Einführung in die statistische Datenanalyse mit R

Logistische Regression

David Benček

Wintersemester 2015/16

Logistische Regression

- auch **logit-Modell** genannt,
- nützlich für Fragestellungen, bei denen die abhängige Variable nur zwei Werte annahmen kann (dichotome Variable).

Beispiele:

- Wahlbeteiligung von Individuen
- Wahlsieg von Kandidaten
- Erfolg einer Studienplatzbewerbung

Demonstration

Beispielhafte Berechnung eines logit-Modells zur Frage, welche Größen sich auf eine erfolgreiche Studienplatzbewerbung auswirken.

```
logit_data <- read.csv("./data/logit_example.csv")

head(logit_data)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

Demonstration II

```
summary(logit_data)
```

```
##      admit             gre             gpa               r
##  Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.
##  1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu
##  Median :0.0000   Median :580.0   Median :3.395   Median
##  Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean
##  3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu
##  Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.
```

Demonstration III

Modellschätzung:

```
logit_data$rank <- factor(logit_data$rank)
logit_model <- glm(admit ~ gre + gpa + rank, data = logit_d
```

Demonstration IV

```
summary(logit_model)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = logit_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gre          0.002264   0.001094   2.070 0.038465 *
## gpa          0.804038   0.331819   2.423 0.015388 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

Exkurs: Odds-Ratio

	erkrankt	nicht.erkrankt
mit Risikofaktor	65	30
ohne Risikofaktor	20	75

$$\text{OR} = \frac{65/30}{20/75} = 8.125$$

Personen mit Risikofaktor haben eine über 8-mal höhere Chance zu erkranken.

Interpretation

- Bei einer Änderung von `gre` um eine Einheit steigt die log-odds der Zulassung um 0.002 an.
- Ein Anstieg des `gpa` um eine Einheit erhöht die log-odds der Zulassung um 0.804.
- Kategoriale Variablen sind relativ zu ihrer Basiskategorie zu interpretieren:
  - Ist die vorherige Hochschule dem Rang 2 zuzuordnen, ändert dies die log-odds der Zulassung um -0.675 im Vergleich zu einer identischen Bewerbung von einer Hochschule vom Rang 1.