Inhalt

3.1 Parametrische Likelihood-Inferenz

Situation: $\mathcal{P} = \{f(\mathbf{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$, $k \ll n$, k konstant für $n \to \infty$. $f(\mathbf{x}|\boldsymbol{\theta})$ ist eine diskrete oder stetige oder allgemeiner eine Radon-Nikodym-Dichte.

Definition 3.1 (Likelihoodfunktion)

Die Likelihoodfunktion von $\boldsymbol{\theta} \in \Theta$,

$$L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}),$$

ist definiert als die Dichte der beobachteten Daten $\mathbf{X} = (X_1, \ldots, X_n) = \mathbf{x} = (x_1, \ldots, x_n)$, betrachtet als Funktion von $\boldsymbol{\theta}$. Mit $L(\boldsymbol{\theta}|\mathbf{x})$ ist auch $\widetilde{L}(\boldsymbol{\theta}|\mathbf{x}) = const \times L(\boldsymbol{\theta}|\mathbf{x})$ eine Likelihoodfunktion.

3.1 Parametrische Likelihood-Inferenz

Eine alternative Definition definiert die Likelihood als die Wahrscheinlichkeit für die beobachteten Daten, aufgefasst als Funktion von $\boldsymbol{\theta}$, das das statistische Modell parametrisiert.

Für diskrete Beobachtungen ist die Übereinstimmung der Definitionen klar.

Für stetige Beobachtungen x, betrachte für kleines ε

$$\mathbb{P}_\theta(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2}) = \int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f(x|\boldsymbol{\theta})dx \approx \varepsilon f(x|\boldsymbol{\theta}).$$

Da ε nicht von $\boldsymbol{\theta}$ abhängt, kann der konstante Faktor in der Likelihood $L(\boldsymbol{\theta}|x)$ ignoriert werden.

3.1 Parametrische Likelihood-Inferenz

Zu unterscheiden sind folgende häufige Situationen:

1. $X_1, \ldots, X_n$ sind i.i.d. wie $X_1 \sim f_1(x|\boldsymbol{\theta})$. Es gilt die Faktorisierung
$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f_1(x_i|\boldsymbol{\theta}).$$

2. $X_1, \ldots, X_n$ – bzw. $y_1|\mathbf{z}_1, \ldots, y_n|\mathbf{z}_n$ im Regressionsfall bei einer Zielvariable y und Kovariablenvektor $\mathbf{z}$ – sind unabhängig, aber nicht mehr identisch verteilt. Es gilt die Faktorisierung
$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i|\boldsymbol{\theta}).$$

3. $\mathbf{X}_1, \ldots, \mathbf{X}_n$ sind unabhängig, die einzelnen Komponenten innerhalb eines Vektors sind jedoch möglicherweise abhängig.

3.1 Parametrische Likelihood-Inferenz

4. Zeitlich korrelierte Daten / Stichprobenvariablen $X_1, \ldots, X_t, \ldots, X_n$ mit Dichtefunktion

$$f(x_1, \ldots, x_t, \ldots, x_n|\boldsymbol{\theta}) =$$
$$f(x_n|x_{n-1}, \ldots, x_t, \ldots, x_1; \boldsymbol{\theta}) \cdot \ldots \cdot f(x_{n-1}|x_{n-2}, \ldots, x_1; \boldsymbol{\theta}) \cdot \ldots$$
$$\cdot f(x_2|x_1; \boldsymbol{\theta}) f(x_1|\boldsymbol{\theta}).$$

Bei Markov-Ketten erster Ordnung mit der Eigenschaft

$$f(x_n|x_{n-1}, \ldots, x_1; \boldsymbol{\theta}) = f(x_n|x_{n-1}; \boldsymbol{\theta})$$

vereinfacht sich die Likelihood zu

$$L(\boldsymbol{\theta}|\mathbf{x}) = \left(\prod_{i=2}^{n} f(x_i|x_{i-1}; \boldsymbol{\theta}) \right) f(x_1|\boldsymbol{\theta}).$$

3.1 Parametrische Likelihood-Inferenz

Beispiel 3.1 (zu diesen vier Situationen)

1. Siehe z.B. Beispiele in 1.1.1
2. Regressionssituationen (Querschnittsdaten) mit unabhängigen Zielvariablen $y_1|\mathbf{z}_1, \ldots, y_n|\mathbf{z}_n$ und Kovariablen $\mathbf{z}_i$: z.B.
   - klassisches lineares Modell: $y_i|\mathbf{z}_i \sim N(\mathbf{z}_i^\top \boldsymbol{\beta}, \sigma^2)$,
   - Logit- oder Probitmodell: $y_i|\mathbf{z}_i \sim \text{Bin}(1, \pi_i = h(\mathbf{z}_i^\top \boldsymbol{\beta}))$,
   - Poisson-Regression: $y_i|\mathbf{z}_i \sim \text{Po}(\lambda_i = h(\mathbf{z}_i^\top \boldsymbol{\beta}))$.
3.
   - multivariate Daten
   - Survivaldaten $\mathbf{X}_i = (T_i, \delta_i)$ mit stetiger Lebensdauer T_i und diskretem Zensierungsindikator δ_i
4. Markov-Ketten, autoregressive Modelle für Zeitreihen/Longitudinaldaten.

3.1 Parametrische Likelihood-Inferenz

Autoregressiver Prozess 1. Ordnung: Sei

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t$$

mit $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ oder – mit zusätzlichem (zeitabhängigen) Kovariablenvektor $\mathbf{z}_t$ –

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \mathbf{z}_t^\top \boldsymbol{\beta} + \varepsilon_t.$$

In letzterem Fall hat die Likelihood die Form

$$L(\boldsymbol{\theta}) = \left(\prod_{t=2}^{n} f_t(y_t | y_{t-1}; \boldsymbol{\theta}) \right) f_1(y_1)$$

mit

$$f_t(y_t | y_{t-1}; \boldsymbol{\theta}) = \phi(y_t | \alpha_0 + \alpha_1 y_{t-1} + \mathbf{z}_t^\top \boldsymbol{\beta}, \sigma^2),$$

wobei $\phi(y|\mu, \tau^2)$ den Wert der Normalverteilungsdichte mit Erwartungswert μ und Varianz τ^2 an der Stelle y bezeichnet.

