Logistische Regression

21.11.2011

Einführung

Beispiele

- Medikament, Phase-I study (FDA):
  Suche Dosis, sd. max. 1/3 von (gesunden) Probanden Nebenwirkungen zeigt.
- Tierversuch:
  Bei welcher Dosis überleben 50% der Mäuse (=LD50)?
- Frühgeburten:
  Wie hängt die Überlebenswahrscheinlichkeit ab von gewissen Variablen (Gewicht, Alter, APGAR score, ...)?
- Technik:
  Bei welchen Bedingungen fallen Geräte aus?
- Customer-Relationship-Management (CRM):
  Was für Massnahmen sind erfolgreich, damit ein Kunde z.B. auf ein teureres Produkt wechselt?
- Google nach "Nate Silver" (Oscar-, Election-prediction etc.)

Gemeinsamkeiten

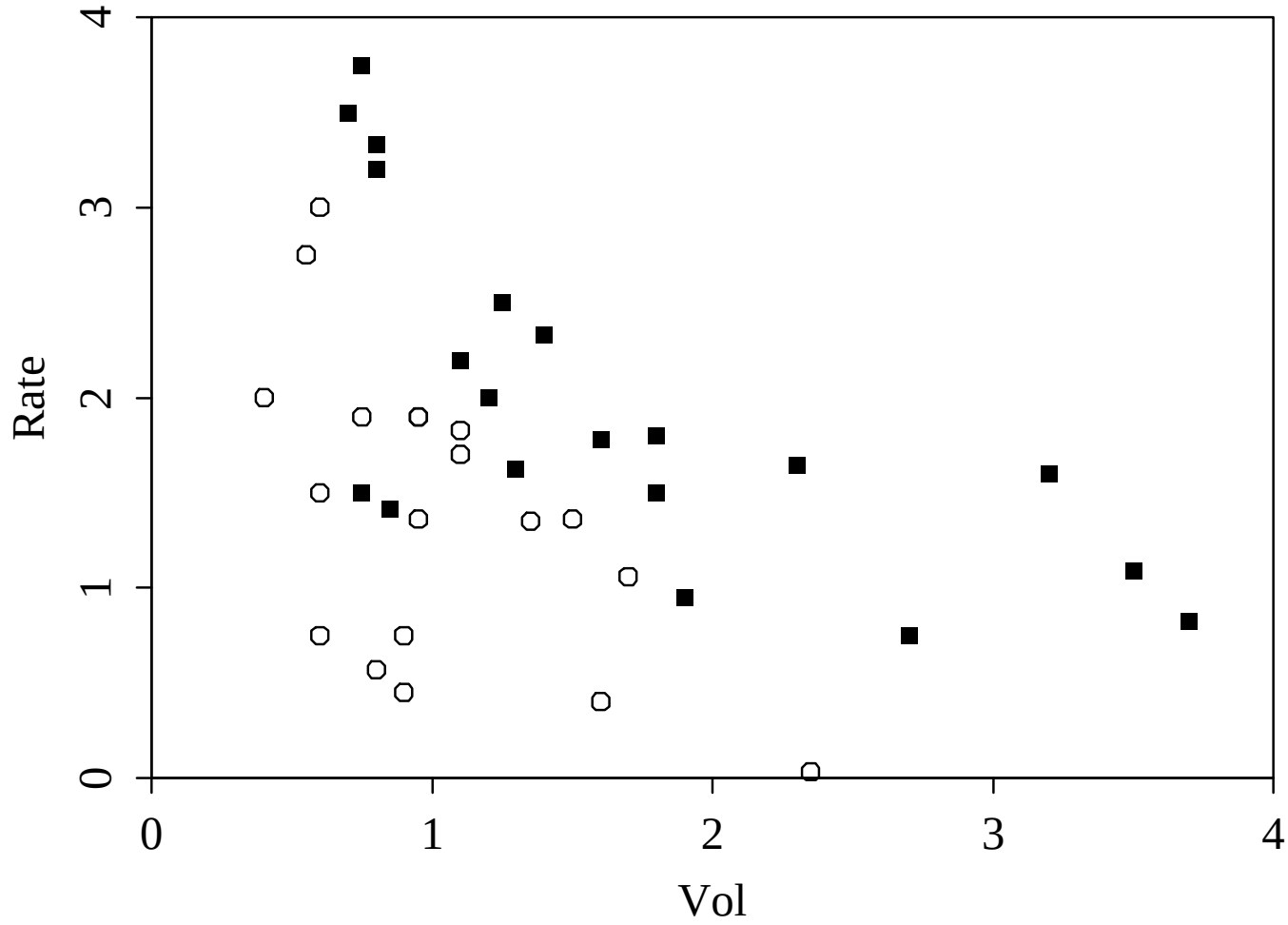
- **Binäre Zielgrösse** (ja/nein, lebt/tot, ...).
- Beliebige erklärende Variablen.

Dies hat die Form eines Regressionsmodells. Was ist neu?

Die Zielgrösse ist nicht mehr stetig, sondern **binär**.

Beispiel Ader-Verengung

Y: Ader-Verengung ja (1) / nein (0)
Erklärende: Atem-Volumen (`Vol`) und Atem-Frequenz (`Rate`)



Ausgefüllte Symbole: Ader-Verengung ja (1).

Modellansatz

Ziel Modelliere $P(Y_i = 1 \mid x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$

Ansatz $P(Y_i = 1 \mid x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)}) = h(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})$

Bemerkung

Für $Y \in \{0, 1\}$ gilt

$$\begin{aligned} E[Y] &= 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) \\ &= P(Y = 1). \end{aligned}$$

D.h. wir modellieren eigentlich

$$E[Y_i \mid x_i^{(1)}, \ldots, x_i^{(m)}] = h(x_i^{(1)}, \ldots, x_i^{(m)})$$

wie in der linearen Regression.

Ansatz mit linearer Regression?

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + E_i$$

Es gilt dann

- $P(Y_i = 1 \,|\, \underline{x}_i) = E(Y_i \,|\, \underline{x}_i) = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)}$
  D.h. die Funktion h wäre linear.
- Geschätzte Werte können daher < 0 und > 1 werden ↯
- Transformation von Y_i? 2 Werte bleiben 2 Werte ↯

Ausweg: Transformation von $E(Y) = P(Y = 1)$!

