Methoden der Statistik

Mathias Trabs

1. September 2015

Inhaltsverzeichnis

Literatur

- Georgii, H.-O.: *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*, de Gruyter, 2007
- James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning (with Applications in R)*, Springer, 2013
- Lehmann, E.L., Romano, J.P.: *Testing Statistical Hypotheses*, Springer, 2005
- Lehmann, E.L. and G. Casella: *Theory of Point Estimation*, Springer, 2003
- Shao, J: *Mathematical Statistics*, Springer, 2003
- Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S-Plus, Springer, 1997
- Wasserman, L.: *All of Statistics*, Springer, 2003
- Witting, H.: *Mathematische Statistik I*, Teubner, 1985

1 Grundbegriffe der Statistik

Während die Wahrscheinlichkeitstheorie anhand eines gegebenen Modells die Eigenschaften der (zufälligen) Ereignisse untersucht, ist das Ziel der Statistik genau andersherum: Wie kann man aus den gegebenen Beobachtungen Rückschlüsse auf das Modell ziehen?

Beispiel 1.1 (Werbung)**.** Wir verwenden den "Advertising"-Datensatz aus James et al. (2013). Für 200 Märkte haben wir die Anzahl der verkauften Produkte Y sowie das jeweilige Budget für Fernsehwerbung X^F, für Radiowerbung X^R und für Zeitungsannoncen X^Z gegeben.

Betrachten wir das *Modell*

$$Y_i = aX_i^F + b + \varepsilon_i, \quad i = 1, \dots, 200,$$

wobei die zufälligen Störgrößen ε_i Marktunsicherheiten, externe Einflüsse etc. modellieren. Plausible Annahmen an das Modell sind

(i) (ε_i) sind unabhängig (näherungsweise),

(ii) (ε_i) sind identisch verteilt,

(iii) $\mathbb{E}[\varepsilon_i] = 0$ (kein systematischer Fehler)

(iv) ε_i normalverteilt (wegen ZGWS).

Naheliegende *Ziele/Fragestellungen:*

(i) Es sollen a, b anhand der Daten ermittelt werden. Ein mögliches Schätzverfahren ist der *Kleinste-Quadrate-Schätzer*

$$(\widehat{a}, \widehat{b}) := \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - aX_i - b)^2$$

(wir minimieren die Summe der quadrierten Residuen). Mit $\widehat{a}, \widehat{b}$ erhalten wir die *Regressionsgrade*

$$y = \widehat{a}x^F + \widehat{b}.$$

(ii) Sind die Modellannahmen erfüllt? Histogramm, Boxplot und QQ-Plot (Quantil-Quantil-Plot) der Residuen.

(iii) Wenn wir die Verteilung von $\widehat{a}$ kennen (Verteilungsannahme an ε nötig!), können wir Intervalle der Form $I = [\widehat{a} - c, \widehat{a} + c]$ für $c > 0$ konstruieren, so dass der tatsächlich Parameter a mit vorgegebener Wahrscheinlichkeit in I liegt.

(iv) Wir wollen *testen,* ob es einen Effekt gibt, d.h. gilt die Hypothese $H_0 : a = 0$ oder kann sie verworfen werden? Beispielsweise kann man die Hypothese verwerfen, falls $|\widehat{a}| > c$ für einen kritischen Wert $c > 0$. Um einen sinnvollen Wert zu bestimmen, benötigen wir wieder Verteilungsannahmen an die Fehler (ε_i).

Wir können das Modell auf polynomielle Regression $Y_i = a_0 + a_1 X_i^F + \cdots + a_n (X_i^F)^n + \varepsilon_i$ oder multiple Regression $Y_i = a_0 + a_1 X_i^F + a_2 X_i^R + a_3 X_i^Z + \varepsilon_i$ erweitern. Dies führt auf das Problem der Modellwahl.

