Kategorielle Zielgrössen

5.12.2011

Evaluation Vorlesung Teil 2

Bitte (semi-offizielle) Umfrage unter

`http://metaphor2.ethz.ch/eval/hs11/stat_regr/`

ausfüllen.

Einführung

Bis jetzt haben wir diverse Arten von Zielgrössen Y betrachtet:

- Kontinuierliche (Regression)
- Binäre bzw. binomiale (Logistische Regression)
- Zähldaten (Poisson Regression)

Heute betrachten wir noch den Fall von **kategoriellen Zielgrössen** (Faktoren!).

Wir unterscheiden zwischen **ordinalen** (geordneten) und **nominalen** Zielgrössen.

Wir beginnen mit den nominalen Zielgrössen.

Multinomiale Regression

Beispiel Umweltumfrage

Siehe Vorlesung über kategorielle Variablen.

Wir nehmen als Zielgrösse die Hauptverantwortung für den Umweltschutz: bei den Einzelnen / beim Staat / bei beiden.

Die erklärenden Variablen sind

- Alter
- Schulbildung
- Beeinträchtigung
- Geschlecht
- ...

Wie können wir hier die **W'keiten der Zielgrösse** modellieren als Funktion der erklärenden Variablen?

Multinomiale Regression: Modell

Der Einfachheit halber nummerieren wir die Kategorien der Zielgrösse mit $k = 0, 1, 2, \ldots, K$.

Wir fixieren nun eine **Referenzklasse**, z.B. Kategorie 0.

Wir wählen ein **multinomiales Logit-Modell**, d.h.

$$\log\left(\frac{P(Y_i = k \,|\, \underline{x}_i)}{P(Y_i = 0 \,|\, \underline{x}_i)}\right) = \log\left(\frac{\pi_i^{(k)}}{\pi_i^{(0)}}\right) = \eta_i^{(k)} = \beta_0^{(k)} + \sum_{j=1}^{m} \beta_j^{(k)} x_i^{(j)}$$

für $k = 1, \ldots, K$.

Wie bei der logistischen Regression modellieren wir die **logarithmierten Wettverhältnisse**. Hier haben wir aber **mehrere** davon.

Zusätzlich sollten sich die Wahrscheinlichkeiten jeweils zu 1 addieren: $\sum_{k=0}^{K} \pi_i^{(k)} = 1$.

Da wir für jede Kategorie ein "separates" Modell aufsetzen, haben wir eine **grosse Menge von Parametern**, nämlich $K \cdot (m+1)$!

Man kann zeigen, dass wir eigentlich die W'keiten modellieren als

$$\begin{aligned} \pi_i^{(0)} &= 1 - \sum_{k=1}^{K} \pi_i^{(k)} \\ \pi_i^{(k)} &= \frac{\exp\left\{\eta_i^{(k)}\right\}}{1 + \sum_{l=1}^{K} \exp\left\{\eta_i^{(l)}\right\}}, \; k \geq 1. \end{aligned}$$

Oder anders ausgedrückt:

Das Modell liefert uns die W'keiten, in die einzelnen Klassen zu fallen (in Abhängigkeit von den erklärenden Variablen).

Wenn man sich gruppierte Daten vorstellt, so folgen die Anzahlen also gerade einer **multinomialen Verteilung** mit den obigen W'keiten.

Am Anfang stand eine "willkürliche" Entscheidung: Die Wahl der Referenzklasse.

Man kann wieder zeigen, dass sich das Modell nicht ändert, wenn man die Referenzklasse ändert.

Oder anders gesagt:

Für eine andere Referenzklasse kann man die Parameter "eins zu eins" umrechnen (ähnlich wie bei Referenzlevel von Faktoren).

Das Modell ist recht **flexibel** (viele Parameter!).

Interpretation der Parameter

Für jede Kategorie der Zielgrösse erlaubt es eine **eigene Form der Abhängigkeit** (der W'keit) von den erklärenden Variablen.

Die Interpretation ist wie bei der logistischen Regression.

Ein positiver Koeffizient $\beta_j^{(k)}$ bedeutet eine steigende Neigung zur Kategorie k für zunehmendes $x^{(j)}$ im Verhältnis zur Referenzkategorie 0.

Tests, Standardfehler etc.

