Statistische Modellierung III

-Binäre Regression-

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2019/2020

Binäre Regressionsmodelle

Binäre Regression - Setup

- Y nun nicht mehr metrisch, sondern binär (z.B. „Erfolg/Misserfolg", „ja/nein")
- Codierung mit 0 oder 1, das heißt $Y \in \{0, 1\}$
- Stoch. unabhängige Beobachtungen Y_i, $1, \ldots, n$, Kovariablen $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$
- Interessieren uns nun für die Wahrscheinlichkeiten

$$\pi_i = \mathbb{P}(Y_i = 1|\mathbf{x}_i) = E(Y_i|\mathbf{x}_i)$$

- Da $\pi_i \in (0, 1)$ ist ein herkömmliches lineares Modell nicht sinnvoll

Response- und Linkfunktion

- Verwenden das Regressionsmodell

$$\pi_i = h(\eta_i) \quad \text{mit } \eta_i = \mathbf{x}_i\beta,$$

wobei $h : \mathbb{R} \longrightarrow [0, 1]$ eine streng monoton wachsende Verteilungsfunktion ist

- Man nennt h die *Responsefunktion*
- Bezeichne mit $g = h^{-1}$ die zugehörige Umkehrfunktion, dann folgt:

$$g(\pi_i) = \eta_i = \mathbf{x}_i\beta$$

- Man nennt g die *Linkfunktion*

Spezielle binäre Regressionsmodelle

Das Logit-Modell

- Im Logit-Modell verwendet man die sogen. *logistische Responsefunktion*

$$\pi = h(\eta) = \frac{e^\eta}{1 + e^\eta}$$

- Diese ist die Verteilungsfunktion der logistsichen Verteilung
- Damit erhält man die sog. *Logit-Linkfunktion*

$$g(\pi) = \log \frac{\pi}{1 - \pi} = \text{logit}(\pi) = \eta = \mathbf{x}\beta = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Das Logit-Modell

- Für die Odds von π ergibt sich damit

$$\frac{\pi}{1-\pi} = e^{\mathbf{x}\beta} = e^{\beta_1} e^{\beta_2 x_2} \ldots e^{\beta_k x_k}$$

- Damit sind die Kovariableneffekte von exponentiell-multiplikativer Form
- Die logistische Verteilungsfunktion ist drehsymmetrisch um $(0, 1/2)$, das heißt es gilt $h(-\eta) = 1 - h(\eta)$
- Damit gilt für die Logit-Funktion $g(1-\pi) = -g(\pi)$
- Es ist also relativ egal, ob wir π oder $1-\pi$ modellieren

Das Logit-Modell

- Modelliert man $1-\pi$ statt π, so erhält man dieselben absoluten Regressionskoeffizienten, allerdings haben sie ein umgekehrtes Vorzeichen
- Bei kleinem π ist $1-\pi \approx 1$ und damit die Odds $\frac{\pi}{1-\pi} \approx \pi$
- $\pi \approx e^{\mathbf{x}\beta}$ ist dann ein multiplikatives Modell für die Wahrscheinlichkeit π

Das Probit-Modell

- Im Probit-Modell verwendet man die Responsefunktion die Verteilungsfunktion der Standardnormalverteilung, also $h = \Phi$
- Damit ist $\pi = \Phi(\eta) = \Phi(\mathbf{x}\beta) = \Phi(\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k)$
- Die Linkfunktion $g = h^{-1} = \Phi^{-1}$ ist die Quantilfunktion von $N(0,1)$ und wird *Probit-Transformation* genannt
- Es gilt wieder

$$\Phi(-\eta) = 1 - \Phi(\eta) \quad \text{und} \quad g(1-\pi) = \Phi^{-1}(1-\pi) = -\Phi^{-1}(\pi) = -g(\pi)$$

- Es ist also wieder egal ob π oder $1 - \pi$ modelliert wird

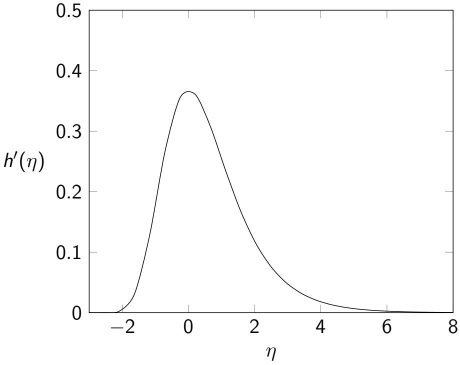
Das Log-Log-Modell

- Im Log-Log-Modell verwendet man die Extremwertverteilung (Standard-Gumbel-Verteilung) für die Responsefunktion:

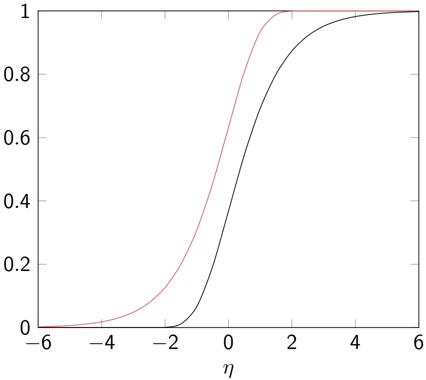
$$h(\eta) = \exp\{-\exp(-\eta)\}, \text{ für } \eta \in \mathbb{R}.$$

- Dies führt zur Linkfunktion $g(\pi) = -\log(-\log(\pi))$, für $0 < \pi < 1$
- Das Log-Log-Modell ist für spezielle Anwendungen interessant, z.B. zur Modellierung von Versagens- oder Schadenswahrscheinlichkeiten.
- Nun gilt **nicht** mehr $h(-\eta) = 1 - h(\eta)$ und entsprechend $g(\pi) \neq 1 - g(\pi)$
- $g(\pi)$ verhält sich deutlich unterschiedlich an den Grenzen 0 und 1, da g unterschiedlich schnell gegen 0 bzw. 1 geht. Die Koeffizienten für $1 - \pi$ sind daher nicht einfach das Negative der Koeffizienten für π.