Am Besten transformieren wir die Wahrscheinlichkeiten so, dass wir **keine Einschränkungen** mehr haben.

Logistisches Regressionsmodell

Benutze **Logit-Funktion** $g : [0, 1] \to \mathbb{R}$

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

- Die Funktion g transformiert die Wahrscheinlichkeiten auf die gesamte reelle Achse!
- $\log\left(\frac{\pi}{1-\pi}\right)$ sind die **log-odds** (siehe Vorlesung kategorielle Variablen)!

Modell

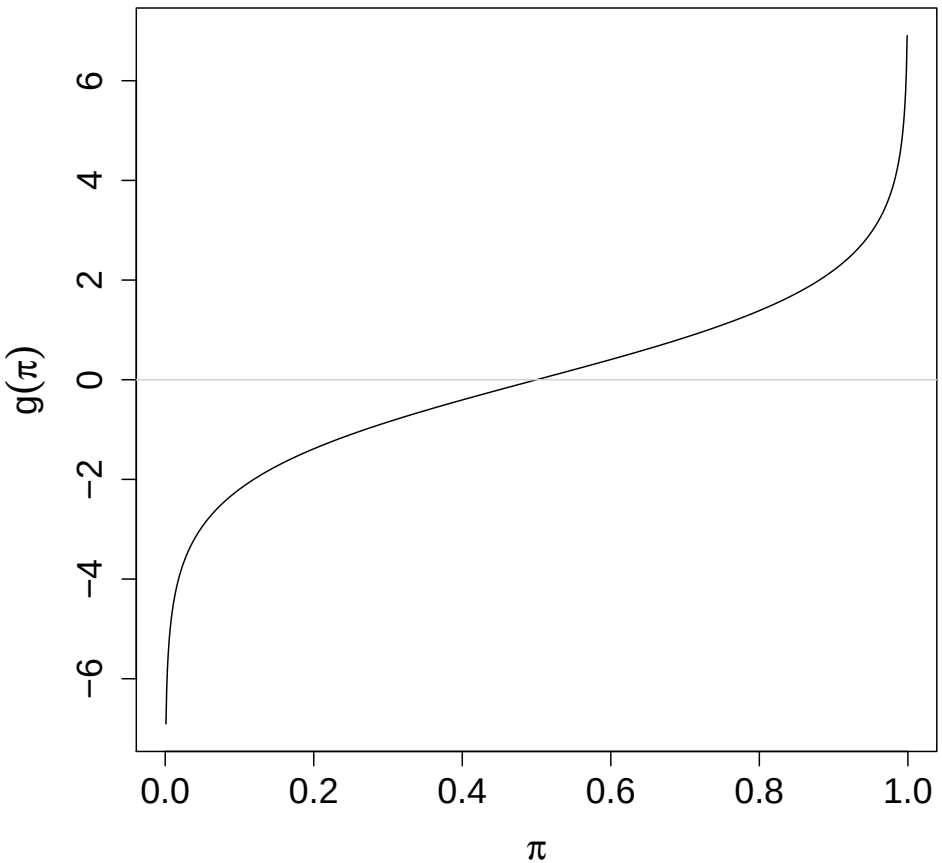
Auf der transformierten Skala verwenden wir den "alten" linearen Ansatz, d.h.

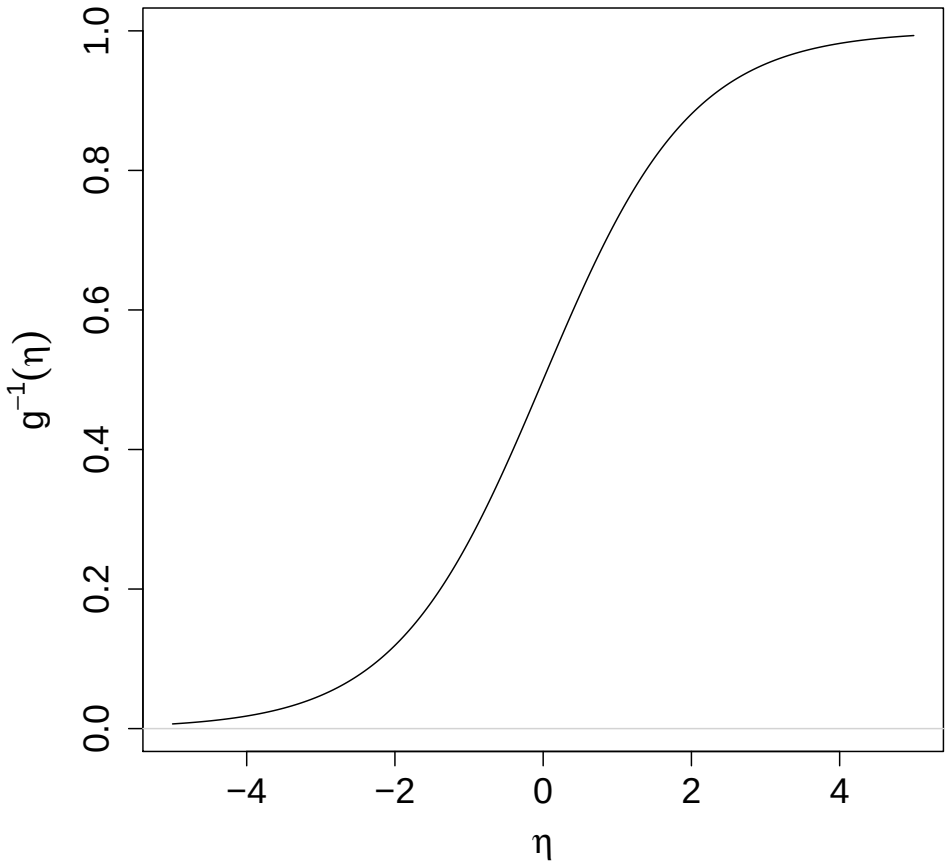
$$\begin{aligned} g(P(Y_i = 1 \mid \underline{x}_i)) &= \log\left(\frac{P(Y_i = 1 \mid \underline{x}_i)}{P(Y_i = 0 \mid \underline{x}_i)}\right) \\ &= \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} \\ &= \underline{x}_i^T \underline{\beta} \\ &= \eta_i. \end{aligned}$$

Terminologie

- $\eta_i = \underline{x}_i^T \underline{\beta}$ heisst **linearer Prädiktor**.
- g heisst **Linkfunktion**.
  Die Linkfunktion transformiert den Erwartungswert auf die "geeignete Skala".