Definition 1.2. Ein messbarer Raum $(\mathcal{X}, \mathscr{F})$ versehen mit einer Familie $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ von Wahrscheinlichkeitsmaßen mit einer beliebigen Parametermenge $\Theta \neq \emptyset$ heißt statistisches Experiment oder statistisches Modell. $\mathcal{X}$ heißt Stichprobenraum. Jede $(\mathcal{F}, \mathscr{S})$-messbare Funktion $Y \colon \mathcal{X} \to \mathcal{S}$ heißt Beobachtung oder Statistik mit Werten in $(\mathcal{S}, \mathscr{S})$ und induziert das statistische Modell $(\mathcal{S}, \mathscr{S}, (\mathbb{P}_\vartheta^Y)_{\vartheta \in \Theta})$. Sind die Beobachtungen $Y_1, \dots, Y_n$ für jedes $\mathbb{P}_\vartheta$ unabhängig und identisch verteilt (iid.), so nennt man $Y_1, \dots, Y_n$ eine mathematische Stichprobe.

Beispiel 1.3 (mathematische Stichprobe)**.** Für $n \in \mathbb{N}$ seien $X_1, \dots, X_n$ iid. verteilte Zufallsvariablen mit Werten in $\mathcal{X}$ und Randverteilung $X_1 \sim \mathbb{P}_\vartheta$ mit Parameter $\vartheta \in \Theta$. Dann ist der Stichprobenvektor $(X_1, \dots, X_n)$ gemäß dem Produktmaß $\mathbb{P}_\vartheta^n(\mathrm{d}x) = \prod_{i=1}^n \mathbb{P}_\vartheta(\mathrm{d}x_i)$ auf $(\mathcal{X}^n, \mathscr{F}^{\otimes n})$ verteilt.

Wir werden uns in dieser Vorlesung weitgehend mit (verallgemeinerten) linearen Modellen befassen, d.h. die Abhängigkeit der Zufallsvariablen X_i bzw. deren Verteilung vom unbekannten Parameter kann durch eine lineare Abbildung dargestellt werden.

1.1 Drei grundlegende Fragestellungen

Die meisten statistischen Fragestellungen kann man einer der drei Grundprobleme *Schätzen*, *Testen* und *Konfidenzintervalle* zuordnen. Diese werden im folgenden kurz umrissen und im Laufe der Vorlesung weiter vertieft.

1.1.1 Schätzprobleme

Ziel ist es, aufgrund der vorhandenen Beobachtungen den unbekannten Parameter im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ zu bestimmen, also einen einzelnen (bestmöglichen) Wert anzugeben (*Punktschätzung*). Damit ist ein Schätzer eine Abbildung, die nur von den Beobachtungen abhängt.

Definition 1.4. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell, $\rho : \Theta \to \mathbb{R}^d$ ein (abgeleiteter) d-dimensionaler Parameter, $d \in \mathbb{N}$. Ein Schätzer ist eine messbare Abbildung $\widehat{\rho}\colon \mathcal{X} \to \mathbb{R}^d$. Gilt $\mathbb{E}_\vartheta[\widehat{\rho}] = \rho(\vartheta)$ so heißt $\widehat{\rho}$ unverzerrt oder erwartungstreu (engl.: unbiased).

Beispiel 1.5. Seien $X_1, \dots, X_n$ eine Bernoulli-verteilte mathematische Stichprobe mit Parameter $p \in (0, 1)$. Betrachte den Schätzer $\widehat{p}_n := n^{-1}\sum_{i=1}^n X_i$. Dann gilt $\mathbb{E}_p[\widehat{p}_n] = n^{-1}\sum_{i=1}^n \mathbb{E}[X_i] = p$. Also ist $\widehat{p}_n$ erwartungstreu. Um die Streuung des Schätzers um den wahren Parameter p zu messen, berechnen wir

$$\mathrm{Var}_p(\widehat{p}_n) = n^{-2}\sum_{i=1}^n \mathrm{Var}_p(X_i) = \frac{p(1-p)}{n}.$$

Für größer werdenden Stichprobenumfang konzentriert sich also $\widehat{p}_n$ um p.

Wie gut ein Schätzer ist, wird mithilfe einer Verlustfunktion bestimmt. Diese misst den Abstand zwischen geschätztem und wahrem Parameter.