Das Log-Log-Modell



Dichte der Responsefunktion h



Schwarz: $h(\eta)$, Rot: $1 - h(-\eta)$

Interpretation der Koeffizienten im Log-Log-Modell

- Bei einer Erhöhung von x_j um eine Einheit ändert sich η zu $\eta + \beta_i$
- π verändert sich damit zu

$$\begin{aligned} h(\eta + \beta_j) &= \exp\{-\exp\{-\eta - \beta_j\}\} = \exp\{-\exp(-\eta) \cdot \exp(-\beta_j)\} \\ &= \exp\{-\exp(-\eta)\}^{\exp(-\beta_j)} = \pi^{\exp(-\beta_j)} \gtrless \pi \begin{cases} \beta_j > 0 \\ \beta_j < 0 \end{cases} \end{aligned}$$

- Der Effekt der j-ten Kovariable ist also eine Potenzierung von π

Zusammenhang zwischen den Modellen

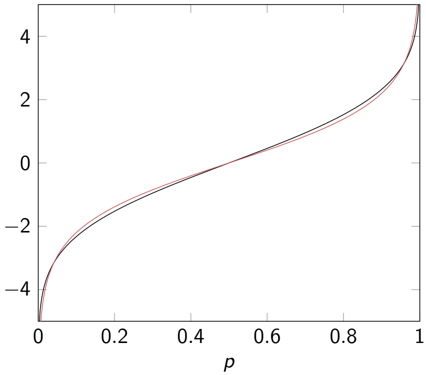
- Erwartungswerte und Varianzen der verschiedenen Verteilungen der Responsefunktionen unterscheiden sich
- Für bessere Vergleichbarkeit werden diese hier so standardisiert, dass sie gleichen Erwartungswert und gleiche Varianz haben
- Betrachten z.B. das skalierte Probit-Modell:

$$\pi = \Phi(\mathbf{X}\beta\sigma^{-1}) = \Phi_\sigma(\mathbf{X}\beta)$$

- Mit $\sigma^2 = Var(\text{Logit-Verteilung}) = \pi^2/3$ haben Φ_σ und die Logit-Verteilung indentische erste und zweite Momente

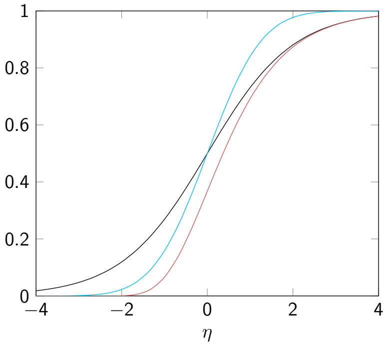
Vergleich Logit- zu Probit-Modell

- Mit dieser Standardisierung erhält man die rechts abgebildeten Linkfunktionen
- Logit- und Probit-Link unterscheiden sich im Bereich 0.02 bis 0.98 nur sehr gering
- Also keine großen Unterschiede, wenn π nicht zu klein oder zu groß ist
- Wegen besserer Interpretierbarkeit in der Praxis häufig Logit bevorzugt

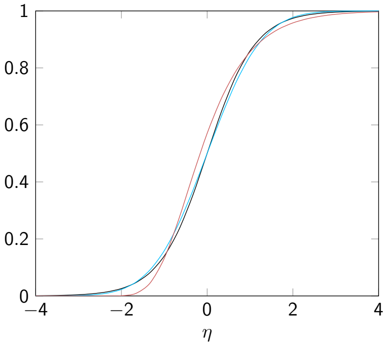


Schwarz: skalierter Probit-Link, Rot: Logit-Link

Zusammenhang zwischen den Modellen



Responsefunktionen des Logit-Modells, des Probit-Modells und des Log-Log-Modells



Standardisierte Responsefunktionen des Logit-Modells, des Probit-Modells und des Log-Log-Modells

Zusammenhang zwischen den Modellen

- Das Verhältnis β_i/β_j für $2 \leq i \neq j \leq k$ ist für alle Modelle identisch
- Vorsicht bei der direkten Interpretation von β_j und der Konstante β_1
- Hängen stark von der Wahl der Response- bzw. der Link-Funktion ab
- Können, wie eben illustriert, so umgerechnet werden, dass sie ähnlich sind

Gruppierte Daten

Gruppierte Binärdaten

- Wieder m Gruppen $G_1, \ldots, G_m$ mit $\mathbf{x}_i = \mathbf{x}_j = \mathbf{z}_l \in \mathbb{R}^k$ für alle i, j in Gruppe G_l
- Y_l nun Zahl der Erfolge in Gruppe l, also $1 \leq Y_l \leq n_l$
- Betrachten auch Mittelwerte $\bar{Y}_l = Y_l / n_l$
- Können Daten wie folgt zusammenfassen:

Gruppe	*Fallzahl*	*absolut*	*Kovariablen*
1	n_1	Y_1	$z_{l1} \ldots z_{lk}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	n_m	Y_m	$z_{m1} \ldots z_{mk}$