Z.B. via Devianzen

Beispiel Umweltumfrage

R: Funktion `multinom` im Package `nnet`

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

Coefficients:
      (Intercept)        Alter SchuleLehre Schuleohne.Abi SchuleAbitur
Staat   0.5986115 -0.002696350  -0.5175446     -0.5003699   -0.6601135
beide  -1.4214438  0.002621571  -0.5623772     -0.2567625    0.3399222
      SchuleStudium Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat    -0.3658393   -0.7220483      -0.7194168  -0.6847353  -0.2439706
beide     0.2204495    0.1354301       0.1057302   0.7156382  -0.1793147

Std. Errors:
      (Intercept)        Alter SchuleLehre Schuleohne.Abi SchuleAbitur
Staat   0.2283652 0.003398536   0.1494304      0.1741211    0.2205693
beide   0.3488268 0.004949062   0.2336575      0.2572836    0.2836734
      SchuleStudium Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat     0.2314224    0.1231731       0.1628122   0.2426448   0.1070307
beide     0.3065064    0.1791422       0.2236203   0.2706047   0.1540745

Residual Deviance: 3384.951
AIC: 3424.951
```

Für den Globaltest muss man leider den Umweg über anova gehen.

```
fit.small <- update(fit, Hauptv ~ . - Schule)
anova(fit, fit.small, test = "Chisq")
```

Das liefert hier

```
Response: Hauptv
                                  Model Resid. df Resid. Dev   Test    Df
1          Alter + Beeintr + Geschlecht      3622   3419.529
2 Alter + Schule + Beeintr + Geschlecht      3614   3384.951 1 vs 2     8

