# 1 Odds and odds-ratio

## 1.1 Odds

Odds: Verhältnis der Wahrscheinlichkeit, dass ein Ereignis eintritt zur Wahrscheinlichkeit, dass es nicht eintritt.

Odds-Ratio: Verhältnis zweier Odds zueinander. $OddsRatio > 1$: Odds in erster Gruppe (Zähler) sind größer. Bei Erkrankungen: Die Wahrscheinlichkeit zu erkranken ist in erster Gruppe größer als in zweiter Gruppe.

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

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

$$p(case) = \frac{68}{278} = 0.2446043$$

$$odds(case) = \frac{p}{1-p} = \frac{0.2446043}{1-0.2446043} = 0.3238095$$

Sensitivität: positiv erkannt — tatsächlich positiv
Spezifität: negativ erkannt — tatsächlich negativ

# 2 Logistische Funktion

Allgemeine Form

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

Umformung erfolgt durch Multiplikation mit $e^{-x}$. $e^{-x} * e^x = 1$
Verdeutlichung der Parameter:

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

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

# 3 Generalisierte Lineare Modelle

## 3.1 Nomenklatur

Leider verwechselbare Bezeichungen.

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 Zuccini) [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)

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

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

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

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 | $\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$ | $\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$ |
| **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:

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

Umkehrung:

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

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

## 4.1 Beispiel GHQ Data Everitt

### 4.1.1 Modell anpassen in R

```
## GHQ Anpassung des Modells
dd <- read.delim("http://www.psych.uni-goettingen.de/mat/mv/everitt-ghq-vp.txt")

# einen Anriss der Tabelle
dd[c(1:5,167:176,273:278),]

# Anzahl der Casese und Codes:
print(c('Anzahl non-Cases (case=0)', length(dd$case[dd$case == 0])))
print(c('Anzahl Cases (case=1)', length(dd$case[dd$case == 1])))

# Kreuztabellierung der Kombinationen
table(dd$sex, dd$case)


# Eine erste Visualisierung, die die Unsinnigkeit einer üblichen MR veranschaulicht
plot(dd[,1:4])

# logistische Regression wird in R mit dem GLM gerechnet
# da GLM verschiedene Auswertungsfamilien beherrscht, muss das spezifiziert werden
# die Angabe data= verweist auf einen DataFrame, aus dem die Daten gelesen werden (kein attach notwendig)
# das angepasste Modell kann wie üblich gespeichert werden

# Fitten der ghq-Werte
m.dd.ghq <- glm(case ~ ghq, data=dd, family=binomial)
# mit Summary
summary(m.dd.ghq)

# nur Geschlecht als Prädiktor hat wenig Erklärungswert
m.dd.sex <- glm(case ~ factor(sex), data=dd, family=binomial)
```

```
summary(m.dd.sex)

# GHQ und Geschlecht als Prädiktor aber ohne Interaktion
m.dd.g.s <- glm(case ~ factor(sex) + ghq, data=dd, family=binomial)
summary(m.dd.g.s)
# wie sind die Koeffizienten
coefficients(m.dd.g.s)
# dieselben Werte in verständlicherer Regressionsparameter-Form
b0 <- coefficients(m.dd.g.s)[1]
b.sex <- coefficients(m.dd.g.s)[2]
b.ghq <- coefficients(m.dd.g.s)[3]

# mit diesen Koeffizienten kann der logit, also der ln(odds) vorhergesagt werden
logit.cases <- b0 + b.sex * dd$sex + b.ghq * dd$ghq

# aus den Logits können die individuellen Erkrankungswahrscheinlichkeiten berechnet werden
p.cases <- exp(logit.cases) / (1 + exp(logit.cases))

# ... und zu Dataframe dd hinzufügen als Spalte
dd$p <- p.cases

# ein Plot: pro Geschlecht: GHQ-Wert gegen Erkrankungswahrscheinlichkeit
  plot(dd$ghq[dd$sex == 1], dd$p[dd$sex == 1], pch='f', cex=1)
points(dd$ghq[dd$sex == 2], dd$p[dd$sex == 2], pch='m', cex=1)

# hier sollten die genealogischen Zeichen kommen, aber der Font macht nicht mit
# plot(dd$ghq[dd$sex == 1], dd$p[dd$sex == 1], pch=-0x2540L, cex=2)
#points(dd$ghq[dd$sex == 2], dd$p[dd$sex == 2], pch=-0x2542L, cex=2)


# beispielhaft ein paar individuelle Vorhersagen:
# Frau (sex=1) mit GHQ 1
logit.p <- b0 + b.sex * 1 + b.ghq * 1
logit.p
exp(logit.p)/(1 + exp(logit.p))

# Frau (sex=1) mit GHQ 9
logit.p <- b0 + b.sex * 1 + b.ghq * 9
logit.p
exp(logit.p)/(1 + exp(logit.p))


# Mann (sex=2) mit GHQ 1
logit.p <- b0 + b.sex * 2 + b.ghq * 1
logit.p
exp(logit.p)/(1 + exp(logit.p))

# Mann (sex=2) mit GHQ 9
logit.p <- b0 + b.sex * 2 + b.ghq * 9
logit.p
exp(logit.p)/(1 + exp(logit.p))
```