Beispiel 3.2.
Wir betrachten unabhängige, aber (teils) unvollständige Ziehungen aus $N(\theta, 1)$.

1. Ziehung: Es sei $x_1 = 2.45$. Dann ist

$$\begin{aligned} L_1(\theta) &= L(\theta | X_1 = 2.45) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(2.45 - \theta)^2\right). \end{aligned}$$

2. Ziehung: Es sei nur $0.9 < x_2 < 4$ bekannt (unvollständige oder intervallzensierte Beobachtung). Die Likelihood lautet dann, wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet:

$$\begin{aligned} L_2(\theta) \;=\; L(\theta|0.9 < X_2 < 4) \;&=\; \mathbb{P}_\theta(0.9 < X_2 < 4) \\ &=\; \Phi(4-\theta) - \Phi(0.9-\theta). \end{aligned}$$

Formal könnte man auch eine binäre Variable

$$X_2^d = \begin{cases} 1, & 0.9 < X_2 < 4, \\ 0, & \text{sonst} \end{cases}$$

mit Dichtefunktion

$$f_2^d(1) = \mathbb{P}(X_2^d = 1) = \Phi(4-\theta) - \Phi(0.9-\theta)$$

definieren.

3. Ziehung: $z_1, \ldots, z_n$ seien i.i.d. Realisierungen aus $N(\theta, 1)$.
Bekannt sei aber nur

$$x_3 = \max_{1 \leq i \leq n} z_i = z_{(n)}.$$

Der Rest sind fehlende Werte („missing values").
Die Verteilungsfunktion von $X_3 = Z_{(n)}$ ist

$$\begin{aligned} F_\theta(z_{(n)}) &= \mathbb{P}_\theta(Z_{(n)} \leq z_{(n)}) = \mathbb{P}_\theta(Z_i \leq z_{(n)} \ \forall \ i) \\ &= [\Phi(z_{(n)} - \theta)]^n. \end{aligned}$$

Die Dichte ergibt sich über Differentiation bezüglich θ:

$$f_\theta(z_{(n)}) = n[\Phi(z_{(n)} - \theta)]^{n-1} \phi(z_{(n)} - \theta),$$

d.h. für zum Beispiel $n = 5$ und $z_{(n)} = x_3 = 3.5$ gilt

$$L_3(\theta) = L(\theta | X_3 = 3.5) = 5[\Phi(3.5 - \theta)]^4 \phi(3.5 - \theta).$$

Die gesamte Likelihood der drei Beobachtungen ist

$$L(\theta|x_1, 0.9 < X_2 < 4, x_3) = L_1(\theta) \cdot L_2(\theta) \cdot L_3(\theta),$$

also das Produkt der Likelihoodfunktionen L_1, L_2 und L_3.

Fazit: Die Likelihood ist sehr allgemein definiert.

Beziehung zur Bayes-Inferenz

- $p(\boldsymbol{\theta})$ sei die Prioriverteilung,
- $L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$ die Likelihood.
- Dann ist

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}|\mathbf{x}) \\ \text{„Posteriori"} &\propto \text{„Priori"} \times \text{Likelihood.} \end{aligned}$$

3.1 Parametrische Likelihood-Inferenz

Likelihood-Quotient

Frage: Wie vergleicht man die Likelihoods $L(\boldsymbol{\theta}_1|\mathbf{x})$ und $L(\boldsymbol{\theta}_2|\mathbf{x})$ für $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$?

Antwort: Man betrachtet den Quotienten (nicht die Differenz), da dieser invariant gegenüber eineindeutigen Transformationen

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{y}(\mathbf{x}) \;\Leftrightarrow\; \mathbf{y} \mapsto \mathbf{x}(\mathbf{y})$$

ist. Für stetige $\mathbf{x}, \mathbf{y}$ gilt mit dem Transformationssatz für Dichten:

$$f_Y(\mathbf{y}|\boldsymbol{\theta}) = f_X(\mathbf{x}(\mathbf{y})|\boldsymbol{\theta}) \left| \det \left(\frac{\partial \mathbf{x}(\mathbf{y})}{\partial \mathbf{y}} \right) \right|$$

und somit

$$L(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}|\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{x}(\mathbf{y})}{\partial \mathbf{y}} \right) \right| \;\Rightarrow\; \frac{L(\boldsymbol{\theta}_2|\mathbf{y})}{L(\boldsymbol{\theta}_1|\mathbf{y})} = \frac{L(\boldsymbol{\theta}_2|\mathbf{x})}{L(\boldsymbol{\theta}_1|\mathbf{x})}.$$

Satz 3.2

1. *Sei* $\mathbf{T} = \mathbf{T}(\mathbf{X})$ *suffizient für* $\boldsymbol{\theta}$. *Dann gilt* $L(\boldsymbol{\theta}|\mathbf{x}) = const \times L(\boldsymbol{\theta}|\mathbf{t})$ *mit* $\mathbf{t} = \mathbf{T}(\mathbf{x})$, *d.h.* $L(\boldsymbol{\theta}|\mathbf{x})$ *und* $L(\boldsymbol{\theta}|\mathbf{t})$ *sind äquivalent.*
2. $L(\boldsymbol{\theta}|\mathbf{x})$ *ist minimalsuffizient.*

Beweis. Folgt unmittelbar aus den Resultaten aus Abschnitt 2. □

3.2 Maximum-Likelihood-Schätzung

Die Maximum-Likelihood-Schätzung ist die populärste Methode zur Konstruktion von Punktschätzern bei rein parametrischen Problemstellungen.

Maximum-Likelihood-Prinzip: Finde Maximum-Likelihood-Schätzwert $\widehat{\boldsymbol{\theta}}$, so dass

$$L(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \geq L(\boldsymbol{\theta}|\mathbf{x}) \text{ für alle } \boldsymbol{\theta} \in \Theta.$$

Dazu äquivalent ist

$$\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) \geq \ell(\boldsymbol{\theta}|\mathbf{x}), \quad \ell(\boldsymbol{\theta}|\mathbf{x}) = \log L(\boldsymbol{\theta}|\mathbf{x})$$

mit der Log-Likelihood ℓ.