  LR stat.      Pr(Chi)
1
2 34.57792 3.191061e-05
```

Ordinale Regression

Wir betrachten jetzt **geordnete Zielgrössen**.

Beispiel Umweltumfrage

Nehme z.B. die Beeinträchtigung als Zielgrösse. Mögliche Ausprägungen sind “überhaupt nicht” bis “sehr stark”.

Also haben wir einen **ordinalen Faktor** als Zielgrösse.

Ordinale Regression: Modell der latenten Variable

Wie bei der logistischen Regression kann man sich ein Modell mit einer **latenten Variable** Z vorstellen.

Durch **Klassierung** von Z entstehen dann die verschiedenen Kategorien.

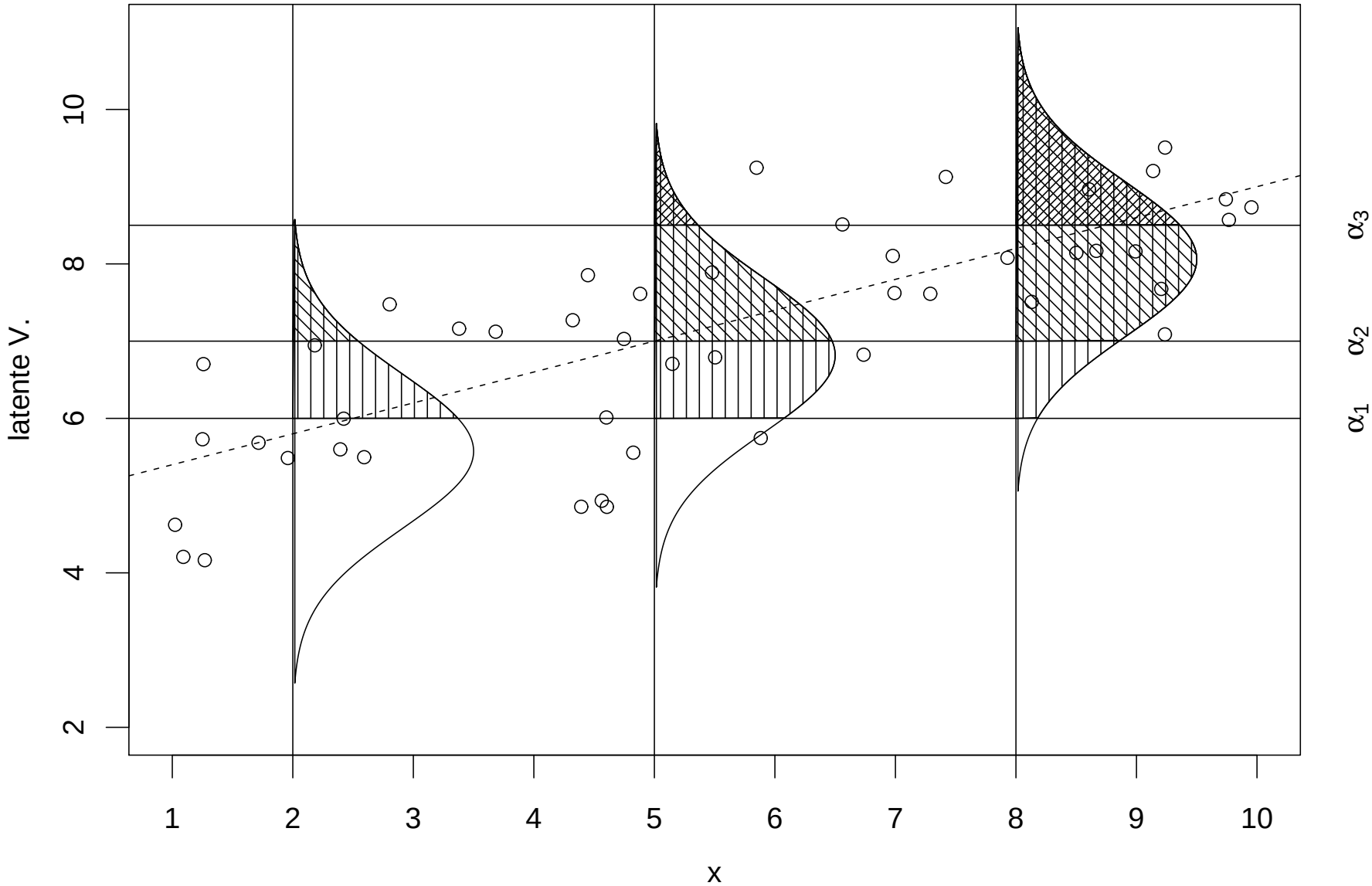
Früher hatten wir nur **einen** Schwellenwert. Jetzt brauchen wir **mehrere** davon (wir haben in der Regel ja mehr als zwei Klassen).

Wir nehmen also an, dass wir eine kontinuierliche **latente Variable** Z und Schwellenwerte α_k haben, sd.

$$\begin{aligned} Y_i = 0 &\iff Z_i \leq \alpha_1 \\ &\dots \\ Y_i = k &\iff \alpha_k < Z_i \leq \alpha_{k+1} \\ &\dots \\ Y_i = K &\iff \alpha_K < Z_i. \end{aligned}$$

Insgesamt haben wir K Schwellenwerte: $\alpha_1 < \alpha_2 < \ldots < \alpha_K$.

Illustration Modell der latenten Variable



Für die latente Variable Z nehmen wir ein gewöhnliches **multiples lineares Regressionsmodell** an, d.h.

$$Z_i = \beta_0 + \sum_{j=1}^{m} x_i^{(j)} \beta_j + E_i,$$

mit einer bestimmten Verteilung für den Fehlerterm.

Betrachten wir nun die (kumulierten) W'keiten, so haben wir

$$\begin{aligned}
\gamma_k &:= P\left(Y_i \geq k \,|\, x_i\right) = P(Z_i > \alpha_k \,|\, x_i) \\
&= P\left(E_i > \alpha_k - (\beta_0 + \underline{x}_i^T \underline{\beta})\right) \\
&= 1 - F_E\left(\alpha_k - (\beta_0 + \underline{x}_i^T \underline{\beta})\right) \\
&= F_{-E}\left(\underline{x}_i^T \underline{\beta} - (\alpha_k - \beta_0)\right).
\end{aligned}$$

Jetzt kann man wieder Annahmen über die Verteilung der Fehler E_i treffen, z.B. logistische Verteilung, Normalverteilung, Extremwertverteilung.

Mit der Linkfunktion $g = F_{-E}^{-1}$ erreichen wir, dass

$$g(\gamma_k) = \underline{x}_i^T \underline{\beta} - (\alpha_k - \beta_0).$$

Wählen wir die logistische Verteilung, so haben wir

$$\log\left(\frac{P(Y_i \geq k \,|\, x_i)}{P(Y_i < k \,|\, x_i)}\right) = \underline{x}_i^T \underline{\beta} - (\alpha_k - \beta_0).$$

Das sieht aus wie ein logistisches Regressionsproblem mit der binären Zielgrösse $\{Y_i \geq k\}$ (ja oder nein).

Es gibt **Identifikationsprobleme**:

Ändert man α_k, so kann man dies mit β_0 kompensieren $\rightsquigarrow$ setze $\beta_0 = 0$.

Multipliziert man Z und alle Schwellenwerte mit einer Konstanten, so ändert sich Y nicht $\rightsquigarrow$ Wir nehmen daher an, dass die Fehlervarianz fix vorgegeben ist.

Was modellieren wir hier eigentlich?

Für jedes Level k haben wir eigentlich ein logistisches Regressionsmodell mit binärer Zielgrösse. Die Zielgrösse ist 1, wenn $Y \geq k$ und sonst 0.

Die Modelle sind aber **miteinander verknüpft**: Die Parameter β_j sind immer die gleichen!

Die Schwellenwerte α_k müssen auch geschätzt werden!

Man spricht von einem **kumulativen Modell**, weil man die W'keiten aufsummiert (von oben her).

Wir können hier auch wieder mit odds arbeiten:

$$\begin{aligned} \text{odds}(Y_i \geq k \,|\, x_i) &= \frac{P(Y_i \geq k)}{P(Y_i < k)} = \frac{\gamma_k}{1-\gamma_k} \\ &= \exp\{-\alpha_k\} \cdot \exp\{\beta_1\}^{x^{(1)}} \cdots \exp\{\beta_m\}^{x^{(m)}}. \end{aligned}$$

Die Interpretation ist wieder analog zur logistischen Regression.

Erhöht man $x^{(j)}$ um eine Einheit, so ändern sich die odds, in die höhere Kategorie zu fallen, um den Faktor $\exp\{\beta_j\}$.

Ein positives β_j bedeutet also, dass man für steigende x-Werte eher in höhere Kategorien fällt.

Oder mit den log odds-ratios

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

Der Einfluss der erklärenden Variablen auf die odds, bzw. die log odds-ratios ist **unabhängig** von k (für alle gleich)!

Das heisst die erklärenden Variablen wirken für alle "Unterteilungen" gleich.

Man spricht daher auch vom **proportional-odds Modell**.

Für die Interpretation der Parameter etc. ist es wohl am einfachsten, das Modell der latenten Variable im Kopf zu haben.

Wenn dort ein Parameter $\beta_j > 0$ ist, so ist klar, dass man für grössere Werte von $x^{(j)}$ eher in höhere Kategorien fällt.

Vergleich mit multinomialer Regression

Verglichen mit dem multinomialen Regressionsmodell haben wir hier **viel weniger Parameter**, nämlich nur $K + m$.

R: Funktion `polr` im Package `MASS`.

Betrachten wir nochmals die Umweltumfrage mit Beeinträchtigung als Zielgrösse.