Link- und inverse Linkfunktion





Kennt man den linearen Prädiktor (oder die β_k's), so erhält man die Wahrscheinlichkeiten durch

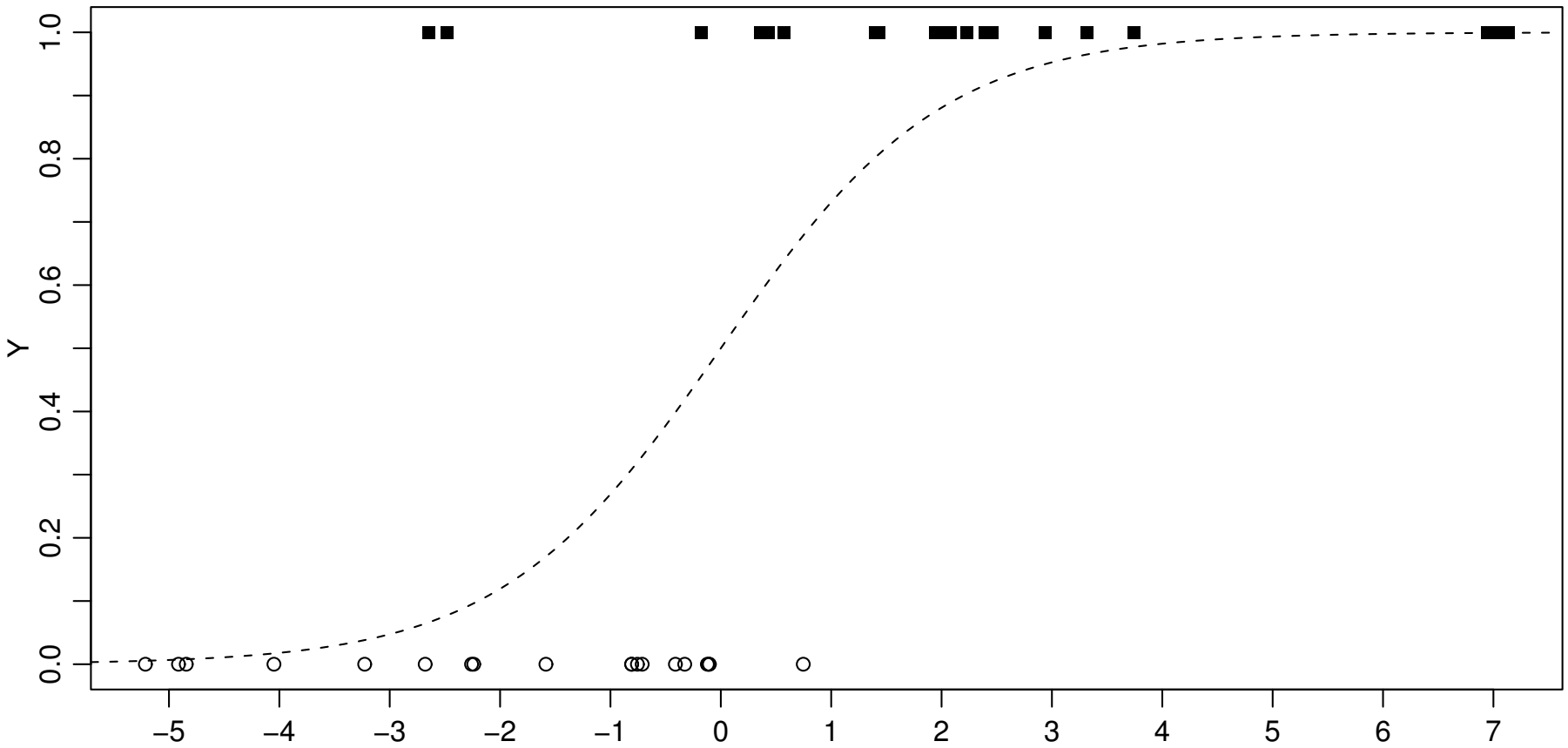
$$P(Y_i = 1) = g^{-1}(\eta_i) = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}},$$

bzw.

$$P(Y_i = 0) = 1 - P(Y_i = 1) = \frac{1}{1 + \exp\{\eta_i\}}.$$

Aus notationellen Gründen lassen wir das Bedingen auf $\underline{x}_i$ weg.