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

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

$$\begin{aligned} logit(\pi_i) &= log(\frac{\pi_i}{1-\pi_i}) \\ &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} \\ &= -2.49351 + -0.93609 * 1 + 0.77910 * 1 \\ &= -2.6505 \end{aligned}$$

Wahrscheinlichkeit für diese Person psychatrisch 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)

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

Diese Frau mit niedrigem GHQ hätte eine 0.07-fach höhere Chance psychiatrisch zu erkranken, als gesund zu bleiben (odds).

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

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

$$\begin{aligned} \pi &= \frac{e^{-2.49351+-0.93609*1+0.77910*9}}{1 + e^{-2.49351+-0.93609*1+0.77910*9}} \\ &= 0.9729415 \end{aligned}$$

$$\begin{aligned} odds = \frac{\pi}{1-\pi} &= \frac{0.9729415}{1 - 0.9729415} \\ &= 35.95695 \end{aligned}$$

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

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

$$\begin{aligned} logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= -2.49351 + -0.93609 * 2 + 0.77910 * 1 \\ &= -3.586595 \end{aligned}$$

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

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

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

$$\begin{aligned} logit(\pi) &= log(\frac{\pi}{1-\pi}) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &= -2.49351 + -0.93609 * 2 + 0.77910 * 9 \\ &= 2.646228 \end{aligned}$$

$$\begin{aligned} \pi &= \frac{e^{-2.49351+-0.93609*2+0.77910*9}}{1+e^{-2.49351+-0.93609*2+0.77910*9}} \\ &= 0.9337781 \end{aligned}$$

$$\begin{aligned} odds = \frac{\pi}{1-\pi} &= \frac{0.9337781}{1-0.9337781} \\ &= 14.10076 \end{aligned}$$

Dieser Mann mit hohem GHQ hätte eine 14.1-fach höhere Chance, psychiatrisch zu erkranken, als gesund zu bleiben.

Bei gleich hohem GHQ-Wert halbiert also ein Mann zu sein die Erkrankungschance im Vergleich zu Frauen.

Als Odds-Ratio ausgedrückt:

$$\begin{aligned} OddsRatio_{GHQ=1} &= \frac{odds_{Frau}}{odds_{Mann}} \\ &= \frac{0.0706159}{0.0276926} \\ &= 2.549992 \end{aligned}$$

Bei GHQ von 1 haben Frauen also ein 2.54-fach höheres Erkrankungsrisiko als Männer

$$\begin{aligned} Odds - Ratio_{GHQ=9} &= \frac{odds_{Frau}}{odds_{Mann}} \\ &= \frac{35.95695}{14.10076} \\ &= 2.550001 \end{aligned}$$

Bei GHQ von 9 haben Frauen ein 2.55-fach höheres Erkrankungsrisiko als Männer

Odds-Ratio bleibt über die Stufen der GHQ-Werte gleich. Odds-Ratio lässt keine Rückschlüsse mehr zu auf die Grundwahrscheinlichkeiten bzw. die zugrunde liegenden Odds.

## 4.2 Beispiel Risiko Herzerkrankung

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)

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

log(odds) = -2.6

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

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

## 4.3 andere Parameter

### 4.3.1 Wald

zum Prüfen von Einzelparametern (Prädiktoren)

$$W = \left(\frac{\beta_j}{s_{\beta_j}}\right)^2$$

$\beta_j$ = Regressionskoeffizient der Variable j
j = unabhängige Variable j
$s_{\beta_j}$ = Standardfehler von $\beta$
Die Wald-Statistik ist $\chi^2$ verteilt