```
Call:
polr(formula = Beeintr ~ Alter + Schule + Geschlecht + Ortsgroesse,
    data = t.d)

Coefficients:
                            Value Std. Error  t value
Alter                    -0.002684   0.002985 -0.89919
SchuleLehre               0.085937   0.139375  0.61659
Schuleohne.Abi            0.630839   0.155462  4.05784
SchuleAbitur              0.818736   0.185023  4.42506
SchuleStudium             1.075209   0.195963  5.48680
Geschlechtw               0.006994   0.091097  0.07677
Ortsgroesse>500000        1.108286   0.215679  5.13859
Ortsgroesse100000-499999  0.875489   0.231673  3.77898
Ortsgroesse2000-4999      0.578782   0.271040  2.13541
Ortsgroesse20000-49999    0.855790   0.271553  3.15147
Ortsgroesse5000-19999     0.582261   0.234546  2.48251
Ortsgroesse50000-99999    0.601429   0.293995  2.04571

Intercepts:
                Value   Std. Error t value
nicht|etwas      0.9950  0.2730      3.6441
etwas|ziemlich   2.5030  0.2779      9.0069
ziemlich|sehr    3.9360  0.2896     13.5923

Residual Deviance: 4114.673
AIC: 4144.673
```

Teste Faktoren mit `drop1(fit, test = "Chisq")`.

```
Model:
Beeintr ~ Alter + Schule + Geschlecht + Ortsgroesse
             Df    AIC    LRT  Pr(>Chi)
<none>          4144.7
Alter         1 4143.5  0.809    0.3683
Schule        4 4196.0 59.368 3.938e-12 ***
Geschlecht    1 4142.7  0.006    0.9388
Ortsgroesse   6 4174.4 41.681 2.125e-07 ***
```