



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

Institut für Statistik



# Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2014



- 5 Metrische Einflußgrößen: Polynomiale Regression, Trigonometrische Polynome, Regressionsplines, Transformationen.
- 6 Modelldiagnose
- 7 Variablenselektion
- Das allgemeine lineare Modell: Gewichtete KQ-Methode, Autokorrelierte und heteroskedastische Störterme
- Das logistische Regressionsmodell**
- 10 Das gemischte lineare Regressionsmodell („Linear mixed Model“)

# Das logistische Regressionsmodell

---

$$\pi_i = P(Y_i = 1|x_i) = G(x_i'\beta) \quad (9.1)$$

$$\ln \frac{\pi_i}{1 - \pi_i} = x_i'\beta \quad (9.2)$$

$$Y_i, i = 1, \dots, n \quad \text{unabhängig (bei gegebenem festen } X) \quad (9.3)$$

$$G(t) = (1 + \exp(-t))^{-1} \quad (9.4)$$

$Y_i$ : binäre Zielgröße

$x_i$ : Vektor der Einflussgrößen

$X$ : Design-Matrix der Einflussgrößen mit vollem Rang

# Bezeichnungen

$\ln \frac{\pi_j}{1-\pi_j}$	:	„Logarithmierte Chance“ Log-odds
$x_j' \beta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$	:	linearer Prädiktor
Funktion $G$	:	Response-Funktion (Inverse Link-Funktion)

Die Wahl von  $G$  (Verteilungsfunktion der logistischen Verteilung) als Responsefunktion ermöglicht folgende Interpretation:  
Einfaches Modell:

$$\begin{aligned}P(Y = 1|x_0) &= G(\beta_0 + \beta_1 * x_0) \\P(Y = 1|x_0 + 1) &= G[\beta_0 + \beta_1 * (x_0 + 1)] \\ \frac{P(Y = 1|x_0 + 1)/(1 - P(Y = 1|x_0 + 1))}{P(Y = 1|x_0)/(1 - P(Y = 1|x_0))} &= \exp(\beta_1) \\ \ln \frac{P(Y = 1|x_0 + 1)}{1 - P(Y = 1|x_0 + 1)} - \ln \frac{P(Y = 1|x_0)}{1 - P(Y = 1|x_0)} &= \beta_1\end{aligned}$$

# Interpretation

Das logistische Regressionsmodell nimmt einen linearen Zusammenhang zwischen den „Log-odds“ von  $Y$  und den Einflussgrößen  $X$  an.

Interpretation:

- Wenn  $x_k$  um einen Einheit steigt, so ändert sich die logarithmierte Chance von  $Y$  um  $\beta_k$ .
- Wenn  $x_k$  um einen Einheit steigt, so ändert sich die Chance von  $Y$  um den Faktor  $\exp(\beta_k)$ .
- Das Odds Ratio (Chancenverhältnis) zwischen  $Y$  bei  $x_k$  und  $Y$  bei  $x_k + 1$  ist  $\exp(\beta)$ .

W'keit	0.01	0.05	0.1	0.3	0.4	0.5	0.6	0.7	0.9	0.95	0.99
Odds	1/99	1/19	1/9	3/7	2/3	1	1.5	7/3	9	19	99
Log odds	-4.6	-2.9	-2.2	-0.85	-0.41	0	0.41	0.85	2.2	2.9	4.6

# Bemerkungen

---

Die Varianten des multiplen linearen Regressionsmodells lassen sich direkt auf das logistische Modell übertragen:

- Behandlung von Gruppenvergleichen (ANOVA) mit Hilfe von Indikatorvariablen
- Behandlung von diskreten Einflussgrößen: verschiedene Codierungen, Interaktionen, etc.
- Behandlung von stetigen Einflussgrößen (Polynome, Splines etc.)

Beachte:

Beim logistischen Modell „fehlt“ der Varianz-Parameter  $\sigma$ ,

da  $Var(Y) = E(Y) * [1 - E(Y)]$

Weiter ist keine Verteilungsannahme nötig, da  $Y$  immer Bernoulli-verteilt ist.