Gruppierte Binärdaten

- Sind Einzelbeobachtungen $B(1, \pi_l)$ verteilt, so gilt $Y_l \sim B(n_l, \pi_l)$
- Also $E(Y_l) = n_l \pi_l$ und $Var(Y_l) = n_l \pi_l (1 - \pi_l)$
- Damit gilt auch $\bar{Y}_l \sim B(n_l, \pi_l)/n_l$ mit $E(\bar{Y}_l) = \pi_l$ und $Var(\bar{Y}_l) = \pi_l(1 - \pi_l)/n_l$
- Für π_l können also die gleichen Regressionsansätze (Logit-, Probit- und Log-Log-Modell) wie vorher verwendet werden

Beispiel: Kaiserschnittgeburten

- Betrachten Datensatz zu Infektionen von Müttern nach Kaiserschnitt-Geburten im Lehrzentrum der Uni Münster
- Untersucht wurden $n = 251$ Frauen
- Kovariablen:

$$\text{NPLAN} = \begin{cases} 1, & \text{Kaiserschnitt nicht geplant} \\ 0, & \text{Kaiserschnitt geplant} \end{cases}$$

$$\text{RISK} = \begin{cases} 1, & \text{Risiko-Faktoren vorhanden} \\ 0, & \text{Risiko-Faktoren nicht vorhanden} \end{cases}$$

$$\text{ANTIB} = \begin{cases} 1, & \text{Antibiotika verabreicht (als Prophylaxe)} \\ 0, & \text{Antibiotika nicht verabreicht} \end{cases}$$

Beispiel: Kaiserschnittgeburten

		Kaiserschnitt geplant		Kaiserschnitt nicht geplant	
		Infektion		Infektion	
		ja	nein	ja	nein
Antibiotika	Risikofaktor	1	17	11	87
	kein Risikofaktor	0	2	0	0
keine Antibiotika	Risikofaktor	28	30	23	3
	kein Risikofaktor	8	32	0	9

Aus Fahrmeir, Kneib und Lang (2009)

Beispiel: Kaiserschnittgeburten

- Mit dem Logit-Modell schätzen wir das folgende Modell für die Log-Odds einer Infektion:

$$\log \frac{P(\text{Infektion})}{1 - P(\text{Infektion})} = -1.89 + 1.07\text{NPLAN} + 2.03\text{RISK} - 3.25\text{ANTIB}$$

- Herleitung der Schätzer in nächstem Abschnitt

- Das Ergebnis kann z.B. wie folgt interpretiert werden: $e^{2.03}$ ist der Faktor, um den sich die Odds vergrößert, wenn Risikofaktoren vorhanden sind (RISK=1).

ML-Schätzung im Logit-Modell: Die logistische Regression

Likelihood bei Binärdaten

- Bei stochastisch unabhängigen Beobachtungen haben die Daten die Likelihood

$$L(\beta) = \Pi_{i=1}^{n} f(Y_i|\mathbf{x}_i\beta_i)$$

wobei $f(\mathbf{Y}_i|\mathbf{x}_i\beta_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$ die (bedingte) Dichte von Y_i mit $\pi_i = E(Y_i|\mathbf{X}_i) = h(\mathbf{x}_i\beta)$ ist

- Damit gilt für die Log-Likelihood:

$$l(\beta) = \sum_{i=1}^{n} \log f(Y_i|\mathbf{x}_i\beta_i) = \sum_{i=1}^{n} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right\}$$

- Also $l(\beta) = \sum_{i=1}^{n} l_i(\beta)$ mit $l_i(\beta) = Y_i \log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i)$

Likelihood im Logit-Modell

- Betrachten nun Logit-Modell, d.h. $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i\beta = \eta_i$
- In diesem Modell ist

$$\log(1-\pi_i) = \log\left(1 - \frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}\right) = \log\left(\frac{1}{1+e^{\mathbf{x}_i\beta}}\right) = -\log(1+e^{\mathbf{x}_i\beta})$$

- Also

$$l_i(\beta) = Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) = Y_i(\mathbf{x}_i\beta) - \log(1+e^{\mathbf{x}_i\beta}) = Y_i\eta_i - \log(1+e^{\eta_i})$$

- Insgesamt erhalten wir

$$l(\beta) = \sum_{i=1}^{n} Y_i\eta_i - \log(1+e^{\eta_i})$$

Score-Funktion im Logit-Modell

- Um ML-Schätzer von β zu finden maximieren wir die Log-Likelihood $l(\beta)$
- Setzen dazu alle partiellen Ableitungen von $l(\beta)$ (nach den β_j) gleich Null
- Vektor der partiellen Ableitungen ist k-dimensionale Funktion in β und wird *Score* genannt. Er ist gegeben durch

$$\mathbf{s}(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i^T (Y_i - \pi_i)$$

- Im Logit-Model ist also folgendes Gleichungssystem zu lösen:

$$\mathbf{s}(\hat{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i^T \left(Y_i - \frac{e^{\mathbf{x}_i \hat{\beta}}}{1 + e^{\mathbf{x}_i \hat{\beta}}} \right) = \mathbf{0}$$

- Lösung numerisch iterativ über Fisher-Scoring oder Newton-Raphson Verfahren

Informationsmatrix

- Für spätere betrachtungen notwendig: Die *beobachtete Informationsmatrix*

$$\mathbf{H}(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = -\left(\frac{\partial^2 l(\beta)}{\partial\beta_i\partial\beta_j}\right)_{i,j}$$

- Und die *erwartete Informationsmatrix*, auch *Fisher-Matrix* genannt

$$\mathbf{F}(\beta) = E[\mathbf{H}(\beta)] = E[\mathbf{s}(\beta)\mathbf{s}(\beta)^T]$$

- Man kann zeigen, dass

$$\mathbf{F}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \pi_i (1-\pi_i) = \left(\sum_{i=1}^{n} x_{il} x_{im} \pi_i (1-\pi_i)\right)_{l,m=1,\dots,k}$$

- $\mathbf{F}$ hängt von β ab, weil $\pi_i = h(\mathbf{x}_i\beta)$

Informationsmatrix

- Im Logit-Modell gilt

$$\mathbf{H}(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T} = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \pi_i (1-\pi_i) = \mathbf{F}(\beta)$$

- $\mathbf{H}(\beta) = \mathbf{F}(\beta)$ gilt nur für das Logit-Modell und nicht für das Probit- oder das Log-Log-Modell.