Definition 1.6. Eine Funktion $L\colon \Theta \times \mathbb{R}^d \to \mathbb{R}_+$ heißt Verlustfunktion, falls $L(\vartheta, \cdot)$ für jedes $\vartheta \in \Theta$ messbar ist. Der erwartete Verlust $R(\vartheta, \widehat{\rho}) := \mathbb{E}_\vartheta[L(\vartheta, \widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ heißt Risiko. Typische Verlustfunktionen sind

(i) der 0-1-Verlust $L(\vartheta, r) = \mathbb{1}_{\{r\neq\rho(\vartheta)\}}$,

(ii) der absolute Verlust $L(\vartheta, r) = |r - \rho(\vartheta)|$ (euklidischer Abstand im $\mathbb{R}^p$) sowie

(iii) der quadratische Verlust $L(\vartheta, r) = |r - \rho(\vartheta)|^2$.

Lemma 1.7 (Bias-Varianz-Zerlegung)**.** *Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\widehat{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ ein Schätzer des Parameters $\rho(\vartheta)$ mit $\mathbb{E}_\vartheta[|\widehat{\rho}|^2] < \infty$ für alle $\vartheta \in \Theta$. Dann gilt für den quadratischen Verlust*

$$\mathbb{E}_\vartheta\left[|\widehat{\rho} - \rho(\vartheta)|^2\right] = \mathrm{Var}_\vartheta(\widehat{\rho}) + |\underbrace{\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)}_{Bias}|^2 \quad \textit{für alle } \vartheta \in \Theta.$$

Beweis. Es gilt

$$\begin{aligned}\mathbb{E}_\vartheta\left[|\widehat{\rho} - \rho(\vartheta)|^2\right] &= \mathbb{E}_\vartheta\left[|\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}] + \mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2\right] \\ &= \mathbb{E}_\vartheta[|\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}]|^2] + 2\mathbb{E}_\vartheta\left[(\widehat{\rho} - \mathbb{E}_\vartheta[\widehat{\rho}])^\top(\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta))\right] + |\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2 \\ &= \mathrm{Var}_\vartheta(\widehat{\rho}) + |\mathbb{E}_\vartheta[\widehat{\rho}] - \rho(\vartheta)|^2. \qquad \square\end{aligned}$$

Beispiel. In der Situation von Beispiel 1.5, betrachten wir den Schätzer $\widetilde{p}_n := (\sum_{i=1}^n X_i + 1)/(n+2)$. Dieser hat den Bias

$$\mathbb{E}[\widetilde{p}_n] - p = \frac{1-2p}{n+2}$$

und die Varianz

$$\mathrm{Var}(\widetilde{p}_n) = \frac{np(1-p)}{(n+2)^2}.$$

Damit hat $\widetilde{p}_n$ einen kleineren quadratischen Fehler als $\widehat{p}_n$, wenn $|p - 1/2| \leqslant 1/\sqrt{8}$.

Bemerkung 1.8. Ein Schätzproblem, bei dem der interessierende Parameter nur endliche viele Werte annehmen kann, heißt auch *Klassifikationsproblem* und der entsprechende Schätzer heißt *Klassifizierer* (mehr dazu in Kapitel 4).

Obwohl wir in dieser Vorlesung keine Asymptotik, d.h. das Verhalten der Schätzer bei Stichprobenumfängen $n \to \infty$, behandeln, seien noch zwei weitere wichtige Grundbegriffe erwähnt.

Definition 1.9. Sei $X_1, \ldots, X_n \overset{iid.}{\sim} \mathbb{P}_\vartheta$ eine mathematische Stichprobe. Dann heißt ein Schätzer $\widehat{\rho}_n$ vom abgeleiteten Parameter $\rho(\vartheta)$ konsistent, falls

$$\widehat{\rho}_n \overset{\mathbb{P}_\vartheta}{\to} \rho(\vartheta) \quad \text{für} \quad n \to \infty.$$

Der Schätzer $\widehat{\rho}_n$ heißt asymptotisch normalverteilt, falls $\mathbb{E}[|\widehat{\rho}_n|^2] < \infty$ und

$$\frac{\widehat{\rho}_n - \mathbb{E}_\vartheta[\widehat{\rho}_n]}{\sqrt{\mathrm{Var}_\vartheta(\widehat{\rho}_n)}} \overset{\mathcal{L}}{\to} \mathcal{N}(0,1) \quad \text{unter} \quad \mathbb{P}_\vartheta.$$

Aufgrund des zentralen Grenzwertsatzes sind viele Schätzer asymptotisch normalverteilt, so auch in Beispiel 1.5. Daher kommt der Untersuchung von statistischen Modellen unter Normalverteilungsannahme eine besondere Bedeutung zu.