# Beispiele

---

- Y: Kreditwürdigkeit,  
X: Personenmerkmale
- Y: Auftreten einer Krankheit innerhalb einer bestimmten Zeit,  
X: Exposition, Geschlecht, Alter etc.
- Y: Auffinden der korrekten Blüte,  
X: Zeit (Trend), Art (Fledermaus)
- Y: Präferenz für eine Partei,  
X: Persönlichkeitsmerkmale
- Y: Bestehen eines Tests,  
X: Lehrmethode, Geschlecht etc.

# Logistische Regression als Klassifikationsproblem

---

- Prognose in der Logistischen Regression entspricht Klassifikationsproblem mit 2 Gruppen
- Analogien zu Verfahren der Diskriminanzanalyse
- Diskriminanzregeln aus logistischer Regression möglich

# ML-Schätzung im logistischen Regressionsmodell

Sei das Modell (9.1)–(9.4) gegeben.

$$\hat{\beta}_{ML} := \arg \max L(\beta) = \arg \max \ln L(\beta) \quad (9.5)$$

$$L(\beta) = \prod_{i=1}^n G(x_i' \beta)^{Y_i} (1 - G(x_i' \beta))^{1 - Y_i} \quad (9.6)$$

$$\ln L(\beta) = \sum_{i=1}^n Y_i \ln(G(x_i' \beta)) + (1 - Y_i) \ln(1 - G(x_i' \beta)) \quad (9.7)$$

Ableiten nach  $\beta$  und Null setzen liefert unter Benutzung von  $G' = G(1 - G)$  die Score-Gleichungen für  $\hat{\beta}_{ML}$ .

$$s(\hat{\beta}_{ML}) := \sum_{i=1}^n (Y_i - G(x_i' \hat{\beta}_{ML})) x_i = 0. \quad (9.8)$$

# Eigenschaften des ML-Schätzers

Die allgemeine Theorie der Maximum-Likelihood-Schätzung liefert:  
Für  $n \rightarrow \infty$  gilt unter Regularitätsbedingungen:

$$\hat{\beta}_{ML} \rightarrow N(\beta, F^{-1}(\beta)) \quad (9.9)$$

$$F(\beta) = X' D(\beta) X \quad (9.10)$$

$$D(\beta) = \text{diag}\{(G(x_i' \beta)(1 - G(x_i' \beta)))\} \quad (9.11)$$

$$\hat{\beta}' F(\beta) \hat{\beta} \rightarrow \chi^2(p') \quad (9.12)$$

- Die asymptotische Varianzmatrix ergibt sich als Inverse der Fischer-Information (negative Ableitung der Score-Funktion)
- Die asymptotische Varianzmatrix entspricht auch der Varianzmatrix aus der gewichteten (heteroskedastischen) Regression, da  $\text{Var}(Y) = D(\beta) = \text{diag}(G(x_i' \beta)(1 - G(x_i' \beta)))$
- Die numerische Berechnung des ML-Schätzers erfolgt nach der Methode der „iterierten gewichteten kleinsten Quadrate“ (IWLS Iteratively Weighted Least Squares)

# Existenz und Eindeutigkeit des ML-Schätzers im logistischen Modell

---

## **Eindeutigkeit:**

Da die Likelihood-Funktion konkav ist, ist die Lösung der Score-Gleichung immer eindeutig

## **Existenz:**

Der ML- Schätzer existiert  $\iff$  Die Werte 0 und 1 sind nicht linear trennbar, d.h. es existiert kein  $\alpha$  mit  $Y = 1$  für  $x'\alpha > 0$  und  $Y = 0$  für  $x'\alpha < 0$

Im Fall der Nicht- Existenz geht mindestens eine Komponente gegen  $\infty$ .  
Im einfachen Modell bedeutet die Bedingung, dass  $Y=1$  für  $x > c$  und  $Y=0$  für  $x < c$ .

# Inferenz im logistischen Regressionsmodell

---

## Beachte:

Alle Aussagen gelten - im Gegensatz zum linearen Regressionsmodell - nur asymptotisch, d.h. für hinreichend große Stichprobenumfänge!