- Üblicherweise verwendet man die Fisher-Matrix $\mathbf{F}(\beta)$ und nicht das beobachtete $\mathbf{H}(\beta)$. Die Fisher-Matrix ist in der Regel leichter zu bestimmen.

Gruppierte Daten

- Bei den gruppierten Daten ($Y_l \sim B(n_l, \pi_l)$) ist der Log-Likelihood-Kern (Log-Likelihood ohne Terme die unabhängig von π_l sind) gegeben als:

$$\sum_{l=1}^{m} \left\{ Y_l \log(\pi_l) - Y_l \log(1-\pi_l) + n_l \log(1-\pi_l) \right\}$$

- Bei gruppierten Daten also kein Unterschied ob Log-Likelihood der Einzeldaten oder Gruppendaten maximiert wird
- Der Score ergibt sich zu $\mathbf{s}(\beta) = \sum_{l=1}^{m} \mathbf{z}_l^T (Y_l - n_l \pi_l) = \sum_{l=1}^{m} n_l \mathbf{z}_l^T (\bar{Y}_l - \pi_l)$
- Die Fisher Matrix ergibt sich zu $\mathbf{F}(\beta) = \sum_{l=1}^{m} \mathbf{z}_l^T \mathbf{z}_l n_l \pi_l (1-\pi_l)$

Einschub: Newton-Raphson-Verfahren

Newton-Raphson-Verfahren

- Zum Finden der Nullstellen des Scores kann das Newton-Raphson-Verfahren benutzt werden
- Verfahren kann allgemein für jede zweimal stetig differenzierbare reelle Funktion $f : \mathbb{R} \longrightarrow \mathbb{R}$ mit eindeutiger Nullstelle x^* angewendet werden
- Nehmen an, dass $f'(x^*) \neq 0$

Newton-Raphson-Verfahren - Algorithmus

- Betrachten die Tangentenfunktion von f an einer Stelle $x^{(0)}$ (Startwert),

$$Tf(x; x^{(0)}) = f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$$

- Nullstelle der Tangentenfunktion ist gegeben durch:

$$0 = Tf(x^{(1)}; x^{(0)}) = f(x^{(0)}) + f'(x^{(0)})(x^{(1)} - x^{(0)}) \quad \Longrightarrow \quad x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}$$

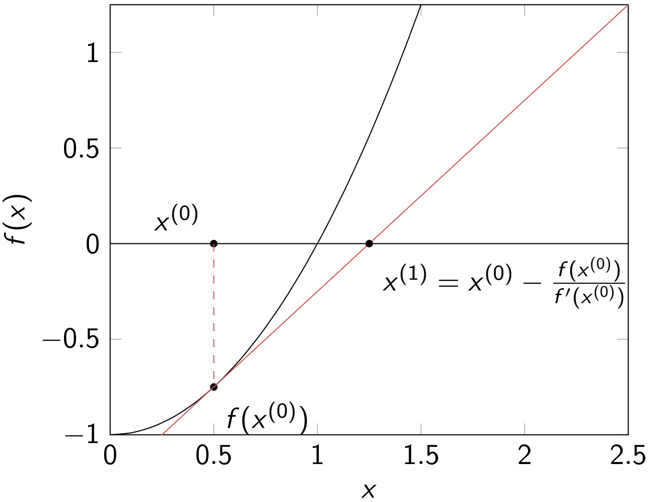
- Ersetze nun x^0 durch x^1 und wiederhole den Vorgang:

$$x^{(2)} = x^{(1)} - \frac{f(x^{(1)})}{f'(x^{(1)})}$$

- Im $k+1$-ten Schritt berechnen wir also

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

Newton-Raphson-Verfahren



Newton-Raphson-Verfahren - Eigenschaften

- Man kann zeigen, dass wenn der Startwert $x^{(0)}$ nahe genug an der (unbekannten) Nullstelle x^* von f liegt, $x^{(k)}$ gegen x^* konvergiert
- **Problem:** Nullstelle x^* und wie nahe $x^{(0)}$ bei x^* sein muss sind unbekannt
- Wahl des Startwerts kann kritisch sein, d.h. der Algorithmus konvergiert bei falschem Startwert nicht zu x^*

Newton-Raphson-Verfahren - mehrdimensional

- Die Score-Funktion ist bekanntlich mehrdimensional $\mathbf{s} : \beta \in \mathbb{R}^k \mapsto \mathbb{R}^k$
- Können das Newton-Raphson-Verfahren auf diesen Fall verallgemeinern
- Das iterative Newton-Verfahren liefert die Rekursion

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \mathbf{H}^{-1}(\hat{\beta}^{(k)})\mathbf{s}(\hat{\beta}^{(k)}).$$

- Meist wird $\mathbf{H}(\beta)$ durch die Fisher-Matrix ersetzt. Das liefert das Fisher-Scoring-Verfahren:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \mathbf{F}^{-1}(\hat{\beta}^{(k)})\,\mathbf{s}(\hat{\beta}^{(k)}).$$