Invarianz des ML-Schätzers
Wenn $\widehat{\boldsymbol{\theta}}$ der ML-Schätzer von $\boldsymbol{\theta}$ ist und $\mathbf{h}(\cdot)$ eine eineindeutige Funktion, so ist $\mathbf{h}(\widehat{\boldsymbol{\theta}})$ der ML-Schätzer von $\mathbf{h}(\boldsymbol{\theta})$.

3.2 Maximum-Likelihood-Schätzung

3.2.1 Schätzkonzept

Meist sucht man nach (lokalen) Maxima von $\ell(\boldsymbol{\theta}|\mathbf{x})$ durch Nullsetzen der Score-Funktion

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_k} \right)^\top$$

(soweit die 1. Ableitung der Log-Likelihood existiert!) als Lösung der sogenannten *ML-Gleichung*

$$\mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = \mathbf{0}.$$

Dies funktioniert (meist) unter Annahme von Fisher-Regularität. Nur in einfachen Fällen ist die Lösung analytisch zugänglich.

3.2 Maximum-Likelihood-Schätzung

3.2.1 Schätzkonzept

Beispiel 3.3 (Lineares Modell)
Betrachte

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{mit} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

- Likelihood:

$$L(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2\right)$$

- Log-Likelihood:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \underbrace{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2}_{\text{KQ-Kriterium}}$$

3.2 Maximum-Likelihood-Schätzung

3.2.1 Schätzkonzept

- Score-Funktion:

$$\begin{aligned} \mathbf{s}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2) &= \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{Z}^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \\ s_{\sigma^2}(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2 \end{aligned}$$

Man verifiziert leicht, dass $\mathbb{E}[\mathbf{s}_{\boldsymbol{\beta}}] = \mathbf{0}, \mathbb{E}[s_{\sigma^2}] = 0$ ist. Aus den ML-Gleichungen, die sich durch Nullsetzen der Score-Funktionen ergeben, folgt:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\mathsf{ML}} &= (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}, \\ \widehat{\sigma}^2_{\mathsf{ML}} &= \frac{1}{n} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2. \end{aligned}$$

- Score-Funktion (fortgeführt):
  Der ML-Schätzer für $\boldsymbol{\beta}$ entspricht also dem KQ-Schätzer. Der ML-Schätzer für σ^2 ist verzerrt, aber asymptotisch erwartungstreu.

  Der Restricted Maximum Likelihood (REML) Schätzer

$$\sigma^2_{\text{REML}} = \frac{1}{n-p}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2$$

  ist erwartungstreu für σ^2. Dabei ist p die Dimension von $\boldsymbol{\beta}$.

3.2 Maximum-Likelihood-Schätzung

3.2.1 Schätzkonzept

- Informationsmatrizen:

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{\partial s_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}^\top} = \frac{1}{\sigma^2} \mathbf{Z}^\top \mathbf{Z} \quad = \left(\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \right)^{-1} \text{ (von } \mathbf{y} \text{ unabh.)}$$

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \sigma^2} = \frac{1}{\sigma^4} \mathbf{Z}^\top (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \qquad \Rightarrow \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta} \partial \sigma^2} \right] = \mathbf{0}$$

$$-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2 \partial \sigma^2} = -\frac{n}{2(\sigma^2)^2} + \frac{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2}{(\sigma^2)^3} \qquad \Rightarrow \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2 \partial \sigma^2} \right] = \frac{n}{2\sigma^4}$$

Der letzte Erwartungswert folgt aus

$$\|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2 = \sum_{i=1}^{n} \varepsilon_i^2 \ \sim \ \sigma^2 \chi^2(n).$$

3.2 Maximum-Likelihood-Schätzung

3.2.1 Schätzkonzept

Beispiel 3.4 (GLMM für Longitudinaldaten)
Sei $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})$ mit bedingt unabhängigen Komponenten $y_{ij} \sim f(y_{ij}|\mu_{ij})$ und $\mu_{ij} = h(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i)$.
Die $\mathbf{b}_i$ sind z.B. individuenspezifische Intercepts b_i ($\mathbf{z}_{ij} \equiv 1$) mit Prioriverteilung $b_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$. Die Likelihood des Parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2)$ lautet

$$L(\boldsymbol{\beta}, \tau^2) = \int \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \mathbf{b}_i) p(\mathbf{b}_i|\tau^2)\, d\mathbf{b}_i.$$

Lösungsansätze für die Maximierung der Likelihood:
EM-Algorithmus oder numerische Integration bzw. Bayes-Inferenz.
Siehe die Vorlesungen Gemischte Modelle und Analyse Longitudinaler Daten.

3.2 Maximum-Likelihood-Schätzung

3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Wenn keine analytische Lösung existiert, geschieht die numerische Lösung der ML-Gleichung über Verfahren wie Newton-Raphson, Quasi-Newton, Fisher-Scoring oder den EM-Algorithmus.

Newton-Raphson
ist ein allgemeines Verfahren zur Nullstellensuche einer stetig differenzierbaren Funktion $\mathbf{g}(\boldsymbol{\theta})$. Im skalaren Fall liefert eine Taylorentwicklung um den aktuellen Iterationswert $\theta^{(t)}$

$$g(\theta) \approx g(\theta^{(t)}) + g'(\theta^{(t)})(\theta - \theta^{(t)}) \stackrel{!}{=} 0$$

und damit als approximative Lösung θ den nächsten Iterationswert

$$\theta^{(t+1)} = \theta^{(t)} - \frac{g(\theta^{(t)})}{g'(\theta^{(t)})}.$$

Das mehrdimensionale Analogon für die Lösung der ML-Gleichung $\mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = \mathbf{0}$ angewandt, ergibt

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + [\mathbf{J}(\boldsymbol{\theta}^{(t)}|\mathbf{x})]^{-1}\mathbf{s}(\boldsymbol{\theta}^{(t)}|\mathbf{x}),$$

da die beobachtete Informationsmatrix

$$\mathbf{J}(\boldsymbol{\theta}|\mathbf{x}) = -\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = -\frac{\partial \mathbf{s}(\boldsymbol{\theta}|\mathbf{x})}{\partial\boldsymbol{\theta}^\top}$$

der negativen Ableitung von $\mathbf{s}(\boldsymbol{\theta}|\mathbf{x})$ entspricht.

