Verallgemeinerte lineare Modelle in R: Logistische Regression

Achim Zeileis

2009-02-20

Um die Analyse der Vorlesung zu reproduzieren, wird zunächst der `BBBClub` Datensatz geladen

```
R> load("BBBClub.rda")
```

Um den Zusammenhang zwischen der Kaufentscheidung `choice` und dem Geschlecht `gender` zu untersuchen, können wir uns wie schon im Tutorium *EDA2.pdf* die entsprechende Kontingenztafel berechnen

```
R> tab <- xtabs(~gender + choice, data = BBBClub)
R> tab
```

```
        choice
gender    no yes
  female 273 183
  male   627 217
```

sowie die zugehörigen bedingten relativen Häufigkeiten

```
R> prop.table(tab, 1)
```

```
        choice
gender          no       yes
  female 0.5986842 0.4013158
  male   0.7428910 0.2571090
```

Der Odds Ratio ergibt sich hier als

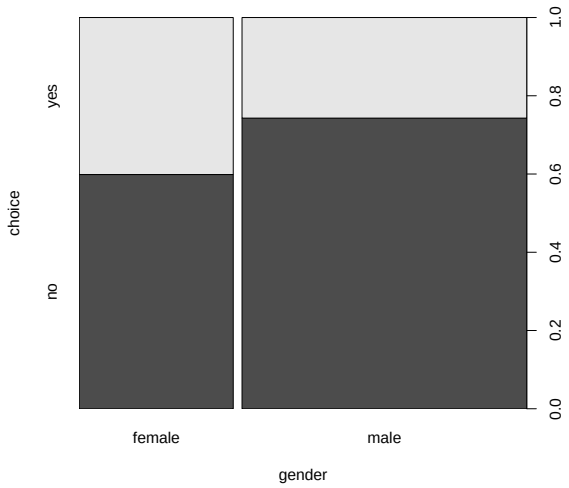
```
R> (273 * 217)/(627 * 183)
```

```
[1] 0.5163019
```

Die Chancen für einen Kauf des Kunstbandes sind bei den Männern nur halb so hoch wie bei den Frauen.

Die Daten können mittels eines Mosaikplots visualisiert werden

```
R> plot(choice ~ gender, data = BBBClub)
```



aus dem ebenfalls abzulesen ist, daß die Kaufwahrscheinlichkeit mit dem Geschlecht sinkt.

Das zugehörige verallgemeinerte lineare Modell (GLM) wird in R mit der Funktion `glm` angepaßt. Dabei spezifiziert man, wie gehabt, eine Formel und die Daten und zusätzlich noch die Verteilungsfamilie. Für letztere wird bei binären Daten die Binomialverteilung verwendet.

```
R> fmG <- glm(choice ~ gender, data = BBBClub, family = binomial)
R> fmG

Call:  glm(formula = choice ~ gender, family = binomial, data = BBBClub)

Coefficients:
(Intercept)   gendermale
     -0.400       -0.661

Degrees of Freedom: 1299 Total (i.e. Null);  1298 Residual
Null Deviance:            1605
Residual Deviance: 1576         AIC: 1580
```

Der Print-Output liefert uns wieder die Koeffizientenschätzer (sowie zusätzliche Information, die in folgenden Tutorien noch genauer behandelt werden wird).

Eine ausführlichere Zusammenfassung inklusive der üblichen t-Tests für jeden Koeffizienten liefert wieder die `summary`-Funktion.

```
R> summary(fmG)

Call:
glm(formula = choice ~ gender, family = binomial, data = BBBClub)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.013  -0.771  -0.771   1.351   1.648

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -0.39999    0.09554  -4.187 2.83e-05 ***
gendermale  -0.66106    0.12382  -5.339 9.34e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1604.8  on 1299  degrees of freedom
Residual deviance: 1576.4  on 1298  degrees of freedom
AIC: 1580.4

Number of Fisher Scoring iterations: 4
```