- Stoppen in beiden Verfahren, wenn $||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||/||\hat{\beta}^{(k)}|| \leq \epsilon$ für ein kleines ϵ (z.B. 10^{-6})

Newton-Raphson-Verfahren - Bemerkungen

- Im Logit-Modell ist $\mathbf{F} = \mathbf{H}$. Damit sind Newton-Raphson und Fisher-Scoring-Verfahren identisch
- Damit das Fisher-Scoring-Verfahren angewendet werden kann, muss $\mathbf{F}(\beta)$ für β invertierbar sein. Da für das Logit-Modell $\mathbf{F}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \pi_i (1 - \pi_i)$ gilt, folgt das bereits aus Rang$(\mathbf{X}^T\mathbf{X}) = k$ und $0 < \pi_i < 1$ für alle i
- In der Regel konvergiert der Fisher-Scoring-Algorithmus und stoppt nach wenigen Iterationsschritten in der Nähe des MLE
- Manchmal existiert der MLE nicht (die Likelihood hat also kein Maximum). Dann divergiert der Algorithmus, $||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||$ wird größer statt kleiner. Nicht-Existenz des MLE tritt in der Regel bei sehr ungünstigen Datenkonstellationen auf, vor allem, wenn n klein im Vergleich zu k ist
- In der Praxis sollte geprüft werden, ob $||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||$ tatsächlich fällt oder steigt

Beispiel: Infektionen nach Kaiserschnittgeburten

- Hatten bereits das lineare Modell ohne Interaktionsterme betrachtet, also

$$\eta = \beta_0 + \beta_1 \text{NPLAN} + \beta_2 \text{RISK} + \beta_3 \text{ANTIB}.$$

- Bei Hinzunahme von der Interaktion zwischen RISK und ANTIB oder NPLAN und RISK tritt das Problem der Divergenz auf

- Grund: In den Gruppen

$$\text{NPLAN} = 0,\ \text{RISK} = 0,\ \text{ANTIB} = 1 \text{ und } \text{NPLAN} = 1, \text{RISK} = 0, \text{ANTIB} = 0$$

  treten jeweils keine Interaktionen auf

- Die MLEs für die entsprechenden Terme sind unbeschränkt, da das Risiko für eine Infektion in diesen Gruppen auf Null geschätzt wird

Eigenschaften der MLEs

Asymptotische Eigenschaften der MLEs

Unter realtiv schwachen Annahmen kann man zeigen:

- Der MLE $\hat{\beta}$ existiert
- $\mathbf{F}(\beta)^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{I})$
- $\mathbf{F}(\hat{\beta})^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{I})$
- Also kann Verteilung von $\hat{\beta}$ durch $N(\beta, \mathbf{F}^{-1}(\hat{\beta}))$ approximiert werden
- Insbesondere $\hat{\beta}_j \overset{a}{\sim} N(\beta_j, a_{jj})$, wobei a_{jj} das j-te Diagonalelement von $\mathbf{F}^{-1}(\hat{\beta})$ ist

Testen linearer Hypothesen im Logit-Modell

Lineare Hypothesen

- Betrachten wieder Hypothesen der Form

$$H_0 : \mathbf{C}\beta = \mathbf{d} \quad \text{gegen} \quad H_1 : \mathbf{C}\beta \neq \mathbf{d},$$

  wobei $\mathbf{C}$ eine $(r \times k)$-Restriktionsmatrix mit $r < k$ und $\mathbf{d} \in \mathbb{R}^k$

- Werden 3 Methoden zum Testen dieser Hypothesen beschreiben

Lineare Hypothesen - Beispiele

(1) Globaler Test: Hier testen wir

$$H_0 : \beta_2 = \ldots = \beta_k = 0 \text{ vs. } H_1 : \beta_j > 0 \text{ für mindestens ein } j = 2, \ldots, k$$

Das entspricht einer linearen Hypothese mit Restriktionsmatrix $\mathbf{C}$ und $\mathbf{d}$

$$\mathbf{C} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ \vdots & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \ldots & \ldots & 1 \end{pmatrix}}_{(k-1)\times k\text{-dimensional}}, \qquad \mathbf{d} = \underbrace{\begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}}_{(k-1)-\text{dimensional}}.$$

Lineare Hypothesen - Beispiele

(2) $H_0 : \beta_j = 0$ gegen $H_1 : \beta_j \neq 0$ für ein festes j.

$$\mathbf{C} = e_j^T, \quad \mathbf{d} = 0,$$

wobei $e_j^T = (0, \ldots, 1, \ldots, 0)$. $\mathbf{C} = e_j^T$ ist damit eine $(1 \times k)$-dimensionale Matrix und $\mathbf{d} = 0$ entsprechend eindimensional.

(3) Es sei $S \subseteq \{2, \ldots, k\}$ eine Teilmenge von unabhängigen Variablen, $s = |S|$.

$$H_0 : \beta_j = 0 \text{ für alle } j \in S \quad \text{gegen} \quad H_1 : \beta_j \neq 0 \text{ für mindestens ein } j \in S$$

$\mathbf{C} = (\mathbf{e}_j^T)_{j \in S}$ ist die $s \times k$-dimensionale Matrix mit Zeilen $\mathbf{e}_j^T$ für $j \in S$. Entsprechend ist $\mathbf{d} = \mathbf{0} \in \mathbb{R}^s$.