Ader-Verengung: Illustration W'keiten vs. linearer Prädiktor



Beispiel Ader-Verengung

Angepasstes Modell liefert

$$g(P(Y=1)) = -9.53 + 3.88 \cdot \texttt{Vol} + 2.65 \cdot \texttt{Rate}.$$

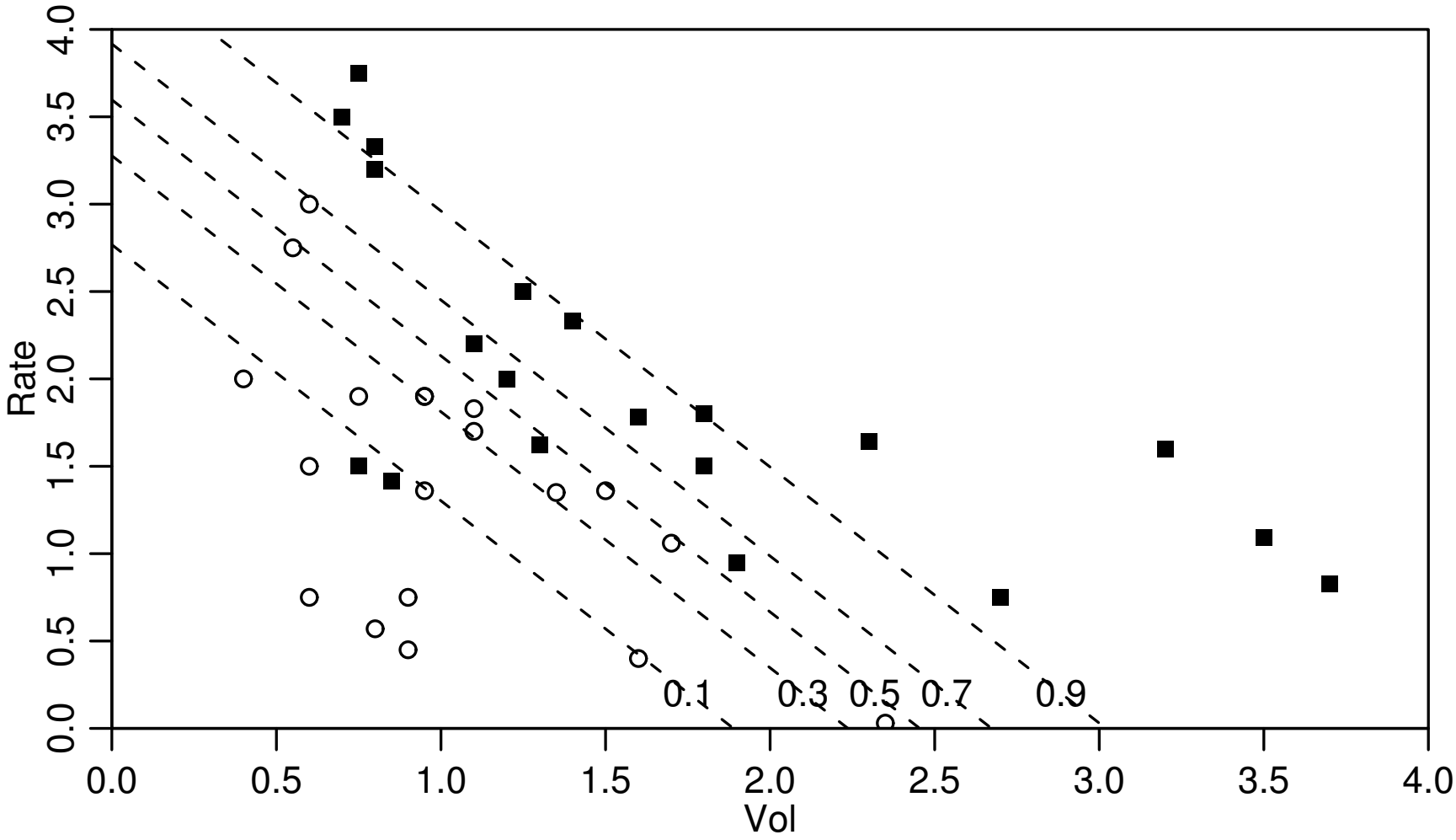
Punkte mit gleichen W'keiten haben die Eigenschaft, dass

$$-9.53 + 3.88 \cdot \texttt{Vol} + 2.65 \cdot \texttt{Rate} = \text{const}.$$

D.h. `Rate` hängt dann linear von `Vol` ab.

Bzw. Punkte mit gleichen W'keiten liegen auf (parallelen) Geraden im Raum (`Vol, Rate`).

Illustration angepasstes Modell für Ader-Verengung



Punkte auf gestrichelten Geraden liefern alle die gleichen W'keiten.

Interpretation der Parameter

Wir betrachten die **odds**

$$\begin{aligned}\text{odds}\,(Y=1\,|\,\underline{x}) &= \frac{P(Y=1\,|\,\underline{x})}{P(Y=0\,|\,\underline{x})}\\ &= \exp\left\{\beta_0+\beta_1 x^{(1)}+\cdots+\beta_m x^{(m)}\right\}.\\ &= \exp\{\beta_0\}\cdot\exp\{\beta_1\}^{x^{(1)}}\cdots\exp\{\beta_m\}^{x^{(m)}}.\end{aligned}$$

- Wenn man $x^{(j)}$ um eine Einheit erhöht (und alles andere fix lässt), so ändern sich die odds um den Faktor $\exp\{\beta_j\}$.
- Das Doppelverhältnis (**odds ratio**) ist $\exp\{\beta_j\}$

$$\frac{\text{odds}\left(Y=1\,|\,x^{(j)}=c_j+1\right)}{\text{odds}\left(Y=1\,|\,x^{(j)}=c_j\right)}=\exp\{\beta_j\},\ \text{für beliebiges } c_j.$$

- Das **log odds-ratio** ist dann entsprechend β_j.

Gruppierte Daten

Manchmal hat man zu den gleichen erklärenden Variablen **mehrere Beobachtungen** (Replikate) der Zielvariable.

Notation

m_ℓ Beobachtungen Y_i zu gleichen $\underline{x}_i = \underline{\widetilde{x}}_\ell$.

Definiere $\widetilde{Y}_\ell = \frac{1}{m_\ell} \sum_{i:\underline{x}_i = \underline{\widetilde{x}}_\ell} Y_i$ (Anteil "Erfolge").

Es gilt dann

- $m_\ell \widetilde{Y}_\ell \sim Bin(m_\ell, \widetilde{\pi}_\ell)$
- $E[\widetilde{Y}_\ell] = \widetilde{\pi}_\ell$,

wobei $\widetilde{\pi}_\ell = P(Y = 1 \,|\, \underline{\widetilde{x}}_\ell)$, $g(\widetilde{\pi}_\ell) = \underline{\widetilde{x}}_\ell^T \underline{\beta}$.

Wir verwenden das gleiche Modell wie vorher.