Zwei wichtige Konstruktionsprinzipien von Schätzern sind die Momentenmethode und Maximum-Likelihood-Schätzer:

Methode 1: Momentenmethode. Sei $X_1, \ldots, X_n$ eine mathematische Stichprobe reeller Zufallsvariablen mit $\mathbb{E}[|X_1|^d] < \infty$. Offensichtlich hängen i.A. die Momente einer Verteilung $m_k = m_k(\vartheta) := \mathbb{E}_\vartheta[X_1^k], k \in \mathbb{N}$, von ihrem Parameter $\vartheta \in \mathbb{R}^d$ ab. Aufgrund des Gesetzes der großen Zahlen ist der kanonische Schätzer von m_k gegeben durch das Stichprobenmoment $\widehat{m}_k := \frac{1}{n}\sum_{j=1}^n X_j^k$. Der Momentenschätzer $\widehat{\vartheta}$ von ϑ ist definiert als die Lösung der d-Gleichungen

$$\begin{aligned} m_1(\widehat{\vartheta}) &= \widehat{m}_1, \\ m_2(\widehat{\vartheta}) &= \widehat{m}_2, \\ \vdots \quad &\vdots \quad \vdots \\ m_d(\widehat{\vartheta}) &= \widehat{m}_d. \end{aligned}$$

Beispiel 1.10. Sei $X_1, \ldots, X_n \overset{iid.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Dann ist $m_1 = \mathbb{E}_{\mu,\sigma^2}[X_1] = \mu$ und $m_2 = \mathbb{E}_{\mu,\sigma^2}[X_1^2] = \mathrm{Var}_{\mu,\sigma^2}(X_1) + \mathbb{E}_{\mu,\sigma^2}[X_1]^2 = \sigma^2 + \mu^2$. Folglich müssen wir die Gleichungen

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^n X_j \quad \text{und} \quad \widehat{\sigma}^2 + \widehat{\mu}^2 = \frac{1}{n}\sum_{j=1}^n X_j^2$$

lösen. Bezeichnen wir das Stichprobenmittel mit $\overline{X}_n := \frac{1}{n}\sum_{j=1}^n X_j$, erhalten wir die Lösung

$$\widehat{\mu} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X}_n)^2.$$

Die Momentenmethode kann auf die Erwartungswerte allgemeinerer Funktionale verallgemeinert werden (siehe Übung ☑). Für die zweite Methode benötigen wir etwas mehr Struktur, die wir auch im weiteren Verlauf der Vorlesung immer wieder aufgreifen.

Definition 1.11. Ein statistisches Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt dominiert, falls es ein σ-endliches Maß μ gibt, so dass $\mathbb{P}_\vartheta$ absolut stetig bzgl. μ ist ($\mathbb{P}_\vartheta \ll \mu$) für alle $\vartheta \in \Theta$. Die durch ϑ parametrisierte Radon-Nikodym-Dichte

$$L(\vartheta, x) := \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x), \quad \vartheta \in \Theta, x \in \mathcal{X}$$

heißt Likelihoodfunktion, wobei diese meist als durch x parametrisierte Funktion in ϑ aufgefasst wird.

Beispiel 1.12.

(i) $\mathcal{X} = \mathbb{R}, \mathscr{F} = \mathscr{B}(\mathbb{R}), \mathbb{P}_\vartheta$ ist gegeben durch die Lebesguedichte f_ϑ, beispielsweise $\mathbb{P}_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$ oder $\mathbb{P}_\vartheta = \mathcal{U}([0, \vartheta])$. Dann ist $L(\vartheta, x) = f_\vartheta(x)$.

(ii) Jedes statistische Modell auf dem Stichprobenraum $(\mathbb{N}, \mathscr{P}(\mathbb{N}))$ oder allgemeiner auf einem abzählbaren Raum $(\mathcal{X}, \mathscr{P}(\mathcal{X}))$ ist vom Zählmaß dominiert. Die Likelihoodfunktion ist durch die Zähldichte gegeben.