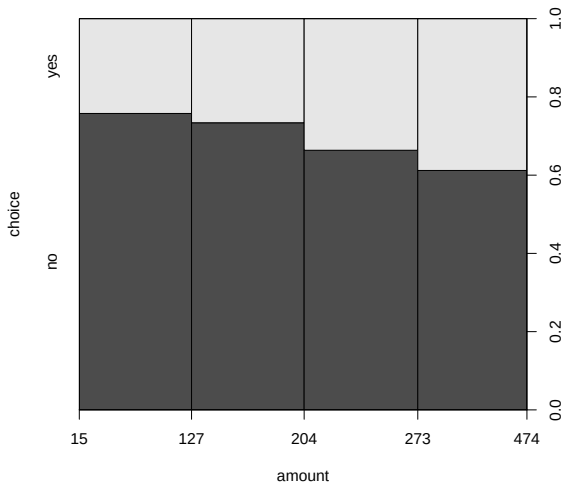
Hieraus ist abzulesen, daß das Geschlecht einen hochsignifikanten Einfluß auf die Kaufwahrscheinlichkeit hat und zwar so, daß diese von Frauen zu Männern sinkt (wegen des negativen Vorzeichens). Genauergesagt, sind die Chancen auf Kauf bei Männern um rund 48% niedriger als bei den Frauen:

```
R> exp(coef(fmG)[2])

gendermale
 0.5163019
```

Völlig analog kann man bei der Analyse mit Hilfe metrischer Erklärungsvariablen vorgehen. Betrachten wir bspw. den Einfluß der Gesamtausgaben auf die Kaufwahrscheinlichkeit, so ist aus dem Mosaikplot abzulesen, daß diese mit den Gesamtausgaben steigt.

```
R> plot(choice ~ amount, data = BBBClub, breaks = fivenum(BBBClub$amount))
```



Das entsprechende GLM wird wieder mit `glm` angepaßt:

```
R> fmA <- glm(choice ~ amount, data = BBBClub, family = binomial)
R> fmA

Call:  glm(formula = choice ~ amount, family = binomial, data = BBBClub)
```

```
Coefficients:
(Intercept)       amount
  -1.453283     0.003109

Degrees of Freedom: 1299 Total (i.e. Null);  1298 Residual
Null Deviance:            1605
Residual Deviance: 1581         AIC: 1585
```

Aus der zugehörigen Zusammenfassung läßt sich ablesen, daß die Gesamtausgaben ebenfalls einen hochsignifikanten Einfluß auf die Kaufwahrscheinlichkeit haben.

```
R> summary(fmA)

Call:
glm(formula = choice ~ amount, family = binomial, data = BBBClub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1847  -0.8839  -0.7734   1.3967   1.8004

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4532828  0.1499533  -9.692  < 2e-16 ***
amount       0.0031088  0.0006477   4.800 1.59e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1604.8  on 1299  degrees of freedom
Residual deviance: 1581.3  on 1298  degrees of freedom
AIC: 1585.3

Number of Fisher Scoring iterations: 4
```

Auch hier läßt sich ein Odds Ratio angeben und zwar für den Fall, daß sich die Gesamtausgaben um eine Einheit ändern (hier also um USD 1).

```
R> exp(coef(fmA)[2])

  amount
1.003114
```

Mit jedem zusätzlich ausgegebenen US Dollar steigen also die Chancen auf Kauf im Durchschnitt um 0.3%. Bei so kleinen Veränderungen in den Ausgaben, sind die Veränderungen der Chancen natürlich sehr klein. Einfacher zu interpretieren ist deshalb bspw.

```
R> exp(100 * coef(fmA)[2])

  amount
1.364619
```

Ein Kunde, der USD 100 mehr ausgegeben hat als ein anderer Kunde, hat also im Schnitt eine um 36% höhere Chance den Kunstband zu kaufen.