Lösungen zu Blatt 5

Aufgabe 1: Multinomiales Modell

(i) Link-Funktionen:

$$log\left(\frac{P(y=i|\underline{x})}{P(y=cust|\underline{x})}\right) = \beta_0^{(i)} + x_{ed}\beta_1^{(i)} + x_g\beta_2^{(i)} + x_{min}\beta_3^{(i)} \text{ mit i} \in \{admin, manage\}$$

Response-Funktionen:

$$P(y=i) = \frac{exp(\underline{x}^T\underline{\beta}^{(i)})}{1 + exp(\underline{x}^T\underline{\beta}^{(admin)}) + exp(\underline{x}^T\underline{\beta}^{(manage)})} \text{ mit i} \in \{admin, manage\}$$

$\Longrightarrow$ entsprechend der Form zweier (eines für admin, eines für manage) binärer Logit-Modelle und daher analoge Interpretation.

Unterschiede zur binären Logit-Modellierung:

- $P(y=i|\underline{x}) + P(y=cust|\underline{x}) \overset{i.A.}{\neq} 1$, also nicht komplementär
  [aber: $P(y=manage|\underline{x}) + P(y=admin|\underline{x}) + P(y=cust|\underline{x}) = 1$]
- Responsefunktion jeweils abhängig von admin UND manage

(ii) Laden des Datensatzs

```
library("AER")
data("BankWages")
```

Schätzung des Modells

```
library("nnet")
multinomial <- multinom(job ~ education + minority + gender,
                        data = BankWages)
```

```
summary(multinomial)
```

```
## Call:
## multinom(formula = job ~ education + minority + gender, data = BankWages)
##
## Coefficients:
##        (Intercept) education minorityyes genderfemale
## admin       -4.767    0.5541     -0.4267        12.54
## manage     -32.809    2.3183     -2.7450        11.63
##
## Std. Errors:
##        (Intercept) education minorityyes genderfemale
## admin        1.173   0.09908      0.5027       0.2233
## manage       4.448   0.29294      0.9353       0.2233
##
## Residual Deviance: 276.6
## AIC: 292.6
```

(iii) Relevante Größen für die Interpretation

```
exp(coef(multinomial))
```

```
##        (Intercept) education minorityyes genderfemale
## admin    8.503e-03      1.74     0.65264       280380
## manage   5.638e-15     10.16     0.06425       112176
```

aus (i) folgt:

$$\frac{P(y=i|\underline{x})}{P(y=cust|\underline{x})} = exp(\beta_0^{(i)}+x_{ed}\beta_1^{(i)}+x_g\beta_2^{(i)}+x_{min}\beta_3^{(i)}) = exp(\beta_0^{(i)})\cdot exp(x_{ed}\beta_1^{(i)})\cdot ...$$

- Interpretationen für die Kategorie `"admin"` im Verhältnis zu `"custodial"`:
  - Steigt die Bildung bei Konstanthaltung aller anderen Variablen um eine Einheit, so erhöht sich die Chance auf einen Verwaltungsjob im Verhältnis zu einer Stelle als Aufsichtsperson um den Faktor 1.74
  - Für eine Frau erhöht sich bei Konstanthaltung aller anderen Variablen, im Vergleich zu einem Mann, die Chance auf einen Verwaltungsjob im Verhältnis zu einer Stelle als Aufsichtsperson um den Faktor 280381
- Gehe für die Kategorie `"manage"` analog vor!
- `education` und `gender` mit höchstsignifikantem Einfluss!

(iv) Dummy-Kodierung für die Variable `job` mit:

- Referenzkategorie: custodial
- Link-Funktion für custodial: $log\left(\frac{P(y=cust|\underline{x})}{P(y=cust|\underline{x})}\right) = log(1) = 0$
- Response-Funktion für custodial: $P\left(y=cust\right) = \frac{1}{1+exp(\underline{x}^T\underline{\beta}^{(admin)})+exp(\underline{x}^T\underline{\beta}^{(manage)})}$
- grobes Prinzip der Dummy-Kodierung: setze für die Referenzkategorie k immer $x_k = 0$ und für die jeweils interessierende Kategorie i $x_i = 1$ (alle übrigen Kategorien sind ebenfalls auf 0 gesetzt)

Aufgabe 2: Proportional-Odds-Modell

(i) Es wurde ein Proportional-Odds-Modell mit der Funktion `polr` gefittet.
Modellformeln:

$$1.\ log\left(\frac{P(y \leq custodial)}{P(y \geq admin)}\right) = \beta_0^{(1)} - \underline{x}^T\underline{\beta}$$

$$2.\ log\left(\frac{P(y \leq admin)}{P(y \geq manage)}\right) = \beta_0^{(2)} - \underline{x}^T\underline{\beta}$$

Eigenschaften:

- Unterschiedliche Intercepts werden geschätzt
- Identische Steigungsparameter werden geschätzt

- Beachte das Minus im linearen Prädiktor!

Schätzung des Modells

```
library("MASS")
proportional <- polr(formula = job ~ education + minority,
                     data = BankWages,subset = gender == "male",
                     Hess = "TRUE")
```

```
summary(proportional)
```

```
## Call:
## polr(formula = job ~ education + minority, data = BankWages,
##     subset = gender == "male", Hess = "TRUE")
##
## Coefficients:
##              Value Std. Error t value
## education     0.87     0.0931    9.35
## minorityyes -1.06     0.4120   -2.56
##
## Intercepts:
##                  Value  Std. Error t value
## custodial|admin  7.951  1.077      7.383
## admin|manage    14.172  1.474      9.612
##
## Residual Deviance: 260.64
## AIC: 268.64
```

(ii) Relevante Größen für die Interpretation

```
exp(proportional$zeta)
```

```
## custodial|admin    admin|manage
##            2839         1428485
```

```
exp(-coef(proportional))
```

```
##   education minorityyes
##       0.419       2.876
```

- Interpretation der Intercept-Terme:
  - Die Chance auf eine Stelle als Aufsichtsperson (oder niedriger) im Verhältnis zu einer Stelle im Verwaltungswesen (oder höher) liegt bei 2839, falls man keiner Minderheit angehört und ein Ausbildungsniveau von 0 aufweist.
  - Die Chance auf eine Stelle im Verwaltungwesen (oder niedriger) im Verhältnis zum Managerposten (oder höher) liegt bei 1428485, falls man keiner Minderheit angehört und ein Ausbildungsniveau von 0 aufweist.
- Interpretation der übrigen Koeffizienten (Steigungsparameter):
  - Die Chance auf einen Job der Kategorie i oder niedriger im Verhältnis zu einer höheren Jobkategorie fällt - mit dem Zuwachs

des Ausbildungsniveaus um eine Einheit - um den Faktor 0.419; dieser Sachverhalt wird als unabhängig von den betrachteten Kategorien angenommen.

  - Die Chance auf einen Job der Kategorie i oder niedriger im Verhältnis zu einer höheren Jobkategorie ist für ein Mitglied einer Minderheit um den Faktor 2.876 höher als für jemanden, der keiner Minderheit angehört; dieser Sachverhalt wird als unabhängig davon betrachtet, um welche Jobkategorien es sich handelt.

(iii)
- Vorteile des Proportional-Odds-Modells:
  - deutlich parametersparsamer
  - den Jobkategorien angepasst, da es sich dabei um ordinale Daten handelt
    $\Rightarrow$ gute Interpretierbarkeit
- Nachteile
  - eventuell unzureichende Anpassung an die Daten aufgrund der kategorienunabhängigen Steigungsparameter
  - eventuell zu parametersparsam