

# 1 Odds and odds-ratio

## 1.1 Odds

$$\frac{p}{1-p}$$

Zahlen Everitt: 278 Vpn, 68 Fälle, 210 gesund, 174 Frauen, 104 Männer

$$p(\text{case}) = \frac{68}{278} = 0.2446043 \frac{p}{1-p} = \frac{0.2446043}{1-0.2446043} = 0.3238095$$

## 2 Logistische Funktion

$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Umformung durch Multiplikation mit  $e^{-x}$

$$f(x) = \frac{1}{1+e^{-(x+h)*s}}$$

Lage des Wendepunktes (und damit Verschiebung der Kurve in der Waagerechten) läßt sich über Parameter h beeinflussen, Steilheit der Kurve über Parameter s. Beide sind Konstanten so dass x die einzige Variable bleibt.

## 3 Generalisierte Lineare Modelle

### 3.1 Nomenklatur

ALM (allgemeine lineare Modelle = GLM (general linear models))

GLM (generalisierte lineare Modelle) = VLM (verallgemeinerte lineare Modelle) = GLZ (generalized linear models) = GzLM (in SPSS) = GLiM (bei Zucchini) [manchmal leider auch im Englischen als GLM bezeichnet]

### 3.2 Grundidee

Verallgemeinerung klassischer linearer Modelle.

Erwartungswert der Response-Variable (Mittelwert) gegeben die Erklärungsvariablen (response variables)

$$E(y|x_1, x_2, \dots, x_q) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

Bedingung:  $y$  ist stetige Variable und normalverteilt.

Wenn Bedingung nicht erfüllt: Obiges Modell passt nicht. Lösung: Generalisiertes Lineares Modell (GLiM). Ansatz: Eine Link-Funktion vermittelt zwischen Erklärungsvariablen und der Responsevariable.

Allgemein

$$g(E(y|x_1, x_2, \dots, x_q)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q = \beta_0 + \sum_{k=1}^q \beta_k x_k$$

GLZ sind verwendbar, wenn Responsevariable eine andere als Normalverteilung vorliegt (Verteilungen aus der 'Exponentialfamilie' wie Poisson, Gamma, ...)

Die sog. Linkfunktion verknüpft die Responsevariable mit der Linearkombination der Erklärungsvariablen.

Die Parameter stehen ausschließlich in linearer Form in den Modellgleichungen.  $\exp(\beta_2)$  oder  $\alpha\beta$  sind nicht möglich.

Die Linkfunktion und die Verteilung der Responsevariablen stehen in engem Zusammenhang.

Distribution	Name	Link Function	Mean Function
Normal	Identity	$X\beta = \mu$	$\mu = X\beta$
Exponential	Inverse	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Gamma			
Poisson	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$
Binomial	Logit	$X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(X\beta)}{1 + \exp(X\beta)} = \frac{1}{1 + \exp(-X\beta)}$
Multinomial			

## 4 Logistische Funktion als Beispiel

Bei binärer response-variable  $\pi$  gibt es nur die Ausgänge 0 und 1 und sie sind bernoulli-verteilt (hier binomialverteilt). Problem: Klassisches Lineares Modell passt nicht. Lösung: Generalisiertes Lineares Modell (GLZ). Ansatz: Eine Link-Funktion vermittelt zwischen Erklärungsvariablen und der Responsevariable. Allgemein

Am Beispiel der binären Responsevariablen ist die Link-Funktion die Logit-Funktion. Logit-Wert L ist der Logarithmus naturalis eines Odds, also Wahrscheinlichkeit p durch Gegenwahrscheinlichkeit p-1:

$$L = \text{logit}(p) = \log(\text{odds}) = \log\left(\frac{p}{1-p}\right)$$

Umkehrung:

$$p = \frac{e^L}{1 + e^L} = \frac{1}{1 + e^{-L}}$$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q = \beta_0 + \sum_{k=1}^q \beta_k x_k$$