Quasi-Newton arbeitet mit Approximationen an $\mathbf{J}(\boldsymbol{\theta}|\mathbf{x})$,
Fisher-Scoring mit der erwarteten Informationsmatrix
$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{E}_\theta[\mathbf{J}(\boldsymbol{\theta}|\mathbf{X})]$.

EM (Expectation-Maximization)-Algorithmus
Der EM-Algorithmus ist eine Alternative zu Newton-Raphson, Fisher-Scoring usw., vor allem in Modellen mit unvollständigen Daten oder latenten (nicht direkt beobachtbaren) Variablen oder Faktoren (vgl. Computerintensive Methoden).

Beispiele:

- Wie in 1.1.6 a) werden Lebensdauern beobachtet. Für die rechtszensierten Beobachtungen ist die genaue Lebensdauer nicht bekannt.
- Die Daten stammen aus einer Mischverteilung, aus welcher Mischungskomponente jede Beobachtung stammt, ist jedoch unbekannt.
- Gemischte Modelle

Notation:

- $\mathbf{x}$ beobachtbare („unvollständige") Daten
- $\mathbf{z}$ unbeobachtbare Daten/latente Variablen
- $(\mathbf{x}, \mathbf{z})$ vollständige Daten
- $L(\boldsymbol{\theta}|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta})$ Likelihood der beobachtbaren Daten
- $L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) = f(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ Likelihood der vollständigen Daten

Der EM-Algorithmus ist insbesondere nützlich, wenn $L(\boldsymbol{\theta}|\mathbf{x})$ schwierig zu berechnen und $L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$ leichter zu handhaben ist.

3.2 Maximum-Likelihood-Schätzung

3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Intuition:

$$\begin{aligned} \log f(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) &= \log f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}) + \log f(\mathbf{x}|\boldsymbol{\theta}) \\ \Rightarrow \underbrace{\mathbb{E}_{\theta^{(k)}}[\ell(\boldsymbol{\theta}|\mathbf{x}, \mathbf{Z})]}_{Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})} &= \underbrace{\mathbb{E}_{\theta^{(k)}}[\ell(\boldsymbol{\theta}|(\mathbf{Z}|\mathbf{x}))]}_{C(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})} + \ell(\boldsymbol{\theta}|\mathbf{x}), \end{aligned}$$

wobei der Erwartungswert $\mathbb{E}_{\theta^{(k)}}$ bzgl. $f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}^{(k)})$ genommen wird.

Es gilt $C(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) \geq C(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)})$ wegen der Informationsungleichung.

Daher folgt aus $Q(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})$ auch

$$\ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{x}) \geq \ell(\boldsymbol{\theta}^{(k)}|\mathbf{x}).$$

Vorgehen: Für gegebenes $\boldsymbol{\theta}^{(k)}$ maximiere $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ über $\boldsymbol{\theta}$ und erhalte $\boldsymbol{\theta}^{(k+1)}$. Neues $\boldsymbol{\theta}^{(k+1)}$ erhöht die Likelihood. Iteriere.

Algorithmus 1 : EM-Algorithmus

Startwert: $\boldsymbol{\theta}^{(0)}$

- **E**-Schritt: Berechne

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}) = \underbrace{\mathbb{E}_{\theta^{(0)}}[\ \ell(\boldsymbol{\theta}|\mathbf{x}, \mathbf{Z})\ |\mathbf{x}]}_{\text{bzgl. der Vtlg. von } \mathbf{Z}|\mathbf{x} \text{ mit Parameter } \boldsymbol{\theta}^{(0)}}.$$

- **M**-Schritt: Berechne $\boldsymbol{\theta}^{(1)}$, so dass $Q(\boldsymbol{\theta})$ maximiert wird:

$$\boldsymbol{\theta}^{(1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}).$$

Iteriere **E**/**M**-Schritte: $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$ bis zur Konvergenz.

Satz 3.3

Unter relativ allgemeinen Annahmen gilt $\boldsymbol{\theta}^{(m)} \rightarrow \widehat{\boldsymbol{\theta}}_{ML}$ *für* $m \rightarrow \infty$.

Eigenschaften des EM-Algorithmus:

- Monotonie: $\ell(\boldsymbol{\theta}^{(m+1)}|\mathbf{x}) \geq \ell(\boldsymbol{\theta}^{(m)}|\mathbf{x})$.
- Langsame Konvergenz.
- Der Standardfehler des resultierenden Schätzers ist schwierig zu bestimmen, die Informationsmatrix ist nicht direkt zugänglich wie beim Fisher-Scoring.

Eine Alternative bietet u.a. die Bayes-Inferenz.

Beispiel 3.5 (Mischverteilungen)
Seien $X_1, \ldots, X_n$ i.i.d. wie $X \sim f(x|\boldsymbol{\theta})$, dabei darf X auch multivariat sein. Betrachte die Mischverteilung

$$f(x|\boldsymbol{\theta}) = \sum_{j=1}^{J} \pi_j f_j(x|\boldsymbol{\vartheta}_j) \quad \text{mit} \quad \boldsymbol{\theta} = (\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_J,\ \pi_1, ..., \pi_J). \quad (3.2)$$

Dabei sind

- π_j unbekannte Mischungsanteile, $\sum_{j=1}^{J} \pi_j = 1$,
- $f_j(x|\boldsymbol{\vartheta}_j)$ die j-te Mischungskomponente,
- $\boldsymbol{\vartheta}_j$ der unbekannte Parameter(-vektor) für Komponente j.

