davon ein 3-4-stufiger Faktor) und einer Zielgrösse mit 3 Werten hat alarmierende Resultate gebracht: Die ausgewiesenen Standardfehler waren um einen Faktor von 2 bis 3 zu klein. Die Resultate von `polr` stimmten zudem schlecht mit einer alternativen Berechnungsmethode überein, die gleich geschildert wird. Die Resultate sind also mit äusserster Vorsicht zu geniessen. Es ist bis auf Weiteres angezeigt, die Bootstrap-Methode zu benützen, um die Unsicherheiten zu erfassen. Für Vorhersagen der richtigen Klasse sind die Methoden vermutlich zuverlässiger.

Die Resultate für das **Beispiel der Umweltumfrage** zeigen auch hier, dass Schulbildung und Ortsgrösse einen klaren Einfluss auf die Beurteilung der Beeinträchtigung haben, während Alter und Geschlecht keinen Einfluss zeigen. (Die P-Werte für die beiden letzteren konnten schon in der ersten Tabelle abgelesen werden, da beide nur einen Freiheitsgrad haben.)

m* Man kann das Modell auch **mit Hilfe einer Funktion für die logistische Regression** anpassen. Dazu muss man allerdings die Daten speziell arrangieren. Aus jeder Beobachtung Y_i machen wir k^* Beobachtungen Y^*_{ik} nach der Regel

$$\widetilde{Y}^*_{ik} = \begin{cases} 1 & \text{falls} \quad Y_i \geq k \\ 0 & \text{falls} \quad Y_i < k \end{cases}$$

oder, tabellarisch,

	Y^*_{i1}	Y^*_{i2}	Y^*_{i3}
$Y_i = 0$	0	0	0
1	1	0	0
2	1	1	0
3	1	1	1

Gleichzeitig führt man als Eingangs-Variable einen Faktor $X^{(Y)}$ ein, dessen geschätzte Haupteffekte die Schwellenwerte α_k sein werden. Die neue Datenmatrix besteht jetzt aus n Gruppen von k^* Zeilen. Die k-te Zeile der Gruppe i enthält Y^*_{ik} als Wert der Zielgrösse, k als Wert von $X^{(Y)}$ und die $x_i^{(j)}$ als Werte der anderen Regressoren. Mit diesen $n \cdot k^*$ „Beobachtungen" führt man nun eine logistische Regression durch.

n* Wie bei der binären und der multinomialen Regression kann man Beobachtungen mit gleichen Werten $\underline{x}_i$ der Regressoren zusammenfassen. Die Zielgrössen sind dann

$$\widetilde{Y}_\ell^{(k)} = \text{Anzahl}\{i \mid Y_i = k \text{ und } \underline{x}_i = \underline{\widetilde{x}}_\ell\} / m_\ell \ ,$$

also die Anteile der Personen mit Regressor-Werten $\underline{\widetilde{x}}_\ell$, die die $k$te Antwort geben.

Die Funktion `polr` erlaubt die Eingabe der Daten in aggregierter Form mittels dem Argument `weights`.

o Im Vergleich mit dem **multinomialen Logit-Modell** muss man im kumulativen Logit-Modell deutlich weniger Parameter schätzen: Anstelle von $k^* \cdot p$ sind es hier $k^* + p$. Deswegen wird man bei ordinalen Kategorien das kumulative Modell vorziehen. Wenn die Annahme der gleichen Steigungen verletzt ist, ist es jedoch sinnvoll, auch ordinale Daten mit einem multinomialen Regressions-Modell auszuwerten. Diese Überlegung zeigt auch, wie man diese Annahme überprüfen kann: Man passt ein multinomiales Logit-Modell an und prüft mit einem Modellvergleichs-Test, ob die Anpassung signifikant besser ist.

(Wenn man es genau nimmt, sind die beiden Modelle allerdings nicht geschachtelt, weshalb die Voraussetzungen für den Test nicht exakt erfüllt sind.)

p **Residuen-Analyse**. Wie für die ungeordneten Zielgrössen sind dem Autor keine dem Modell angepassten Definitionen für Residuen bekannt. Eine sinnvolle Definition erscheint mir die Differenz zwischen dem bedingten Erwartungswert der latenten Variablen Z, gegeben die beobachtete Kategorie und der lineare Prädiktor, und dem Wert des linearen Prädiktors,

$$R_i = \mathcal{E}\langle Z \mid Y_i, \widehat{\eta}_i\rangle - \widehat{\eta}_i \ .$$

Die entsprechende S-Funktion ist im Package `regr0` eingebaut.

10.S S-Funktionen

a **Funktion `polr`.** Die S-Funktion `polr` (proportional odds linear regression) aus dem Package `MASS` passt Modelle mit geordneter Zielgrösse an.

```
> t.r <- polr(y~x1+x2+..., data=t.d, weigths, ...)
```

Die linke Seite der Formel, `y`, muss ein Faktor sein. Die Niveaus werden in der Reihenfolge geordnet, wie sie unter `levels(t.y)` erscheinen. Damit man keine Überraschungen erlebt, sollte man einen Faktor vom Typ `ordered` verwenden.

```
> t.y <- ordered(t.d$groups, levels=c("low","medium","high"))
```

Gruppierte Daten können nicht als Matrix eingegeben werden. (Man muss die Anzahlen untereinander schreiben und als `weights` angeben. ...)

b **Funktion `multinom`.** Für multinomiale Regression gibt es die Funktion `multinom`. Sie ist im Package `nnet` versorgt, weil die Berechnung Methoden braucht, die auch für „neural networks“ Anwendung finden. Die linke Seite der Formel kann ein Faktor sein oder für gruppierte Daten, analog zur logistischen Regression, eine Matrix mit k^* Spalten, in denen die Anzahlen mit $Y_i = k$ stehen.

c **Funktion `regr`.** Beide Funktionen sind auch über die Funktion `regr` des Packages `regr0` verfügbar und werden automatisch gewählt, wenn die Zielgrösse ein `ordered`-Faktor resp. ein gewöhnlicher `factor` ist.

Im ersten Fall wird mit `plot` eine spezielle Residuen-Darstellung gewählt, s. oben und Dokumentation zu `regr0`.