Likelihood-Quotienten-Test

Likelihood-Quotienten-Test

- Sei $\hat{\beta}$ der MLE von β und $\tilde{\beta}$ der MLE von β mit Restriktion $\mathbf{C}\beta = \mathbf{d}$
- Bestimmung von $\tilde{\beta}$ hängt von $\mathbf{C}$ ab (z.B. Maximierung unter Nebenbedingungen)
- Bilde die *Likelihood-Quotienten-Statistik*:

$$LQ = \frac{L(\hat{\beta})}{L(\tilde{\beta})} = \frac{\max_{\beta'} L(\beta')}{\max_{\beta'', \mathbf{C}\beta''=\mathbf{d}} L(\beta'')}.$$

- Vergleichen also MLE ohne Restriktion mit dem MLE mit Restriktionen
- Da $\max_{\beta'} L(\beta') \geq \max_{\beta'', \mathbf{C}\beta''=\mathbf{d}} L(\beta'')$ ist $LQ \geq 1$
- Große Werte von LQ sprechen gegen H_0 (also für H_1), kleine Werte eher dafür

Likelihood-Quotienten-Test

- Gehen nun zum Logarithmus über und fügen noch den Faktor 2 hinzu:

$$lq = 2 \log LQ = 2\{l(\hat{\beta}) - l(\tilde{\beta})\} = -2\{l(\tilde{\beta}) - l(\hat{\beta})\}$$

- Man kann zeigen, dass unter H_0 gilt:

$$lq \overset{a}{\sim} \chi^2_r$$

- Dies führt zu einem χ^2-Test mit asymptotischen Signifikanzniveau α: Verwerfe H_0, falls $lq \geq Q_r^{\chi^2}(1-\alpha)$ oder (äquivalent) der p-Wert $p_{lq} = 1 - \chi^2_r(lq) \leq \alpha$

- Bei der allgemeinen linearen Regression gilt $lq = ||\hat{\mathbf{e}}_0||^2_{\mathbf{W}} - ||\hat{\mathbf{e}}||^2_{\mathbf{W}}$. Dies ist der Zähler der F-Statistik (bis auf den Quotienten $k - k_0$), der unter H_0 exakt χ^2_r-verteilt war, wenn $r = k - k_0$

Likelihood-Quotienten-Test - Bemerkungen

- In den Beispielen (1) bis (3) kann der MLE $\tilde{\beta}$ unter der Restriktion $\mathbf{C}\beta = \mathbf{d}$ sehr leicht bestimmt werden, da er jeweils einem Modell mit weniger Kovariablen entspricht:
  (1) nur Achsenabschnitt $\Longrightarrow \tilde{\beta}_1 = \frac{1}{n}\sum Y_i$, Gesamtanzahl der Erfolge,
  (2) Kovariable x_j weglassen, d.h. logistische Regression ohne $x_j \Longrightarrow \tilde{\beta}$,
  (3) alle Kovariablen x_j mit $j \in S$ weglassen $\Longrightarrow \tilde{\beta}$.

- Bei komplexerem $\mathbf{C}$ kann das Bestimmen von $\tilde{\beta}$ aufwendiger sein, z.B. wenn

$$H_0 : \pi(\mathbf{x}_0) = \pi_0 \quad \Longleftrightarrow \quad \mathbf{x}_0\beta = \eta_0 = \log \frac{\pi_0}{1-\pi_0},$$

  was dem $(1 \times k)$-dimensionalem $\mathbf{C} = \mathbf{x}_0$ und eindimensionalen $\mathbf{d} = \eta_0$ entspricht. Hier müssen wir bzgl. aller β' maximieren, die die eindimensionale Restriktion $\mathbf{x}_0\beta' = \eta_0$ erfüllen.

Wald-Test

Wald-Test

- Wir haben gesehen, dass

$$\hat{\beta} - \beta \overset{a}{\sim} N(\mathbf{0}, \mathbf{F}^{-1}(\hat{\beta}))$$

- Also gilt unter H_0

$$\mathbf{C}\hat{\beta} - \mathbf{d} = \mathbf{C}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, \mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T).$$

- Hieraus folgt unter H_0

$$W = (\mathbf{C}\hat{\beta} - \mathbf{d})^T[\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\beta} - \mathbf{d}) \overset{a}{\sim} \chi^2_r,$$

weil aus $\mathbf{Z} \sim N(\mathbf{0}, \Sigma)$ folgt, dass $\mathbf{Z}^T\Sigma^{-1}\mathbf{Z} \sim \chi^2_r$, wenn $\mathrm{Rang}(\Sigma) = r$.

Wald-Test

- Wir erhalten somit einen asymptotischen Niveau α-Test, wenn wir H_0 verwerfen, falls $W \geq Q_r^{\chi^2}(1-\alpha)$ oder der p-Wert$= 1 - \chi_r^2(W) \leq \alpha$.
- Im Gegensatz zum $LQ-$Test muss beim Wald-Test der MLE $\tilde{\beta}$ unter der Restriktion in der Nullhypothese nicht bestimmt werden
- Der $LQ-$Test ist in der Regel exakter, da er das Niveau besser einhält

Score-Test

Score-Test

- Man kann zeigen, dass $E[\mathbf{s}(\beta)] = 0$ für das wahre β
- Für den MLE $\hat{\beta}$ gilt nach Konstruktion: $\mathbf{s}(\hat{\beta}) = \mathbf{0}$
- Für den restringierten MLE (unter $H_0 : \mathbf{C}\beta = \mathbf{d}$) $\tilde{\beta}$ ist im Allgemeinen $\mathbf{s}(\tilde{\beta}) \neq \mathbf{0}$
- Je weiter $\mathbf{s}(\tilde{\beta})$ von $\mathbf{0} = \mathbf{s}(\hat{\beta})$ entfernt ist, desto stärker sprechen die Daten gegen H_0.