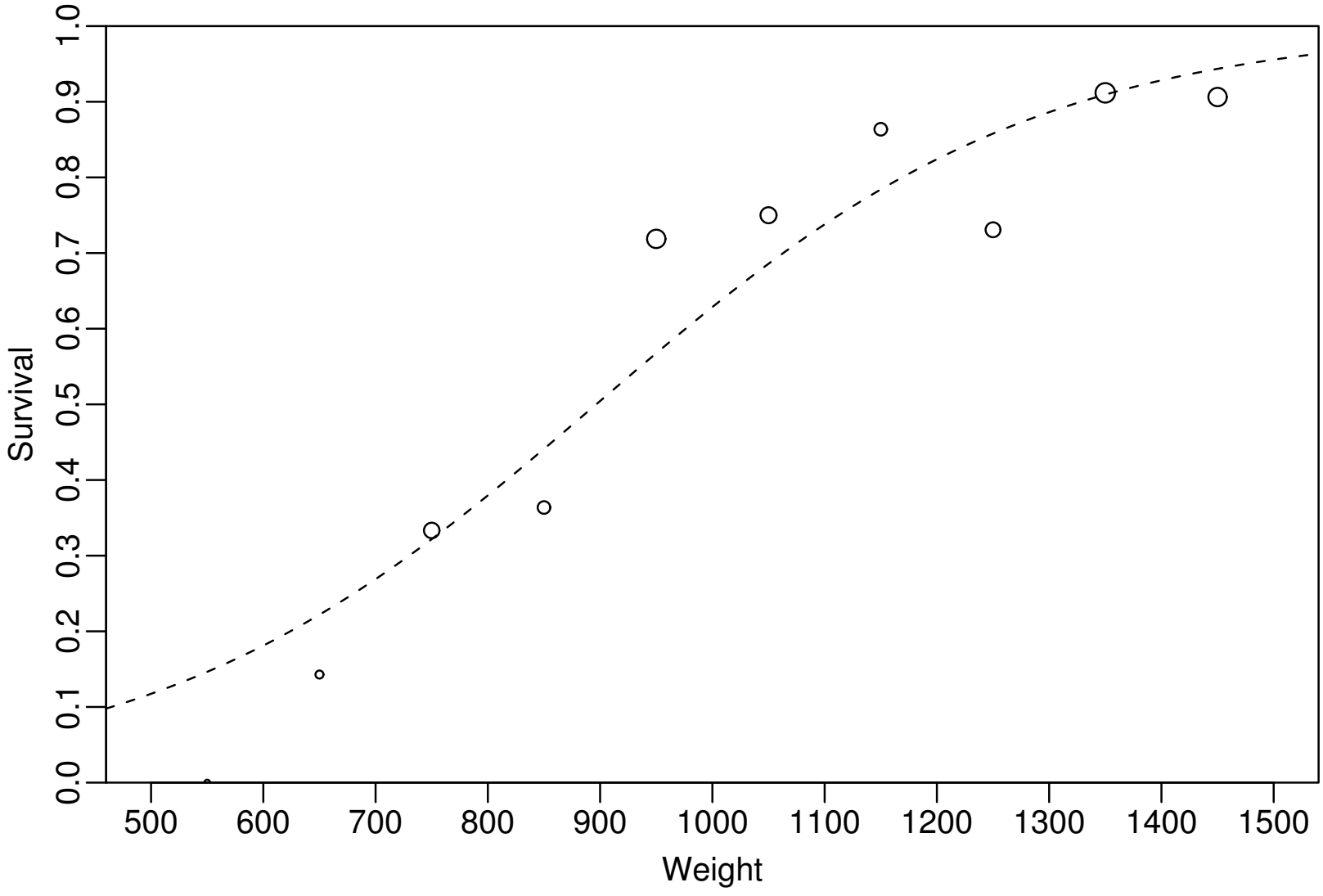
Bei gruppierten Daten hat man den Vorteil, dass man **mehr Informationen** hat. Man könnte eigentlich für jede Gruppe die W'keit einzeln schätzen wenn m_ℓ genügend gross ist.

Beispiel gruppierte Daten: Frühgeburten

- 247 Säuglinge
- Erklärende Variable: Geburtsgewicht
- Klassen von je 100g Gewicht
- Zielvariable: Survival (ja / nein)

	n	Surv.no	Surv.yes	Weight
1	10	10	0	550
2	14	12	2	650
3	27	18	9	750
4	22	14	8	850
5	32	9	23	950
6	28	7	21	1050
7	22	3	19	1150
8	26	7	19	1250
9	34	3	31	1350
10	32	3	29	1450

Illustration: Frühgeburten



Die Fläche der Kreise ist proportional zu der Anzahl Beobachtungen gewählt.

Modell der latenten Variablen

Nehme an, es ex. nicht-beobachtbare Z_i, sd.

$$Z_i = \underline{x}_i^T \underline{\widetilde{\beta}} + E_i.$$

Wir beobachten aber nur, ob Z_i grösser oder kleiner als ein Schwellenwert c ist.

$$Y_i = \begin{cases} 1 & Z_i \geq c \\ 0 & Z_i < c \end{cases}.$$

Jetzt gilt aber:

$$\begin{aligned} \pi_i &= P(Y_i = 1) = P(Z_i \geq c) = P(E_i \geq c - \underline{x}_i^T \underline{\widetilde{\beta}}) \\ &= 1 - F_E\left(c - \left(\widetilde{\beta}_0 + \sum\nolimits_j \widetilde{\beta}_j x_i^{(j)}\right)\right), \end{aligned}$$

wobei F_E: kumulative Verteilungsfunktion der Zufallsfehler E_i.

Definiere nun $\underline{\beta} = [\widetilde{\beta}_0 - c, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_m]$. Jetzt gilt

