Binär- und Binomialdaten

Binäre Zielgrösse:

$$Y_i = \begin{cases} 0 & \text{Misserfolg} \\ 1 & \text{Erfolg} \end{cases}$$

n Beobachtungen

Beispiele:

- Elektronische Komponente: defekt/ nicht defekt
- Insekt: stirbt/ stirbt nicht nach toxischer Exposition
- PatientIn: Übelkeit/keine Übelkeit nach Operation.

Stress der Eltern von ambulant operierten Kindern

Auftreten von Stress hängt eventuell ab vom Geschlecht des Kindes, Nationalität, Wartezeit, . . .

Eltern-Nr.	Geschlecht des Kindes	Deutschsprachig	unerwarteter Schmerz	Stress
1	m	ja	ja	nein
2	m	nein	ja	nein
. . .	. . .	. . .	. . .	. . .

Binärdaten: $p_i = P(Y_i = 1), \quad Y_i \sim \mathcal{B}(1, p_i)$
(ungruppiert)

Insektizid Rotenon

Konzentration (log von mg/l)	Anzahl Insekten (n_i)	Anzahl Getötete (y_i)
0.96	50	6
1.33	48	16
1.63	46	24
2.04	49	42
2.32	50	44

Binomialdaten: $Y_i \sim \mathcal{B}(n_i, p_i)$
(gruppiert)

p_i abhängig von erklärenden Variablen x_1, x_2, x_3

Lineares Regressionsmodell

$$E(Y_i/n_i) = p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

ist schlecht, weil

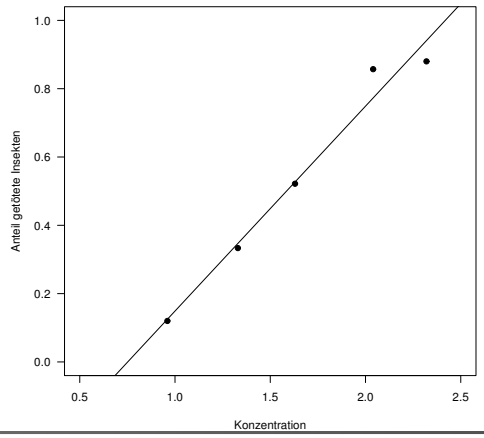
- Angepasste Werte $\hat{p_i}$ können ausserhalb des Intervalls $(0, 1)$ liegen. Prognose wenig sinnvoll!
- Varianz der Zielvariablen Y_i/n_i ist nicht konstant, sondern $p_i(1 - p_i)/n_i$.

Insektizid (Fort.)

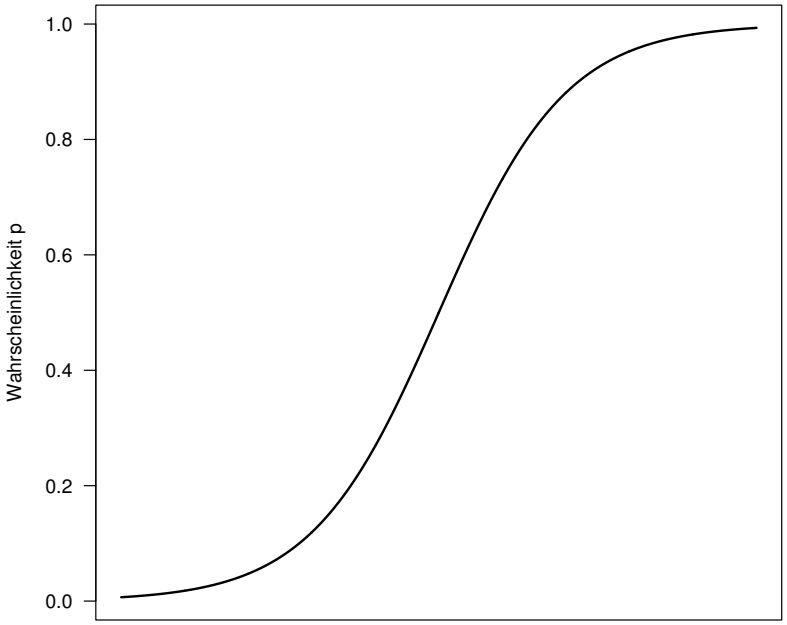
Einfache lineare Regression:

$$\hat{p} = -0.451 + 0.5999 \cdot \text{Konz}$$

Für Konzentrationen über 2.42 wird $\hat{p} > 1$!