Speziell: Bei einer Mischung von Normalverteilungen erhalten wir

$$\begin{aligned} f_j(x|\boldsymbol{\vartheta}_j) &\propto |\boldsymbol{\Sigma}_j|^{-1/2} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(x-\boldsymbol{\mu}_j)\right) \\ X &\sim \pi_1 N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \ldots + \pi_J N(\boldsymbol{\mu}_J, \boldsymbol{\Sigma}_J). \end{aligned}$$

Im univariaten Fall mit zwei Mischungskomponenten also:

$$X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2).$$

Interpretation des Mischungsmodells (3.2): x_i entstammt einer von J Subpopulationen, wobei in Subpopulation j gilt:

$$X_i|j \sim f_j(x_i|\boldsymbol{\vartheta}_j).$$

Definiere die unbeobachtete (latente) Indikatorvariable Z_i für $j = 1, \ldots, J$ durch

$$Z_i = j \;\Leftrightarrow x_i \text{ ist aus Population } j.$$

Die Randverteilung sei $\mathbb{P}(Z_i = j) = \pi_j,\ j = 1, \ldots, J$. Dann lautet die bedingte Verteilung von $X_i|Z_i$:

$$X_i|Z_i = j \;\sim\; f_j(x_i|\boldsymbol{\vartheta}_j).$$

3.2 Maximum-Likelihood-Schätzung

3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

Die Log-Likelihood der beobachteten Daten $\mathbf{x}$ ist

$$\ell(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{J} \pi_j f_j(x_i|\boldsymbol{\vartheta}_j) \right),$$

die der vollständigen Daten $(\mathbf{x}, \mathbf{z})$

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{n} \log f(x_i, z_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log \left(f(x_i|z_i, \boldsymbol{\theta}) \cdot f(z_i|\boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{n} (\log f_{z_i}(x_i|\boldsymbol{\vartheta}_{z_i}) + \log \pi_{z_i}). \end{aligned}$$

E-Schritt:

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\theta^{(m)}}[\ \ell(\boldsymbol{\theta}|\mathbf{x}, \mathbf{Z})\ |\mathbf{x}]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}^{(m)}\{\log \pi_j - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(x_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(x_i - \boldsymbol{\mu}_j)\}$$

wobei wir nur

$$p_{ij}^{(m)} = \mathbb{P}(Z_i = j|x_i, \boldsymbol{\theta}^{(m)}) \overset{\text{Bayes}}{=} \frac{\pi_j^{(m)} f_j(x_i|\boldsymbol{\vartheta}_j^{(m)})}{\sum_{s=1}^{J}\pi_s^{(m)} f_s(x_i|\boldsymbol{\vartheta}_s^{(m)})}.$$

für $i = 1, \ldots, n$, $j = 1, \ldots, J$ tatsächlich in der Praxis berechnen müssen.

3.2 Maximum-Likelihood-Schätzung

3.2.2 Iterative numerische Verfahren zur Berechnung des ML-Schätzers

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j}^{J} p_{ij}^{(m)} \{\log \pi_j - \frac{1}{2} \log |\boldsymbol{\Sigma}_j| - \frac{1}{2} (x_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (x_i - \boldsymbol{\mu}_j)\}$$

M-Schritt: Berechne

$$\pi_j^{(m+1)} = \underset{\pi_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{1.}{=} \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(m)}$$

$$\boldsymbol{\mu}_j^{(m+1)} = \underset{\mu_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{2.}{=} \sum_{i=1}^{n} w_{ij}^{(m)} x_i$$

$$\boldsymbol{\Sigma}_j^{(m+1)} = \underset{\Sigma_j}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}) \overset{3.}{=} \sum_{i=1}^{n} w_{ij}^{(m)} (x_i - \boldsymbol{\mu}_j^{(m+1)})(x_i - \boldsymbol{\mu}_j^{(m+1)})^T$$

mit $w_{ij}^{(m)} = \frac{p_{ij}^{(m)}}{\sum_{i=1}^{n} p_{ij}^{(m)}}$. 1. folgt für $J = 2$ als Maximierer der binomialen Likelihood (für $J > 2$ braucht man Lagrange). 2.+3. folgt als Maximierer der gewichteten Normalverteilungslikelihood.

Beispiel 3.6 (Gemischte Modelle)
Eine Herleitung für E- und M-Schritt in linearen gemischten Modellen findet sich in Pawitan, Kapitel 12.8.

Satz 3.4

Seien $X_1, \ldots, X_n$ i.i.d. aus einer Dichte $f(x|\boldsymbol{\theta})$, die folgenden Annahmen genügt:

- *$f(x|\boldsymbol{\theta})$ ist Fisher-regulär.*
- *Die Informationsmatrix $\mathcal{I}(\boldsymbol{\theta})$ ist positiv definit im Inneren von Θ.*
- *Es existieren Funktionen M_{jkl} derart, dass*

$$\left| \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(x|\boldsymbol{\theta}) \right| \leq M_{jkl}(x)$$

und

$$\mathbb{E}_{\theta_0}[M_{jkl}(X)] < \infty$$

für alle j, k und l, wobei $\boldsymbol{\theta}_0$ den wahren Wert des Parameters bezeichnet.

Satz 3.4 (fortgeführt)

Dann gilt unter weiteren, relativ schwachen Regularitätsannahmen:

- *Die Likelihood-(ML-)Gleichungen haben für* $n \to \infty$ *mit Wahrscheinlichkeit 1 eine Lösung* $\widehat{\boldsymbol{\theta}}_n$ *(d.h.* $\mathbb{P}(\widehat{\boldsymbol{\theta}}_n \text{ existiert}) \to 1$*) mit* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$*; die konsistente Lösung* $\widehat{\boldsymbol{\theta}}_n$ *ist eindeutig und* $\mathbb{P}(\widehat{\boldsymbol{\theta}}_n \text{ ist (lokales) Maximum}) \to 1$.
- $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N(\boldsymbol{\theta}_0, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_0))$ *bzw.* $\boldsymbol{\mathcal{I}}^{1/2}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k)$,
- $\widehat{\boldsymbol{\theta}}_n \overset{a}{\sim} N(\boldsymbol{\theta}_0, \mathbf{J}^{-1}(\boldsymbol{\theta}_0|\mathbf{X}))$ *bzw.* $\mathbf{J}^{1/2}(\boldsymbol{\theta}_0|\mathbf{X})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k)$,

d.h. ML-Schätzer sind asymptotisch erwartungstreue BAN-Schätzer.