Score-Test

- Man kann zeigen, dass $\mathbf{s}(\beta) \sim N(\mathbf{0}, \mathbf{F}(\beta))$
- Damit folgt

$$\mathbf{s}(\tilde{\beta}) \overset{a}{\sim}_{H_0} N(\mathbf{0}, \mathbf{F}(\tilde{\beta}))$$

- Somit ist

$$U = \mathbf{s}(\tilde{\beta})^T \mathbf{F}^{-1}(\tilde{\beta})\mathbf{s}(\tilde{\beta}) \overset{a}{\sim}_{H_0} \chi^2_r$$

- Verwerfen wir also H_0, falls $U \geq Q_r^{\chi^2}(1-\alpha)$ oder mit dem p-Wert $p_{\mathbf{s}} = 1 - \chi^2_r(U) \leq \alpha$, dann erhalten wir einen asymptotischen Niveau α-Test

Beispiele

Beispiel globale Nullhypothese - LQ-Test

- $l(\beta') = \sum_{i=1}^{n} Y_i \eta_i + \log(1 - \pi_i)$
- $l(\hat{\beta})$ durch Einsetzten von $\hat{\eta}_i = \mathbf{x}_i \hat{\beta}$ und $\hat{\pi}_i = e^{\hat{\eta}_i}/(1 + e^{\hat{\eta}_i})$ für η_i und π_i
- $l(\tilde{\beta})$ durch Einsetzen von $\hat{\pi}_0 = \frac{1}{n}\sum_{j=1}^{n} Y_j$ und $\hat{\eta}_0 = \log\left(\hat{\pi}_0/(1 - \hat{\pi}_0)\right)$ für alle η_i und π_i
- Die Likelihood-Quotienten-Statistik ist dann $lq = 2\{l(\hat{\beta}) - l(\tilde{\beta})\}$

Beispiel globale Nullhypothese - Wald-Test

- Bei der globalen Nullhypothese ist

$$\mathbf{C} = (\mathbf{0}, \mathbf{I}_{k-1}), \quad \mathbf{d} = \mathbf{0} \quad \Rightarrow \quad \mathbf{C}\hat{\beta} = \begin{pmatrix} \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \hat{\beta}_{-1}$$

- $\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T$ entsteht aus $\mathbf{F}^{-1}(\hat{\beta})$ durch Löschung der ersten Spalte und ersten Zeile
- Damit lässt sich die Wald-Statistik berechnen

$$W = \hat{\beta}_{-1}^T \, [\mathbf{C}\,\mathbf{F}^{-1}(\hat{\beta})\,\mathbf{C}^T]^{-1} \hat{\beta}_{-1}$$

Beispiel globale Nullhypothese - Score-Test

- Wir haben gesehen, dass $\mathbf{s}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i^T (Y_i - \pi_i)$
- Zudem gilt unter der globalen Nullhypothese $\hat{\pi}_1 = \ldots = \hat{\pi}_n = \hat{\pi}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i$
- Also $\mathbf{s}(\tilde{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i^T (Y_i - \hat{\pi}_0)$
- Ähnlich sieht man, dass

$$\mathbf{F}(\tilde{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \hat{\pi}_0 (1 - \hat{\pi}_0) = (\mathbf{X}^T \mathbf{X}) \hat{\pi}_0 (1 - \hat{\pi}_0)$$

- Also

$$\mathbf{F}^{-1}(\tilde{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \frac{1}{\hat{\pi}_0 (1 - \hat{\pi}_0)}$$

Beispiel globale Nullhypothese - Score-Test

- Damit erhalten wir für die Score-Statistik ingesamt:

$$
\begin{aligned}
U &= \mathbf{s}(\tilde{\beta})^T \mathbf{F}^{-1}(\tilde{\beta})\mathbf{s}(\tilde{\beta}) \\
&= \frac{1}{\hat{\pi}_0(1-\hat{\pi}_0)} \left\{ \sum_{i=1}^{n} (Y_i - \hat{\pi}_0)\mathbf{x}_i \right\} (\mathbf{X}^T\mathbf{X})^{-1} \left\{ \sum_{j=1}^{n} \mathbf{x}_j^T (Y_j - \hat{\pi}_0) \right\} \\
&= \frac{1}{\hat{\pi}_0(1-\hat{\pi}_0)} \sum_i \sum_j (Y_i - \hat{\pi}_0)\mathbf{x}_i (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_j^T (Y_j - \hat{\pi}_0) \\
&= \frac{1}{\hat{\pi}_0(1-\hat{\pi}_0)} (\mathbf{Y} - \hat{\pi}_0 \mathbb{1})^T \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\pi}_0 \mathbb{1})
\end{aligned}
$$

Beispiel $H_0 : \beta_j = 0 \; \forall j \in S$ - LQ-Test

- Erhalten $\tilde{\eta}_i = \mathbf{x}_i \tilde{\beta}$ und $\tilde{\pi}_i = e^{\tilde{\eta}_i}/(1 + e^{\tilde{\eta}_i})$ aus logistischer Regression ohne die $\mathbf{x}^j$ für $j \in S$, aber mit den anderen $\mathbf{x}^l$ $(l \notin S)$
- Bilden $l(\tilde{\beta}) = \sum_{i=1}^{n} \{Y_i \tilde{\eta}_i + \log(1 - \tilde{\pi}_i)\}$ und $l(\hat{\beta})$ wie vorher
- Schließlich ist $lq = 2\{l(\hat{\beta}) - l(\tilde{\beta})\}$