Zusammenhang zwischen x und p



Link-Funktionen

Transformation: $p \in (0,1) \mapsto g(p) \in (-\infty, \infty)$

mit $g(p) = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$

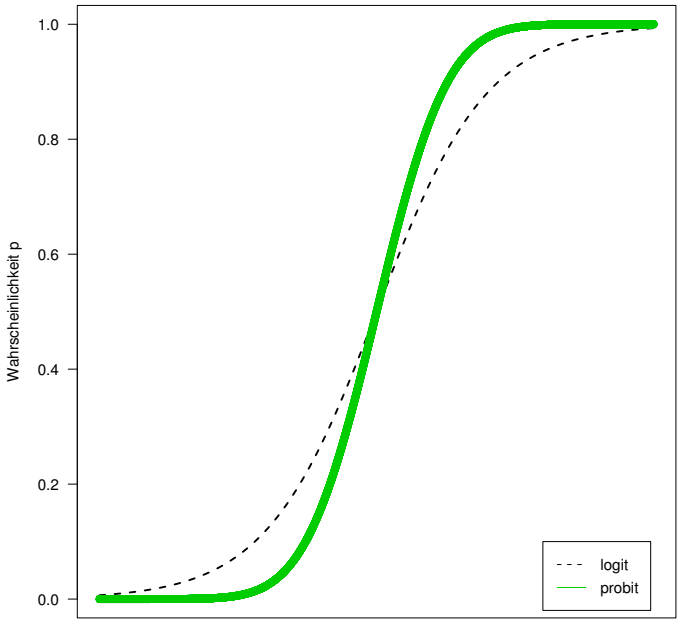
Logit-Transformation:

$$logit(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

$$g(p) = \log(\tfrac{p}{1-p}), \quad p = \tfrac{exp(\eta)}{1+exp(\eta)}$$

Probit-Transformation: $g(p) = \Phi^{-1}(p), \quad p = \Phi(\eta)$

Logit- und Probit-Transformation



Lineares logistisches Modell

Gegeben sind n unabhängige binomialverteilte Zielgrössen Y_i

mit Erfolgswahrscheinlichkeit $p_i = E(Y_i/n_i)$

und p_i hängt von erklärenden Variablen $x_1, x_2, \ldots$ in der folgenden Form ab:

$$logit(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

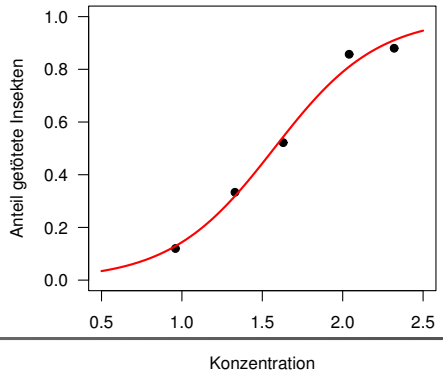
Maximum Likelihood-Schätzungen für die Koeffizienten führt auf nichtlineares Gleichungssystem. Lösung durch IWLS.

R-Output für Insektizid

```
> glm1=glm(cbind(y,n-y)~conz,family=binomial)
> summary(glm1)
Call:
glm(formula=cbind(y, n - y)~conz, family = binomial)

Deviance Residuals:
      1        2        3        4        5
-0.1963   0.2099  -0.2978   0.8726  -0.7222

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8923     0.6426  -7.613 2.67e-14 ***
conz          3.1088     0.3879   8.015 1.11e-15 ***
---
Signif.codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

R-Output für Insektizid (Fort.)

```
(Dispersion parameter for binomial f.. taken to be 1)

    Null deviance: 96.6881  on 4  degrees of freedom
Residual deviance:  1.4542  on 3  degrees of freedom
AIC: 24.675
Number of Fisher Scoring iterations: 4
```



Interpretation der Koeffizienten

Fit: $\log(\frac{\hat{p}}{1-\hat{p}}) = -4.8923 + 3.1088 \cdot \text{Konz}$

Interpretation von $\hat{\beta}_1$ schwierig,
Retourtransformation nötig.

$$\frac{\hat{p}}{1-\hat{p}} = \exp(-4.8923 + 3.1088 \cdot \text{Konz})$$

= **Odds** getötet zu werden in Abhängigkeit von einer Konzentration.

Odds ratios

$\hat{p_0}$= Wahrscheinlichkeit getötet zu werden bei einer Konzentration von Konz$_0$
$\hat{p_1}$= Wahrscheinlichkeit getötet zu werden bei einer Konzentration von Konz$_0 + 1$

$$\frac{\frac{\hat{p_1}}{1-\hat{p_1}}}{\frac{\hat{p_0}}{1-\hat{p_0}}} = \frac{\exp(-4.8923 + 3.1088 \cdot (\text{Konz}_0 + 1))}{\exp(-4.8923 + 3.1088 \cdot \text{Konz}_0)}$$

$$= e^{3.1088} = 22.39$$

R-Output für Stress

```
> summary(stress)

Call: glm(formula = stress ~ sex + narkose + schmerz,
binomial, data = daten2)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7368     0.2554  -6.801 1.04e-11 ***
sexw         -1.1078     0.4376  -2.532 0.011353 *
narkose       0.8663     0.3450   2.511 0.012029 *
schmerz       1.7537     0.5243   3.345 0.000824 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
```

R-Output für Stress (Fort.)

```
(Dispersion parameter for binomial family taken to be

    Null deviance: 244.98  on 250  degrees of freedom
Residual deviance: 223.34  on 247  degrees of freedom
AIC: 231.34

Number of Fisher Scoring iterations: 5
```


Odds ratios

$\hat{p_0}$= Wahrscheinlichkeit für Stress bei einem Mädchen
$\hat{p_1}$= Wahrscheinlichkeit für Stress bei einem Knaben

$$\frac{\frac{\hat{p_1}}{1-\hat{p_1}}}{\frac{\hat{p_0}}{1-\hat{p_0}}} = \frac{\exp(-1.7368+\ldots)}{\exp(-1.7368+\ldots-1.1078)}$$

$$= e^{1.1079} = 3.028.$$