## Wald-Konfidenzintervalle:

Wir benutzen die asymptotische Normalität und erhalten folgende Konfidenzintervalle für  $\beta$  zum Niveau  $\alpha$  :

$$\hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} z_{1-\alpha/2}$$

$$\hat{\sigma}_{\hat{\beta}_k} = \sqrt{c_{kk}} \text{ (k-tes Diagonalelement der Matrix } F^{-1}(\hat{\beta}))$$

Für die Odds-ratios  $\exp(\beta_k)$  ergibt sich das transformierte Konfidenzintervall zum Niveau  $\alpha$  :

$$\exp \left[ \hat{\beta}_k \pm \hat{\sigma}_{\hat{\beta}_k} z_{1-\alpha/2} \right]$$

Da kein Varianzparameter zu schätzen ist, kommt die t-Verteilung hier nicht vor.

# Wald-Test für die lineare Hypothese

---

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben.

$H_0 : A\beta = c$  mit  $\text{rg}(A) = a$ .

Analog zum linearen Modell wird folgende quadratische Form betrachtet:

$$W = (A\hat{\beta} - c)'(AF^{-1}(\hat{\beta})A')^{-1}(A\hat{\beta} - c)$$

$W$  heißt Wald-Statistik.

Aus der asymptotischen Normalität folgt unmittelbar:

$$W \underset{\sim}{\text{as}} \chi^2(a)$$

Mit dieser Statistik lässt sich die allgemeine lineare Hypothese testen.

# Likelihood-Quotienten-Test für die lineare Hypothese

Sei das logistische Regressionsmodell (9.1)–(9.4) gegeben.

$H_0 : A\beta = c$  mit  $\text{rg}(A)=a$ .

Wir definieren folgende Teststatistik:

$$LQ = -2 \left\{ \ln L(\hat{\hat{\beta}}) - \ln L(\hat{\beta}) \right\}$$

$\hat{\hat{\beta}}$  : ML-Schätzer unter  $H_0$

Aus der allgemeine Theorie von Likelihood-Quotienten-Tests folgt:

Es gilt unter  $H_0$ :

$$LQ \underset{\sim}{\stackrel{\text{as}}{\sim}} \chi^2(a)$$

**Beachte:** Der LQ-Test ist mit dem Wald-Test für endliche Stichproben nicht äquivalent. Äquivalenz gilt nur asymptotisch.

# Likelihood-Quotienten-Konfidenzintervalle

---

Mit Hilfe des LQ-Tests lassen sich auch Konfidenzintervalle zum Niveau  $\alpha$  konstruieren:

$KI := \{\tilde{\beta}_k \mid H_0 : \beta_k = \tilde{\beta}_k \text{ wird mit LQ-Test zum Niveau } \alpha \text{ nicht abgelehnt}\}$

# Devianz im logistischen Modell

---

Analog zur ANOVA-Tafel betrachtet man im logistischen Modell die Log-Likelihood als Maß für die Modellgüte: Dabei wird definiert:

Modell mit Konstante (SST):	$P(Y_i = 1) = G(\beta_0)$
Modell (SSE):	$P(Y_i = 1) = G(x_i' \beta)$
"volles Modell"	$P(Y_i = 1) = p_i$

Von diesen Modellen wird jeweils der Wert von  $-2 \log(L)$  verglichen.

# Das logistische Modell für gruppierte Daten I

---

Wir betrachten das logistische Regressionsmodell mit **gruppierten** Daten:

Jeweils  $n_j$  Datenpunkte werden zu einer Gruppe zusammengefasst. Dabei sind in einer Gruppe die Kovariablen identisch.

Sei  $\hat{\pi}_j := G(x_j' \hat{\beta})$ .

$Y_j$  : Anzahl der Erfolge in Gruppe  $j$ .

Das Modell ist dann:

$$Y_j | x_j \sim B(n_j, G(x_j' \beta)), \quad j = 1, \dots, g \quad (9.13)$$

# Das logistische Modell für gruppierte Daten II

---

$$D = -2 \sum_{j=1}^g (\ln L(\hat{\beta}) - \ln L(y_j)) \quad (9.14)$$

heißt **Devianz**.