Bemerkung

1. Es sind auch andere Varianten von Regularitätsannahmen möglich.
2. Der Satz gilt unter stärkeren Regularitätsannahmen auch für i.n.i.d. und abhängige $X_1, \ldots, X_n$
3. $\mathcal{I}(\theta_0)$ und $\mathbf{J}(\theta_0|\mathbf{x})$ können auch durch $\mathcal{I}(\widehat{\theta}_n)$ bzw. $\mathbf{J}(\widehat{\theta}_n|\mathbf{x})$ ersetzt werden.

Beweis. Erfolgt lediglich skizzenhaft.

- *Existenz (für skalares θ):* Es gilt für alle $\theta \neq \theta_0$, dass

$$\mathbb{P}_{\theta_0}\left(\prod_{i=1}^{n} f(x_i|\theta_0) > \prod_{i=1}^{n} f(x_i|\theta)\right) \to 1 \text{ für } n \to \infty. \quad (2)$$

Beweis: Logarithmieren liefert

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(f(x_i|\theta)/f(x_i|\theta_0)\right) < 0.$$

Nach dem Gesetz der großen Zahlen konvergiert die linke Seite in Wahrscheinlichkeit gegen die Kullback-Leibler-Distanz

$$\mathbb{E}_{\theta_0}[\log\left(f(X|\theta)/f(X|\theta_0)\right)].$$

Anwendung der Ungleichung von Jensen liefert, dass

$$\mathbb{E}_{\theta_0}[\log\left(f(X|\theta)/f(X|\theta_0)\right)] < \log \mathbb{E}_{\theta_0}[f(X|\theta)/f(X|\theta_0)] = 0,$$

woraus die Behauptung (2) folgt.

Wähle nun $a > 0$ klein genug, so dass $(\theta_0 - a; \theta_0 + a)$ vollständig in Θ enthalten ist. Setze

$$S_n = \{\mathbf{x} : L(\theta_0|\mathbf{x}) > L(\theta_0 - a|\mathbf{x}) \text{ und } L(\theta_0|\mathbf{x}) > L(\theta_0 + a|\mathbf{x})\}.$$

Für beliebige Stichproben $\mathbf{x} \in S_n$ existiert somit ein Punkt $\widehat{\theta}_n \in (\theta_0 - a; \theta_0 + a)$, der die Likelihood (lokal) maximiert, d.h. $s(\widehat{\theta}_n|\mathbf{x}) = 0$. Aus eben bewiesener Hilfsaussage (2) folgt, dass $\mathbb{P}_{\theta_0}(S_n) \to 1$ für jedes beliebige a.

3.2 Maximum-Likelihood-Schätzung

3.2.3 Asymptotische Eigenschaften

- *Konsistenz und Eindeutigkeit (für skalares θ):* Huzurbazar (1948): The likelihood equation, consistency and the maxima of the likelihood function. *Ann. Eugenics* **14**, 185-200.

- *Asymptotische Normalität der Score-Funktion:* Aus der Fisher-Regularität folgt, dass der Erwartungswert und die Kovarianzmatrix existieren und durch $\mathbb{E}_{\theta_0}[\mathbf{s}_i(\boldsymbol{\theta}_0|X_i)] = \mathbf{0}$ und $\mathrm{Cov}_{\theta_0}(\mathbf{s}_i(\boldsymbol{\theta}_0|X_i)) = \mathbf{i}(\boldsymbol{\theta}_0)$ gegeben sind.
  Der zentrale Grenzwertsatz liefert $\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X}) \overset{a}{\sim} N(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0))$ bzw.

$$\mathcal{I}(\boldsymbol{\theta}_0)^{-1/2}\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X}) = (n\,\mathbf{i}(\boldsymbol{\theta}_0))^{-1/2}(\sum_{i=1}^{n}\mathbf{s}_i(\boldsymbol{\theta}_0|X_i) - \mathbf{0}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_k).$$

- *Asymptotische Normalität von $\widehat{\theta}_n$:* Eine Taylorentwicklung von $\mathbf{s}(\widehat{\theta}_n|\mathbf{x}) = \mathbf{0}$ um θ_0 führt zu

$$\mathbf{0} = \mathbf{s}(\widehat{\theta}_n|\mathbf{x}) = \mathbf{s}(\theta_0|\mathbf{x}) - \mathbf{J}(\theta_0|\mathbf{x})(\widehat{\theta}_n - \theta_0) + o(\widehat{\theta}_n - \theta_0).$$

Wegen des Satzes von Slutsky können wir im Folgenden auch $\mathbf{J}(\theta_0|\mathbf{x})$ durch $\mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}[\mathbf{J}(\theta_0|\mathbf{X})]$ ersetzen, da $\frac{1}{n}\mathbf{J}(\theta_0|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n \mathbf{j}_i(\theta_0|X_i) \xrightarrow{\mathbb{P}} \mathbf{i}(\theta_0)$. Dies liefert

$$\mathbf{s}(\theta_0|\mathbf{X}) \overset{a}{\sim} \mathcal{I}(\theta_0)(\widehat{\theta}_n - \theta_0) \quad \text{bzw.} \quad \widehat{\theta}_n - \theta_0 \overset{a}{\sim} \mathcal{I}^{-1}(\theta_0)\mathbf{s}(\theta_0|\mathbf{X})$$

und somit

$$\widehat{\theta}_n - \theta_0 \overset{a}{\sim} N(\mathbf{0}, \mathcal{I}^{-1}(\theta_0)\mathcal{I}(\theta_0)\mathcal{I}^{-1}(\theta_0)) = N(\mathbf{0}, \mathcal{I}^{-1}(\theta_0)).$$

□

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Betrachte lineare Hypothesen

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}\boldsymbol{\theta} \neq \mathbf{d},$$

wobei $\mathbf{C}$ vollen Zeilenrang $s \leq k = \dim(\boldsymbol{\theta})$ besitze.