(iii) Ist $\Theta = \{\vartheta_1, \vartheta_2, \dots\}$ abzählbar, so ist $\mu = \sum_i c_i \mathbb{P}_{\vartheta_i}$ mit $c_i > 0$ und $\sum_i c_i = 1$ ein dominierendes Maß.

Methode 2: Maximum-Likelihood-Prinzip. Für ein dominiertes statistisches Modell mit Likelihoodfunktion $L(\vartheta, x)$ heißt eine Statistik $\widehat{\vartheta}\colon \mathcal{X} \to \Theta$ (Θ trage eine σ-Algebra) Maximum-Likelihood-Schätzer (MLE: maximum likelihood estimator), falls

$$L(\widehat{\vartheta}, x) = \sup_{\vartheta\in\Theta} L(\vartheta, x) \quad \text{für } \mathbb{P}_\vartheta\text{-f.a. } x \in \mathcal{X} \text{ und alle } \vartheta \in \Theta.$$

Beispiel 1.13. Betrachten wir wieder eine mathematische Stichprobe $X_1, \dots, X_n$ normalverteilter Zufallsvariablen. Dann ist $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbb{P}^n_{\mu,\sigma^2})$ mit $\mathbb{P}_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$ ein vom Lebesguemaß auf $\mathbb{R}^n$ dominiertes Modell mit Likelihoodfunktion, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$L(\mu, \sigma^2; x) = (2\pi\sigma^2)^{-n/2} \prod_{j=1}^n \exp\Big(-\frac{(x_j-\mu)^2}{2\sigma^2}\Big).$$

Um den Maximum-Likelihood-Schätzer zu berechnen, nutzen wir die Monotonie des Logarithmus und betrachten

$$\log L(\mu, \sigma^2; x) = -\frac{n}{2}(\log(2\pi) + \log\sigma^2) - \sum_{j=1}^n \frac{(x_j-\mu)^2}{2\sigma^2} \to \max_{\mu,\sigma^2}.$$

Ableiten nach μ und σ^2 führt auf die Gleichungen

$$0 = \sigma^{-2}\sum_{j=1}^n (x_j - \mu), \quad \frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{j=1}^n (x_j-\mu)^2.$$

Umstellen der ersten Gleichung nach μ liefert $\widehat{\mu} = \overline{X}_n$ und Einsetzen in die zweite Gleichung ergibt $\widehat{\sigma}^2 = n^{-1}\sum_j (X_j - \overline{X}_n)^2$. Es ist leicht nachzuprüfen, dass $\widehat{\mu}$ und $\widehat{\sigma}^2$ tatsächlich das Maximierungsproblem lösen (und messbar sind). In diesem Fall stimmt der Maximum-Likelihood-Schätzer also mit dem Momentenschätzer überein.

Beispiel 1.14. Sei $X_1, \dots, X_n$ eine Poisson-verteilte mathematische Stichprobe mit Parameter $\lambda > 0$, d.h. $\mathcal{X} = \mathbb{Z}_+^n, \mathcal{F} = \mathscr{P}(\mathcal{X})$ und $\mathbb{P}_\lambda(X_1 = k) = \frac{\lambda^k e^{-k}}{k!}$. Dann ist die gemeinsame Verteilung gegeben durch

$$\mathbb{P}_\lambda(X_1 = k_1, \dots, X_n = k_n) = \frac{\lambda^{\sum_i k_i} e^{-n\lambda}}{(k!)^n}, \quad k_1, \dots, k_n \in \mathbb{Z}_+.$$

Ableiten nach λ und null setzen führt auf den Maximum-Likelihood-Schätzer $\widehat{\lambda} = \overline{X}_n$ (hinreichende Bedingung prüfen!).

1.1.2 Hypothesentests

Häufig interessiert man sich weniger für die gesamte zugrunde liegende Verteilung, als die Frage, ob eine bestimmte Eigenschaft erfüllt ist, oder nicht. Beispielsweise möchte man wissen, ob eine neue Behandlungsmethode I besser ist als die alte bisher genutzte Methode II. Aufgrund einer Beobachtung soll entschieden werden, ob die Hypothese "I ist besser als II" akzeptiert werden kann oder verworfen werden sollte.