Es gilt:

$$D = 2 \sum_{j=1}^g y_j \ln \frac{y_j/n_j}{G(x_j' \hat{\beta})} + (n_j - y_j) \ln \frac{(n_j - y_j)/n_j}{(1 - G(x_j' \hat{\beta}))} \quad (9.15)$$

# Anpassungstests

---

## a) Pearson-Statistik

$$\chi_P^2 = \sum_{j=1}^g n_j \frac{(y_j/n_j - G(x' \hat{\beta}))^2}{\hat{\pi}_j(1 - \hat{\pi}_j)}$$

## b) Devianz (siehe (9.14)) Verteilungsapproximation ( $n_i/n \rightarrow \lambda_i$ )

$$\chi_p^2, D, \stackrel{(a)}{\sim} \chi^2(g - p')$$

## c) Bei kleinen Gruppenumfängen oder im Fall $n_i = 1$ :

### Hosmer-Lemeshow-Test

Bilde ca.  $g = 10$  Gruppen nach der Größe des linearen Prädiktors  $x' \hat{\beta}$  und bilde Anpassungsstatistik wie unter a). Die Testverteilung ist ein  $\chi^2(g - 2)$ -Verteilung.

# Residualanalyse I

---

Wir betrachten wie oben das logistische Regressionsmodell mit gruppierten Daten. Sei  $\hat{\pi}_j := G(x_j' \hat{\beta})$ .

## a) Devianz-Residuen

$$d_j = \text{sign}(y_j - n_j \hat{\pi}_j) \sqrt{y_j \ln \frac{y_j/n_j}{\hat{\pi}_j} + (n_j - y_j) \ln \frac{(n_j - y_j)/n_j}{(1 - \hat{\pi}_j)}} \quad (9.16)$$

## b) Pearson-Residuen

$$r_j = \frac{y_j - n_j \hat{\pi}_j}{\sqrt{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \quad (9.17)$$

# Residualanalyse II

## c) Standardisierung der Residuen

$$H := D^{\frac{1}{2}} X (X' D X)^{-1} X' D^{\frac{1}{2}} \quad (9.18)$$

$$D = \text{diag}(n_j \hat{\pi}_j (1 - \hat{\pi}_j)) \quad (9.19)$$

$$d_j^* := d_j / \sqrt{1 - h_{jj}} \quad (9.20)$$

$$r_j^* := r_j / \sqrt{1 - h_{jj}} \quad (9.21)$$

## d) Likelihood-Residuen

$$lr_j := \text{sign}(y_j - n_j G(x_j' \beta)) \sqrt{2(\ln L(\tilde{\beta}, \hat{\gamma}_j) - \ln L(\hat{\beta}))} \quad (9.22)$$

$L(\tilde{\beta}, \hat{\gamma}_j)$  : Likelihood des Modells mit dem der zusätzlichen Indikatorvariablen für die Beobachtung  $j$  mit zugehörigem Parameter  $\gamma_j$ .

# ROC-Kurve

---

Allgemein

$Y = 1 \longrightarrow$  Ausfall (krank)

$Y = 0 \longrightarrow$  kein Ausfall (gesund)

In der medizinischen Literatur ist das Testergebnis  $m$ :

$$\hat{Y}_i = 1 \Leftrightarrow m_i \geq c \quad (9.23)$$

In der Literatur zum Kreditrisiko ist der Score  $s$ :

$$\hat{Y}_i = 1 \Leftrightarrow s_i \leq c \quad (9.24)$$

$c$  ist dabei ein Grenzwert. Beide Ansätze sind offensichtlich äquivalent:

Betrachte dazu  $m_i = -s_i$

# Sensitivität und Spezifität

---

**Richtig Positiv = Sensitivität:**

$$(m) \quad P(\hat{Y} = 1|Y = 1) = P(m \geq c|Y = 1) = S_1(c) \quad (9.25)$$

$$(k) \quad P(\hat{Y} = 1|Y = 1) = P(s \leq c|Y = 1) = F_1(c) \quad (9.26)$$

$S_1(c)$  stellt die Survivorfunktion dar,  $F_1(c)$  die Verteilungsfunktion.