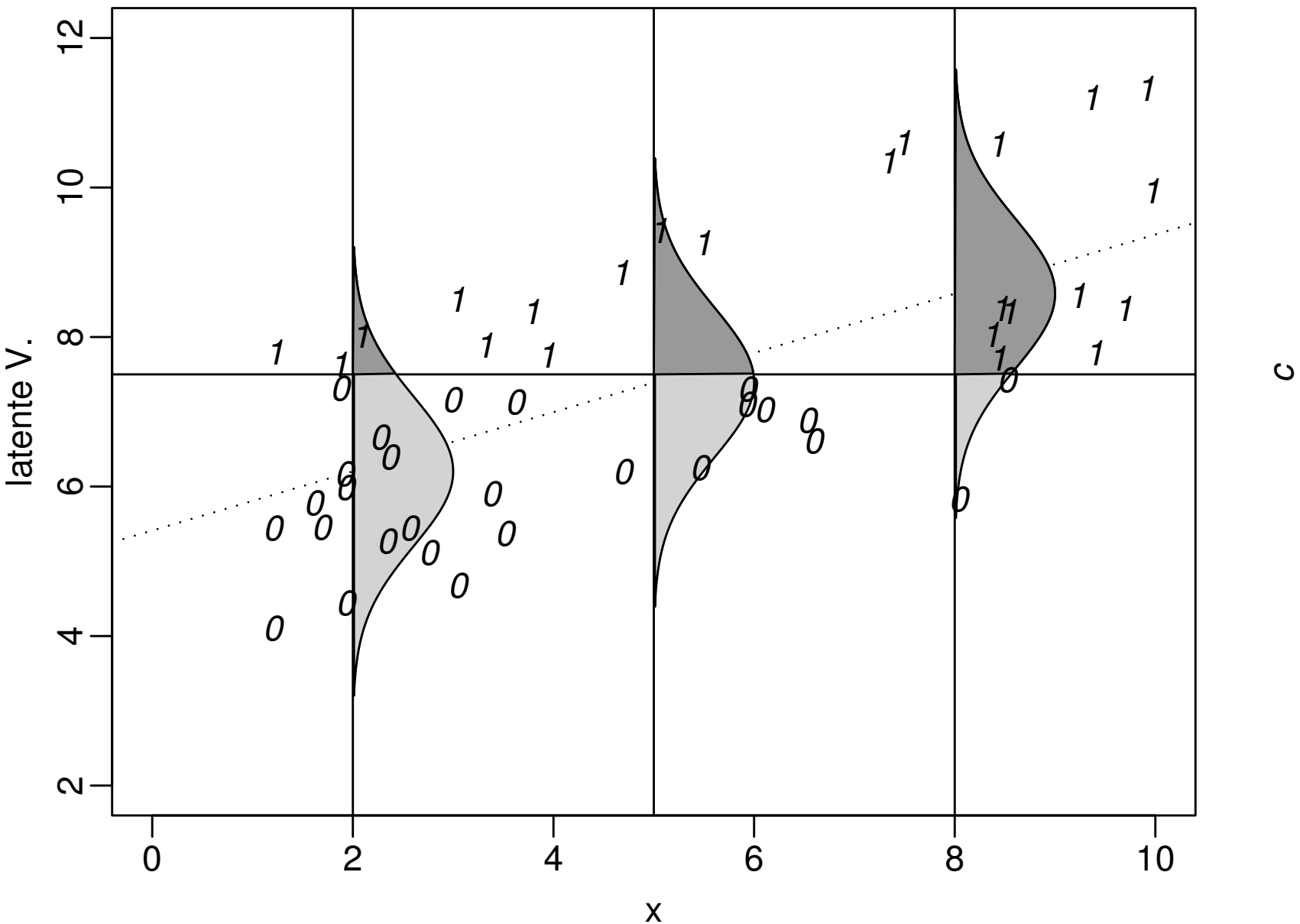
$$P(Y_i = 1) = g^{-1}(\underline{x}_i^T \underline{\beta})$$

mit $g^{-1}(\eta) = 1 - F_E(-\eta) = F_{-E}(\eta)$.

$E_i \sim$ Logistische Verteilung	**Logistische Regressionsmodell**
$E_i \sim$ Normalverteilung	**Probitmodell**
$E_i \sim$ Extremwertverteilung	**Komplementäres log-log Modell**

Das logistische Regressionsmodell hat den Vorteil, dass die Parameter die "schöne" Interpretation mit den odds-ratio haben.

Illustration latente Variable



Schätzungen und Tests

Verwende **Maximum-Likelihood** Prinzip.

Wähle $\underline{\beta}$ so, dass die Wahrscheinlichkeit des beobachteten Ereignisses maximal wird.

Wir haben für unsere n unabhängigen Beobachtungen

$$\ell(\underline{\beta}) = P_{\underline{\beta}}(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P_{\underline{\beta}}(Y_i = y_i),$$

mit $P_{\underline{\beta}}(Y_i = y_i) = \pi_i^{y_i}(1-\pi)^{1-y_i}$. π_i hängt natürlich von $\underline{\beta}$ ab.

Um das Produkt zu vermeiden, arbeitet man mit der **log-likelihood**

$$\ell\ell(\underline{\beta}) = \log(\ell(\underline{\beta})).$$

Schlussendlich erhält man

$$\ell\ell(\underline{\beta}) = \sum_{i=1}^{n} y_i \underline{x}_i^T \underline{\beta} - \log\left(1 + \exp\{\underline{x}_i^T \underline{\beta}\}\right).$$

Maximiere dies bzgl. $\underline{\beta} \rightsquigarrow$ Schätzer $\widehat{\underline{\beta}}$.

- Im Gegensatz zur linearen Regression haben wir **keine geschlossen darstellbare Lösung** ↯
- **Iterative numerische Verfahren** werden benötigt.
- Grundidee: Aproximiere das Problem mit einem gewichteten linearen Regressionsproblem und löse dann sukzessive viele gewichtete lineare Regressionen.
- Entsprechende Routinen sind in allen Statistikpaketen implementiert (Stichwort: Generalized Linear Models, GLM)

Verteilung der geschätzten Koeffizienten

Approximation mit linearen Regressionsproblemen liefert auch (approximative) Verteilung der Parameter ($\rightsquigarrow$ Standardfehler).

- $\underline{\widehat{\beta}}$ ist approximativ normalverteilt mit Erwartungswert β und einer Kovarianzmatrix V.
- Teststatistiken sind dann

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{V}_{jj}}} \overset{approx.}{\sim} N(0,1).$$

Verwende hier Normal- statt t-Verteilung.

Dies ist ein sogenannter **Wald-Test**.

R-Output im Beispiel der Ader-Verengungen

```
Call:
glm(formula = Y ~ Volume + Rate, family = binomial, data = vaso)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5296     3.2332  -2.947  0.00320 **
Volume        3.8822     1.4286   2.717  0.00658 **
Rate          2.6491     0.9142   2.898  0.00376 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.772  on 36  degrees of freedom
AIC: 35.772

Number of Fisher Scoring iterations: 6
```

Devianzen

- **Residuen-Devianz** (früher: Residuenquadratsumme)

  Für gruppierte Daten $\widetilde{Y}_l$. Vergleiche log-likelihood des geschätzten Modells mit derjenigen des maximalen Modells.

$$D(\underline{\widetilde{y}}\,;\underline{\widehat{\pi}}) = 2\left(\ell\ell^{(M)} - \ell\ell\left(\underline{\widehat{\beta}}\right)\right).$$

  $\ell\ell^{(M)}$: Kann für jede Gruppe $\widetilde{\pi}_\ell$ frei wählen (grösstes mögliches Modell).