Wichtiger Spezialfall:

$$H_0 : \boldsymbol{\theta}_s = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta}_s \neq \mathbf{0},$$

wobei $\boldsymbol{\theta}_s$ einen beliebigen s-dimensionalen Subvektor von $\boldsymbol{\theta}$ bezeichnet, zum Beispiel in einem GLM, wo $\boldsymbol{\beta}_s = \mathbf{0}$ bedeutet, dass die zugehörigen Kovariablen keinen Einfluss haben.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Likelihood-Quotienten-Statistik

Die Likelihood-Quotienten-Statistik

$$\lambda = 2\left(\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - \ell(\widetilde{\boldsymbol{\theta}}|\mathbf{x})\right) = 2\log\left[\frac{L(\widehat{\boldsymbol{\theta}}|\mathbf{x})}{L(\widetilde{\boldsymbol{\theta}}|\mathbf{x})}\right]$$

vergleicht das unrestringierte Maximum der Log-Likelihood $\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ (über Θ) mit dem Maximum der Log-Likelihood unter der H_0-Restriktion, d.h. $\widetilde{\boldsymbol{\theta}}$ maximiert $\ell(\boldsymbol{\theta}|\mathbf{x})$ unter der Nebenbedingung $\mathbf{C}\boldsymbol{\theta} = \mathbf{d}$. Die Struktur eines zugehörigen Tests lautet:

$$\lambda \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

Nachteil: Es ist eine numerische Maximierung von $\ell(\boldsymbol{\theta}|\mathbf{x})$ unter linearer Nebenbedingung notwendig, um $\widetilde{\boldsymbol{\theta}}$ zu erhalten.

Wald-Statistik
Die Wald-Statistik

$$w = (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{d})^\top (\mathbf{C}\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})\mathbf{C}^\top)^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{d})$$

misst die (gewichtete, quadrierte) Distanz zwischen der unrestringierten Schätzung $\mathbf{C}\widehat{\boldsymbol{\theta}}$ von $\mathbf{C}\boldsymbol{\theta}$ und dem hypothetischen Wert $\mathbf{d}$ unter H_0. Ein Test wird so konstruiert, dass

$$w \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

Vorteil gegenüber λ: Keine Berechnung von $\widetilde{\boldsymbol{\theta}}$ nötig.

Score- (oder Rao-) Statistik
Die Score-Statistik lautet

$$u = \mathbf{s}(\widetilde{\boldsymbol{\theta}}|\mathbf{x})^\top \boldsymbol{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\theta}})\mathbf{s}(\widetilde{\boldsymbol{\theta}}|\mathbf{x}),$$

wobei $\mathbf{s}(\boldsymbol{\theta}|\mathbf{x})$ die Scorefunktion des vollen Modells unter H_1 ist.

Idee: Für $\widehat{\boldsymbol{\theta}}$ gilt $\mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = \mathbf{0}$. Falls H_1 richtig ist, wird $\mathbf{s}(\widetilde{\boldsymbol{\theta}}|\mathbf{x})$ deutlich von $\mathbf{0} = \mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ verschieden sein, d.h.

$$u \text{ wird groß} \Rightarrow H_0 \text{ ablehnen.}$$

Die Statistik berechnet also den Abstand $\mathbf{s}(\widetilde{\boldsymbol{\theta}}|\mathbf{x})$ vom Ursprung, gewichtet mit $\boldsymbol{\mathcal{I}}^{-1}(\widetilde{\boldsymbol{\theta}})$.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Beispiel 3.7 (Test für einen Subvektor)
Betrachte

- $H_1: \eta = \mathbf{x}^\top \boldsymbol{\beta}$ Prädiktor im vollen GLM,
- $H_0: \eta_s = \mathbf{x}_s^\top \boldsymbol{\beta}_s$ Prädiktor im reduzierten GLM (nach Weglassen von Kovariablen).

Die Log-Likelihood $\ell(\boldsymbol{\beta}_s)$ im reduzierten Submodell werde durch $\widehat{\boldsymbol{\beta}}_s$ maximiert. Mit $\widehat{\boldsymbol{\beta}}_s$ und $\widehat{\boldsymbol{\beta}}$ lässt sich die Likelihood-Quotienten-Statistik bestimmen.

Für die Wald-Statistik ergibt sich

$$w = (\widehat{\boldsymbol{\beta}})_s^\top [(\widehat{\mathbf{A}})_s]^{-1} (\widehat{\boldsymbol{\beta}})_s,$$

dabei bezeichne $(\widehat{\boldsymbol{\beta}})_s$ die Elemente des Subvektors $\boldsymbol{\beta}_s$ in $\widehat{\boldsymbol{\beta}}$ und $(\widehat{\mathbf{A}})_s$ sei die Teilmatrix von $\widehat{\mathbf{A}} = \mathcal{I}^{-1}(\widehat{\boldsymbol{\beta}})$, die diesen Elementen entspricht.

Satz 3.5

Unter H_0 und den gleichen Regularitätsannahmen wie in Satz 3.4 gilt:

$$\lambda,\ w,\ u \overset{a}{\sim} \chi^2(s).$$

D.h. man lehnt H_0 ab, falls $\lambda, w, u > \chi^2_{1-\alpha}(s)$ ist. Für finite Stichproben besitzen λ, w, u aber unterschiedliche Werte; im Zweifelsfall sollte man λ bevorzugen.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

- *Beweis für w:* Es gilt

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N(\boldsymbol{\theta}, \mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}}))$$

und damit

$$\mathbf{C}\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N(\mathbf{C}\boldsymbol{\theta}, \mathbf{C}\mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}})\mathbf{C}^\top).$$

Unter H_0 folgt

$$\mathbf{C}\widehat{\boldsymbol{\theta}} - \underbrace{\mathbf{C}\boldsymbol{\theta}}_{\mathbf{d}} \overset{a}{\sim} N(\mathbf{0}, \underbrace{\mathbf{C}\mathcal{I}^{-1}(\widehat{\boldsymbol{\theta}})\mathbf{C}^\top}_{\mathbf{A}}),$$

also

$$\mathbf{A}^{-1/2}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{d}) \overset{a}{\sim} N(\mathbf{0}, \mathbf{I}_s)$$

und somit

$$w = (\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{d})^T \mathbf{A}^{-1}(\mathbf{C}\widehat{\boldsymbol{\theta}} - \mathbf{d}) \overset{a}{\sim} \chi^2(s).$$

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

- *Beweis für λ:* Durch Taylorentwicklung kann gezeigt werden, dass $w \overset{a}{\sim} \lambda$ und somit $\lambda \overset{a}{\sim} \chi^2(s)$. Die Beweisskizze wird hier lediglich für den Spezialfall

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

geführt (das entspricht $\mathbf{C} = \mathbf{I}_k, \mathbf{d} = \boldsymbol{\theta}_0$, $\text{rang}(\mathbf{C}) = k = \dim(\boldsymbol{\theta})$).