Beispiel $H_0 : \beta_j = 0 \ \forall j \in S$ - Wald-Test

- Erhalten $\hat{\beta}_S = \mathbf{C}\hat{\beta}$ aus $\hat{\beta}$ durch Streichung von $\hat{\beta}_j$ für $j \notin S$ und $\mathbf{F}_S^{-1}(\hat{\beta}) = \mathbf{C}\,\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T$ aus $\mathbf{F}^{-1}(\hat{\beta})$ durch Streichung der Zeilen und Spalten j für $j \notin S$
- Also ist $W = \hat{\beta}_S^T\,[\mathbf{F}_S^{-1}(\hat{\beta})]^{-1}\,\hat{\beta}_S$
- Falls $S = \{j\}$ (d.h. $H_0 : \beta_j = 0$), dann ist $W = \hat{\beta}_j^2 / a_{jj}$, wobei a_{jj} das j-te Diagonalelement von $\mathbf{F}^{-1}(\hat{\beta})$ ist

Beispiel $H_0 : \beta_j = 0 \ \forall j \in S$ - Score-Test

- Erhalten $\mathbf{F}(\tilde{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \tilde{\pi}_i (1 - \tilde{\pi}_i)$ mit $\tilde{\pi}_i$ aus dem Modell ohne $\mathbf{x}^j$, $j \in S$
- Bilden die Inverse $\mathbf{F}^{-1}(\tilde{\beta})$ und erhalten mit $\mathbf{s}(\tilde{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i^T (Y_i - \tilde{\pi}_i)$ die Score-Statistik

$$U = \mathbf{s}(\tilde{\beta})^T \, \mathbf{F}^{-1}(\tilde{\beta}) \, \mathbf{s}(\tilde{\beta}).$$

- Sei $\mathbf{s}_j(\tilde{\beta})$ die j-te Komponente von $\mathbf{s}(\tilde{\beta})$
- $\tilde{\beta}$ wird so bestimmt, dass $s_j(\tilde{\beta}) = 0$ für alle $j \notin S$

Beispiel $H_0 : \beta_j = 0 \ \forall j \in S$ - Score-Test

- Insgesamt folgt

$$U = \mathbf{s}_S(\tilde{\beta})^T \, \mathbf{F}_S^{-1}(\tilde{\beta}) \, \mathbf{s}_S(\tilde{\beta}),$$

wobei $\mathbf{s}_S(\tilde{\beta})$ und $\mathbf{F}_S^{-1}(\tilde{\beta})$ wieder aus $\mathbf{s}(\tilde{\beta})$ bzw. $\mathbf{F}^{-1}(\tilde{\beta})$ durch Streichung der Zeilen und Spalten $j \notin S$ entsteht.

Kriterien zur Modellanpassung und Modellwahl

Anpassungsgüte und saturiertes Modell

- Wollen Modellanpassung an die Daten beurteilen
- Sind Daten gruppiert, lässt sich jeder Gruppe l eine eigene Wahrscheinlichkeit π_l zuordnen, die durch den Gruppenmittelwert $\bar{Y}_l$ geschätzt werden kann
- Erhalten so das sog. *saturierte Modell*
- Das saturierte Modell ist maximal an die gruppierten Daten angepasst und somit ein Maßstab zur Beurteilung der Modellanpassung
- Können dann untersuchen, ob die Abweichung in der Modellanpassung zwischen geschätztem und saturierten Modell signifikant ist
- Am häufigsten verwendet: Pearson-Statistik und Devianz

Pearson-Statistik und Devianz

- Mit der **Person-Statistik** vergleicht man die relativen Häufigkeiten $\bar{Y}_l$ in den Gruppen mit den durch das Modell geschätzten Wahrscheinlichkeiten $\hat{\pi}_l$ durch

$$\chi^2 = \sum_{l=1}^{m} \frac{(\bar{Y}_l - \hat{\pi}_l)^2}{\hat{\pi}_l(1 - \hat{\pi}_l)/n_l}$$

- Bei der **Devianz** vergleicht man die Log-Likelihood des in Frage stehenden Modells mit der des saturierten Modells

$$D = -2 \sum_{l=1}^{m} \{l_l(\hat{\pi}_l) - l_l(\bar{Y}_l)\},$$

wobei l_l die Log-Likelihood in der Gruppe l ist

Pearson-Statistik und Devianz

- Falls die Fallzahl n_l in allen Gruppen "groß" ist, dann sind χ^2 und D beide approximativ χ^2_{m-k}−verteilt, wobei k die Anzahl der geschätzten Koeffizienten ist
- Der Modellanpassungstest vergleicht dann den Wert von χ^2 bzw. D mit dem entsprechenden Quantil der χ^2_{m-k}−Verteilung
- Ist n_l zu klein, ist die Anwendung der Teststatistik problematisch. Große Werte von χ^2 und D weisen dann nicht notwendigerweise auf eine schlechte Anpassung hin

Modellwahl bei unterschiedlicher Kovariablenzahl

- Bei Modellen mit unterschiedlicher Zahl an Prädiktoren und Parametern muss ein Kompromiss zwischen der Datenanpassung und der Modellkomplexität gefunden werden

- Zum Beispiel mit dem *Akaikeschen Informationskriterium*

$$AIC = \underbrace{-2\, l(\hat{\beta})}_{Anpassung} + \underbrace{2k}_{Modellkomplexit\ddot{a}t}$$

- Bevorzugt werden dann Modelle mit kleinem AIC