  Bei ungruppierten Daten gilt: $\ell\ell^{(M)} = 0$ (perfekter fit).

  Residuen-Devianz vergleicht geschätztes Modell mit maximalen Modell ("Anpassungstest"). Geht nur bei nicht zu kleinen m_ℓ.

- **Devianz-Differenz** (zum Vergleich von Modellen)

  Likelihood-Ratio Test für Modellvergleich $K \subset G$:

  $$\widetilde{D}(\underline{y}; \underline{\widehat{\pi}}^{(K)}, \underline{\widehat{\pi}}^{(G)}) = D(\underline{y}; \underline{\widehat{\pi}}^{(K)}) - D(\underline{y}; \underline{\widehat{\pi}}^{(G)}) = 2(\ell\ell^{(G)} - \ell\ell^{(K)}).$$

  Asymptotisch χ^2_d-verteilt, wenn das kleine Modell stimmt.

  Anzahl Freiheitsgrade d ist die Differenz der Anzahl Parameter der beiden Modelle: $d = |G| - |K|$.

  Kann also geschachtelte Modelle miteinander vergleichen.

  Diese **Likelihood-Ratio Tests** sind in der Regel den Wald-Tests vorzuziehen.

  R-Befehle

  - Teste z.B. Faktoren

    ```
    > drop1(fit, test = "Chisq")
    ```
  - Allg. zum Vergleich von geschachtelten Modellen

    ```
    > anova(fit.1, fit.2, test = "Chisq")
    ```

- **Null-Devianz** (früher: $\sum_i \left(Y_i - \overline{Y}\right)^2$)

  Kleinstes Modell (Nullmodell): Besteht nur aus Intercept, d.h. π_i ist für alle Beobachtungen gleich: $\widehat{\pi}^{(0)} = \sum_{i=1}^{n} y_k / n$ (globaler Anteil).

$$D(\underline{y}; \underline{\widehat{\pi}}^{(0)}) = 2 \left(\ell\ell^{(M)} - \ell\ell \left(\underline{\widehat{\beta}}^{(0)} \right) \right).$$

  Damit kann man einen Gesamt-Test für das Modell konstruieren (H_0: alle $\beta_j = 0$, $j = 1, \ldots, m$)

$$D(\underline{y}; \underline{\widehat{\pi}}^{(0)}) - D(\underline{y}; \underline{\widehat{\pi}}) = 2 \left(\ell\ell \left(\underline{\widehat{\beta}} \right) - \ell\ell \left(\underline{\widehat{\beta}}^{(0)} \right) \right).$$

  Unter H_0 ist dies approximativ χ^2_{p-1} verteilt.

Residuenanalyse

Schwieriger als früher. Was ist hier ein Residuum?

Es ex. mehrere mögliche Definitionen.

- **Response Residuals, Raw Residuals**

$$R_\ell = \widetilde{Y}_\ell - \widehat{\pi}_\ell, \quad \widehat{\pi}_\ell = g^{-1}(\underline{\widetilde{x}}_\ell^T \underline{\widehat{\beta}}) \quad \text{(gruppierte Daten)}$$

- **Pearson Residuals**

$$R_\ell^{(P)} = R_\ell / \sqrt{\widehat{\pi}_\ell(1-\widehat{\pi}_\ell)/m_\ell} \quad \text{(standardisiert)}$$

- **Working Residuals, Link Residuals**

  Berechnung der logistischen Regression: iterativ gewichtete Kleinste Quadrate
  $\rightsquigarrow$ lineare Näherung $\rightsquigarrow$ Residuen: „working residuals“.

- **Devianz-Residuen**

$$R_i^{(D)} = \text{sign}(Y_i - \widehat{\pi}_i)\sqrt{d_i},$$

wobei d_i der entsprechende Summand der Residuendevianz ist.

d_i entspricht R_i^2 in der gewöhnlichen linearen Regression.

Grafische Darstellungen

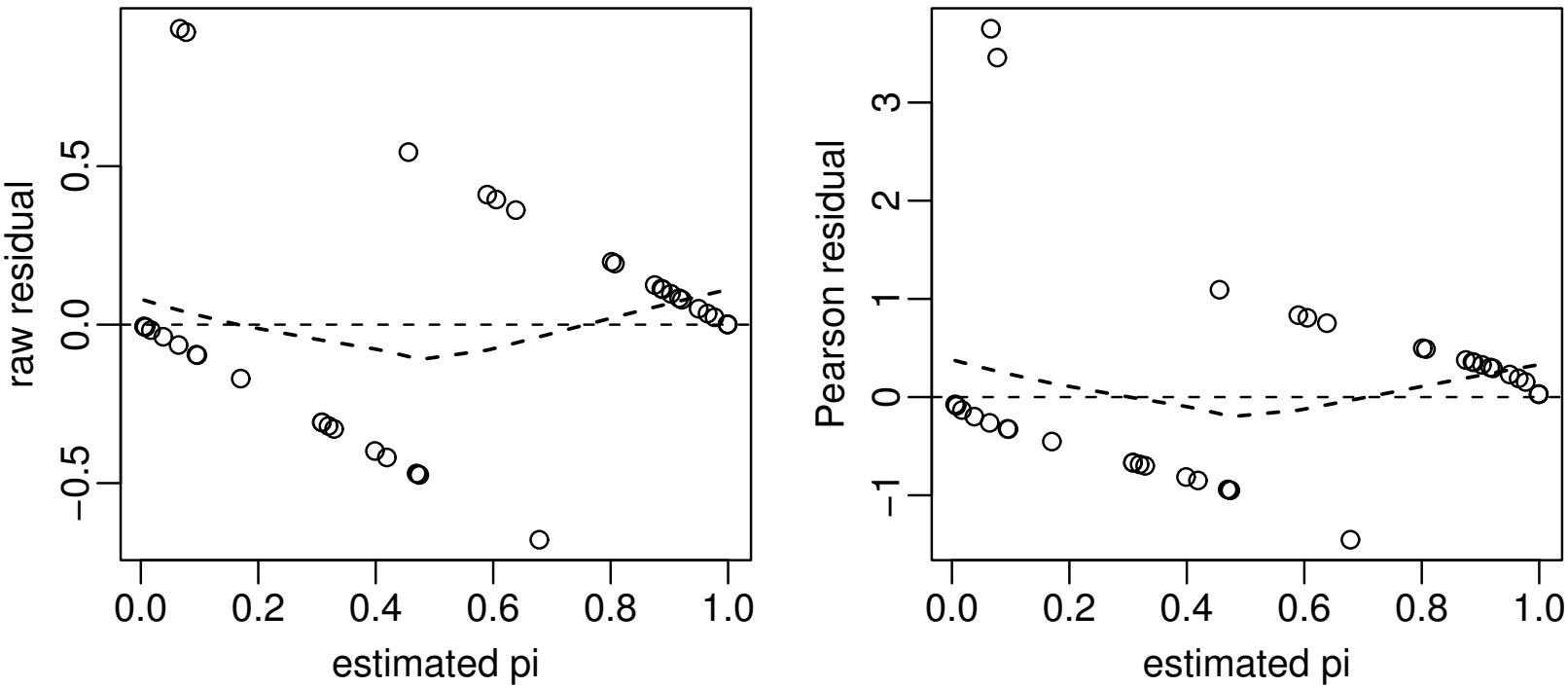
QQ-Plot machen in der Regel keinen Sinn.