Eine Taylorentwicklung 2. Ordnung von $\ell(\boldsymbol{\theta}_0|\mathbf{x})$ um den unrestringierten Maximum-Likelihood-Schätzer $\widehat{\boldsymbol{\theta}}$ liefert

$$\ell(\boldsymbol{\theta}_0|\mathbf{x}) \approx \ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x})^\top(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}})^\top \mathbf{J}(\widehat{\boldsymbol{\theta}}|\mathbf{x})(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}),$$

also wegen $\mathbf{s}(\widehat{\boldsymbol{\theta}}|\mathbf{x}) = \mathbf{0}$ unter H_0

$$\begin{aligned}\lambda = 2\left(\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) - \ell(\boldsymbol{\theta}_0|\mathbf{x})\right) &\approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \mathbf{J}(\widehat{\boldsymbol{\theta}}|\mathbf{x})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = w \overset{a}{\sim} \chi^2(k).\end{aligned}$$

- *Beweis für u:* Wir nehmen denselben Speziallfall wie im Beweis für λ an, also $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Es ist unter H_0

$$\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X}) \overset{a}{\sim} N(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}_0))$$

bzw.

$$\mathcal{I}^{-1/2}(\boldsymbol{\theta}_0)\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X}) \overset{a}{\sim} N(\mathbf{0}, \mathbf{I}_k),$$

also

$$u = \mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X})^\top \underbrace{\mathcal{I}^{-\top/2}(\boldsymbol{\theta}_0)\mathcal{I}^{-1/2}(\boldsymbol{\theta}_0)}_{\mathcal{I}(\boldsymbol{\theta}_0)^{-1}} \mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X}) \overset{a}{\sim} \chi^2(k).$$

□

- Gemeinsamer Konfidenzbereich:

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{a}{\sim} \chi^2(k)$$

$$\Rightarrow \mathbb{P}_\theta \left((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_{1-\alpha}(k) \right) \overset{a}{\approx} 1 - \alpha.$$

Daraus lässt sich ein $(1 - \alpha)$-Konfidenz-Ellipsoid konstruieren.

- Komponentenweise Konfidenzintervalle für θ_j, $j = 1, \ldots, k$:

$$\frac{\widehat{\theta}_j - \theta_j}{\widehat{\sigma}_j} \overset{a}{\sim} N(0,1),$$

wobei $\widehat{\sigma}_j^2$ das j-te Diagonalelement von $\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\theta}})$ ist.

Das zugehörige approximative $(1-\alpha)$-Konfidenzintervall lautet:

$$\widehat{\theta}_j \pm z_{1-\alpha/2}\, \widehat{\sigma}_j.$$

Zum Vergleich verschiedener Modelle existieren Modellwahlkriterien, die die Güte der Anpassung, gemessen durch $\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x})$, und die Modellkomplexität $k = \dim(\boldsymbol{\theta})$ bewerten, indem sie die beiden Größen durch eine Straffunktion $\text{pen}(k)$ in einem Kompromiss zu

$$-\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \text{pen}(k)$$

zusammenführen. Dabei wird $-\ell(\widehat{\boldsymbol{\theta}}|\mathbf{x})$ klein bei guter Anpassung, $\text{pen}(k)$ groß bei stark bzw. überparametrisierten Modellen.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.3 Modellwahl

Am bekanntesten ist *Akaikes Informationskriterium*

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}|\mathbf{x}) + 2k$$

mit $\text{pen}(k) = 2k$.

Motivation: $\{f_\theta(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ parametrisiere die betrachteten Modelle und $g(\mathbf{x})$ sei die wahre Dichte für $\mathbf{X}$.
Ziel: Minimiere die Kullback-Leibler-Distanz

$$D(g, f_\theta) = \mathbb{E}_Z \left[\log \frac{g(\mathbf{Z})}{f(\mathbf{Z}|\boldsymbol{\theta})} \right] \geq 0,$$

bzw. $\mathbb{E}_X[D(g, f_{\hat{\theta}(X)})] = \mathbb{E}_X \mathbb{E}_Z[\log g(\mathbf{Z}) - \log f(\mathbf{Z}|\hat{\boldsymbol{\theta}}(\mathbf{X}))]$, wenn $\boldsymbol{\theta}$ aus gegebenen Daten $\mathbf{X} = \mathbf{x}$ geschätzt wird, $\mathbf{X}, \mathbf{Z} \overset{i.i.d}{\sim} g$.

Die Akaike Information (ohne Konstanten)
$\mathbb{E}_X \mathbb{E}_Z[-\log f(\mathbf{Z}|\hat{\theta}(\mathbf{X}))]$ ist ein prädiktives Maß für zwei unabhängige Realisationen $\mathbf{X}$ und $\mathbf{Z}$ aus g.
Zur Schätzung liegt die maximierte Loglikelihood $-\log f(\mathbf{x}|\hat{\theta}(\mathbf{x}))$ vor, die jedoch nicht erwartungstreu ist, sondern durch die doppelte Verwendung von $\mathbf{x}$ „überoptimistisch“ bzgl. der Anpassung des Modells. Unter Regularitätsbedingungen wie in Satz 3.4 lässt sich zeigen, dass der Bias genau durch $2k$ ausgeglichen wird.

Eine Alternative ist zum Beispiel das *Schwartz- (Bayes-) Informationskriterium*

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}|\mathbf{x}) + k \log n,$$

wobei n die Größe des Datensatzes ist. Für $n \geq 8$ „bestraft“ das BIC die Modellkomplexität stärker als das AIC.

Es lässt sich zeigen, dass die Modellwahl basierend auf dem BIC asymptotisch äquivalent ist zur Modellwahl basierend auf sogenannten Bayes-Faktoren, siehe Held, Kapitel 7.2, für eine Herleitung. Die Bayes-Faktoren vergleichen die Posteriori-Modellwahrscheinlichkeiten mit den Priori-Modellwahrscheinlichkeiten.