## 4.1 Beispiel GHQ Data Everitt

Resultate eines GLZ in R

nach Everitt GHQ-Data Geschlecht (1=Frau, 2=Männer) GHQ-Werte schwanken von 0(gesund) - 10 (krank) volles Modell ohne Interaktion (wie bei Everitt)

$$\beta_0 = -2.49351, \beta_1 = -0.93609, \beta_2 = 0.77910$$

exploratory variables:  $x_1$  Geschlecht  $x_2$  GHQ

### 4.1.1 Frau (1), GHQ niedrig (1)

$$\begin{aligned} \text{logit}(\pi) &= \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= -2.49351 + -0.93609 * 1 + 0.77910 * 1 \\ &= -2.6505 \end{aligned}$$

Wahrscheinlichkeit für diese Person psychiatrisch zu erkranken  $\pi$  ist dann:

$$\begin{aligned} \pi &= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} \\ &= \frac{e^{-2.49351 + -0.93609 * 1 + 0.77910 * 1}}{1 - e^{-2.49351 + -0.93609 * 1 + 0.77910 * 1}} \\ &= \frac{-2.6505}{1 - (-2.6505)} \\ &= 0.0659582 \end{aligned}$$

Die zugehörigen Odds (für diese Person)

$$\begin{aligned} \text{odds} &= \frac{\pi}{1 - \pi} = \frac{0.0659582}{1 - 0.0659582} \\ &= 0.0706159 \end{aligned}$$

#### 4.1.2 Mann (2), GHQ niedrig (1)

$$\begin{aligned}\text{logit}(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= -2.49351 + -0.93609 * 2 + 0.77910 * 1 \\ &= -3.58659\end{aligned}$$

$$\begin{aligned}\pi &= \frac{e^{-2.49351 + -0.93609 * 2 + 0.77910 * 1}}{1 - e^{-2.49351 + -0.93609 * 2 + 0.77910 * 1}} \\ &= 0.02694639\end{aligned}$$

$$\begin{aligned}\text{odds} &= \frac{\pi}{1-\pi} = \frac{0.02694639}{1 - 0.02694639} \\ &= 0.0276926\end{aligned}$$

#### 4.1.3 Frau (1), GHQ hoch (9)

$$\begin{aligned}\text{logit}(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= -2.49351 + -0.93609 * 1 + 0.77910 * 9 \\ &= 3.5823\end{aligned}$$

$$\begin{aligned}\pi &= \frac{e^{-2.49351 + -0.93609 * 1 + 0.77910 * 9}}{1 - e^{-2.49351 + -0.93609 * 1 + 0.77910 * 9}} \\ &= 0.9729409\end{aligned}$$

$$\begin{aligned}\text{odds} &= \frac{\pi}{1-\pi} = \frac{0.9729409}{1 - 0.9729409} \\ &= 35.95614\end{aligned}$$

Diese Frau hätte eine 35.9-fach höhere Chance psychiatrisch zu erkranken, als gesund zu bleiben (odds).

#### 4.1.4 Mann (2), GHQ hoch (9)

### 4.2 Ein Beispiel

Ein Beispiel aus dem englischen Wikipedia: logistische Regression

The application of a logistic regression may be illustrated using a fictitious example of death from heart disease. This simplified model uses only three risk factors (age [years over 50], sex [0=female, 1=male], and blood cholesterol level [over 5]) to predict the 10-year risk of death from heart disease. These are the parameters that the data fit:

$$\beta_0 = -5, \beta_1 = 2, \beta_2 = -1, \beta_3 = 1.2$$

exploratory variables:  $x_1 = 0$  (50 - Alter[50])  $x_2 = 0$  (weiblich)  $x_3 = 2$  (Wert 7 liegt 2 über 5)

$$\begin{aligned} \text{logit}(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= -5 + 2 * 0 + -1 * 0 + 1.2 * 2 \\ &= -2.6 \end{aligned}$$

$$\log(\text{odds}) = -2.6$$

Wahrscheinlichkeit für diese Person in den nächsten 10 Jahren an Herztod  $\pi$  zu sterben ist dann:

$$\begin{aligned} \pi &= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)} \\ &= \frac{e^{-5+2*0+-1*0+1.2*2}}{1 - e^{-5+2*0+-1*0+1.2*2}} \\ &= \frac{-2.6}{1 - (-2.6)} \\ &= 0.06913842 \end{aligned}$$