Tukey-Anscombe Plot am geeignetsten. Z.B.

- Raw Residuals vs. geschätzte $\widehat{\pi}_i$
- Working Residuals vs. linearer Prädiktor $\widehat{\eta}_i$.

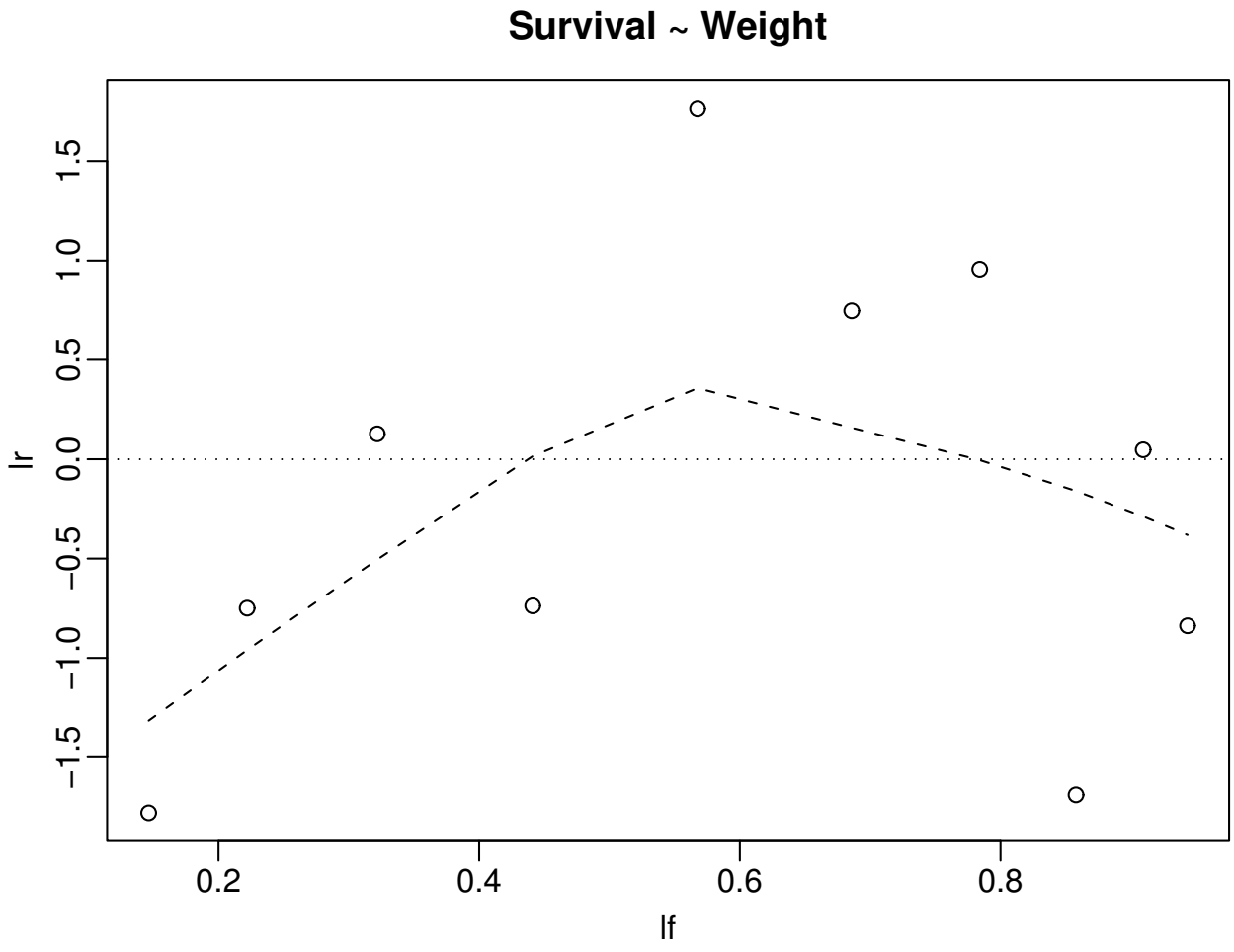
Insbesondere bei nicht gruppierten Daten braucht man einen Glätter (wegen "Artefakten").

Beispiel: Residuenplot bei ungruppierten Daten



Man hat immer die beiden Kurven (Artefakt).

Residuenplot bei gruppierten Daten



Besser interpretierbar.

Merkpunkte

- Logistische Regression für **binäre Zielgrössen**. Gleiche Flexibilität wie gewöhnliche lineare Regression.
- Interpretation mit **odds** bzw. **odds-ratio**:
  log(odds) = linearer Prädiktor
  β_j: log(odds ratio) falls man j-te Variable um eine Einheit erhöht.
- **Schätzungen**, Tests, Vertrauensintervalle via Likelihood-Methoden (Devianzen) und entsprechende Asymptotik.
- **Residuen**: Mehrere Möglichkeiten, wegen Artefakten wird Glätter benötigt