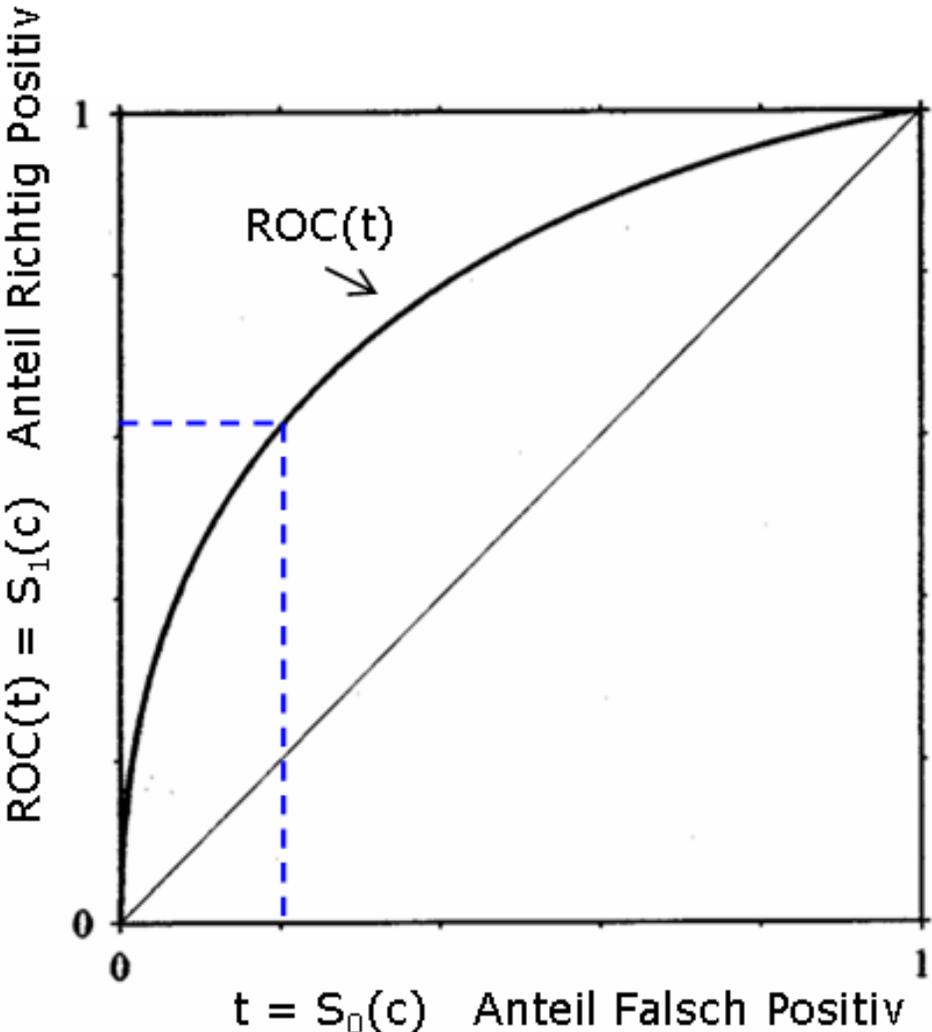
**Falsch Positiv = 1- Spezifität:**

$$(m) P(\hat{Y} = 1|Y = 0) = P(m \geq c|Y = 0) = S_0(c) \quad (9.27)$$

$$(k) P(\hat{Y} = 1|Y = 0) = P(s \leq c|Y = 0) = F_0(c) \quad (9.28)$$

Die ROC-Kurve besteht aus den Punkten  $(S_0(c), S_1(c))$  bzw.  $(F_0(c), F_1(c))$ .

# Beispiel für ROC-Kurve



# Logistische Regression und ROC- Kurve

---

$$m_i = G(x_i' \beta)$$

$$m_i = x_i' \beta$$

Beachte die Invarianz der ROC Kurve bzgl. monotoner Funktionen! Die Verteilung von  $F_0, F_1$  bzw.  $S_0, S_1$  kann aus den Daten geschätzt werden  
 $\Rightarrow$  ROC-Kurve

Alternative: Schätze  $F_0, F_1$  bzw.  $S_0, S_1$  aus Validierungsdaten.

# Maß zur Bewertung der Kurve: AUC

---

$$AUC = \int_{t=0}^1 ROC(t) dt \quad (9.29)$$

Dies stellt die Fläche unter der Kurve dar.

Es gilt:

$$AUC = P(m_1 \leq m_0) \quad (9.30)$$

$m_1$  ist dabei aus der Verteilung  $m|Y = 1$

$m_0$  ist dabei aus der Verteilung  $m|Y = 0$

Daher ist das empirische AUC:

$$\widehat{AUC} = \frac{N_c}{N} \quad (9.31)$$

Dabei bezeichnet  $N_c$  die Anzahl der konkordanten Paare und  $N$  die Anzahl der Paare mit unterschiedlichem  $Y$ .

Mit  $N_d$  gleich der Anzahl der diskordanten Paare ist der GINI dann:

# GINI- Koeffizient

---

Normierte Fläche zwischen Winkelhalbierender und ROC- Kurve

$$GINI = 2 \cdot \left( AUC - \frac{1}{2} \right) = 2 \cdot AUC - 1 \quad (9.32)$$

$$\widehat{GINI} = \frac{N_c - N_d}{N} \quad (9.33)$$

Der empirische GINI entspricht dem Somers D.

# Die logistische Regression für Fall-Kontroll-Studien

---

Sei in der Grundgesamtheit folgende Beziehung gegeben:

$$P_0(Y = 1|X = x) = G(\alpha + \beta x) \quad (9.34)$$

mit  $G(t) = (1 + \exp(-t))^{-1}$

X: Exposition

Y: Erkrankung

Es wird nun aus der GG gezogen:

$n_1$  Fälle ( $Y = 1$ ) und  $n_2$  Kontrollen ( $Y = 0$ )

Gesucht ist

$$P_S(Y = 1|X = x) \quad (9.35)$$

Mit  $P_S$  werden die Wahrscheinlichkeiten (Dichten) in der Stichprobe bezeichnet, mit  $P_0$  die in der GG.

# Berechnung von $P_S$ I

$$\begin{aligned} P_S(Y = 1|X = x) &= \frac{P_S(Y = 1, X = x)}{P_S(X = x)} = \\ &= \frac{P_S(Y = 1)P_S(X = x|Y = 1)}{P_S(Y = 1)P_S(X = x|Y = 1) + P_S(Y = 0)P_S(X = x|Y = 0)} = \\ &= \frac{c_1 P_0(X = x|Y = 1)}{c_1 P_0(X = x|Y = 1) + c_2 P_0(X = x|Y = 0)} \end{aligned}$$

Die letzte Identität gilt wegen

$$P_0(X = x|Y = 1) = P_S(X = x|Y = 1) \quad (9.36)$$

und mit  $c_1 = \frac{n_1}{n_1 + n_2}$  und  $c_2 = \frac{n_2}{n_1 + n_2}$

$$c_1 P_0(X = x|Y = 1) = \frac{c_1}{P_0(Y = 1)} P_0(X = x) P_0(Y = 1|X = x)$$

$$c_2 P_0(X = x|Y = 0) = \frac{c_2}{P_0(Y = 0)} P_0(X = x) P_0(Y = 0|X = x)$$

# Berechnung von $P_S$ II

Mit  $\frac{c_1}{P_0(Y=1)} = d_1$  und  $\frac{c_2}{P_0(Y=0)} = d_2$  folgt:

$$\begin{aligned} P_S(Y=1|X=x) &= \frac{d_1 P(Y=1|X=x)}{d_1 P(Y=1|X=x) + d_2 (1 - P(Y=1|X=x))} \\ &= \frac{1}{1 + \frac{d_2}{d_1} \frac{1 - G(\alpha + \beta x)}{G(\alpha + \beta x)}} \\ &= \frac{1}{1 + \exp(-\alpha - \ln(\frac{d_1}{d_2}) - \beta x)} \end{aligned}$$

Die letzte Gleichung folgt aus  $\frac{1 - G(t)}{G(t)} = \frac{1}{G(t)} - 1 = \exp(-t)$ .

# Berechnung von $P_S$ III

---

Insgesamt gilt:

$$P_S(Y = 1|X = x) = G\left(\alpha + \ln\left(\frac{d_1}{d_2}\right) + \beta x\right) \quad (9.37)$$

Die Auswertung bezüglich  $\beta$  kann also durch eine logistische Regression erfolgen. Der Parameter  $\alpha$  in der Grundgesamtheit kann dabei nicht geschätzt werden.