Um derartige Fragestellungen in einem statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ zu formalisieren, wird die Parametermenge in zwei disjunkte Teilmengen Θ_0 und Θ_1 zerlegt, d.h. $\Theta = \Theta_0 \cup \Theta_1$ und $\emptyset = \Theta_0 \cap \Theta_1$. Das *Testproblem* liest sich dann als

$$H_0 : \vartheta \in \Theta_0 \quad \text{versus} \quad H_1 : \vartheta \in \Theta_1.$$

Dabei werden H_0, H_1 als *Hypothesen* bezeichnet, genauer heißt H_0 *Nullhypothese* und H_1 *Alternativhypothese* oder *Alternative*. Ein statistischer Test entscheidet nun zwischen H_0 und H_1 aufgrund einer Beobachtung $x \in \mathcal{X}$.

Definition 1.15. Ein (nicht-randomisierter) statistischer Test ist eine messbare Abbildung $\varphi\colon (\mathcal{X}, \mathscr{F}) \to (\{0,1\}, \mathscr{P}(\{0,1\}))$, wobei $\varphi(x) = 1$ heißt, dass die Nullhypothese verworfen/ die Alternative angenommen wird und $\varphi(x) = 0$ bedeutet, dass die Nullhypothese nicht verworfen wird/ akzeptiert wird. Die Menge $\{\varphi = 1\} = \{x \in \mathcal{X} : \varphi(x) = 1\}$ heißt Ablehnbereich von φ.

Allgemeiner ist ein randomisierter statistischer Test eine messbare Abbildung $\varphi\colon (\mathcal{X}, \mathscr{F}) \to ([0,1], \mathscr{B}([0,1]))$. Im Fall $\varphi(x) \in (0,1)$ entscheidet ein unabhängiges Bernoulli-Zufallsexperiment mit Erfolgswahrscheinlichkeit $p = \varphi(x)$, ob die Hypothese verworfen wird.

Testen beinhaltet mögliche Fehlerentscheidungen:

(i) Fehler 1. Art (α-Fehler, type I error): Entscheidung für H_1, obwohl H_0 wahr ist,

(ii) Fehler 2. Art (β-Fehler, type II error): Entscheidung für H_0, obwohl H_1 wahr ist.

Definition 1.16. Sei φ ein Test der Hypothese $H_0 : \vartheta \in \Theta_0$ gegen die Alternative $H_1 : \vartheta \in \Theta_1$ im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$. Die Gütefunktion von φ ist definiert als

$$\beta_\varphi\colon \Theta \to \mathbb{R}_+, \vartheta \mapsto \mathbb{E}_\vartheta[\varphi]$$

Ein Test φ erfüllt das Signifikanzniveau $\alpha \in [0,1]$ (oder φ ist Test zum Niveau α), falls $\beta_\varphi(\vartheta) \leqslant \alpha$ für alle $\vartheta \in \Theta_0$. Ein Test φ zum Niveau α heißt unverfälscht, falls $\beta_\varphi(\vartheta) \geqslant \alpha$ für alle $\vartheta \in \Theta_1$.

Somit hat ein nicht-randomisierten Test das Niveau $\alpha \in (0,1)$, falls

$$\mathbb{P}_\vartheta(\varphi = 1) \leqslant \alpha, \quad \text{für alle } \vartheta \in \Theta_0,$$

beschränkt also die Wahrscheinlichkeit des Fehlers 1. Art mit der vorgegeben oberen Schranke α. In der Regel ist es nicht möglich, die Wahrscheinlichkeiten für die Fehler 1. und 2. Art gleichzeitig zu minimieren. Daher werden diese typischerweise asymmetrisch betrachtet:

(i) Begrenzung der Fehlerwahrscheinlichkeit 1. Art durch ein vorgegebenes Signifikanzniveau α.

(ii) Unter der Maßgabe (i) wird die Wahrscheinlichkeit für Fehler 2. Art minimiert.

Eine zum Niveau α statistisch abgesicherte Entscheidung kann also immer nur zu Gunsten von H_1 getroffen werden. Daraus folgt die Merkregel "Was nachzuweisen ist, stets als Alternative H_1 formulieren".

Beispiel 1.17 (Einseitiger Binomialtest)**.** Von den 13 Todesfällen unter 55- bis 65-jährigen Arbeitern eines Kernkraftwerkes im Jahr 1995 waren 5 auf einen Tumor zurückzuführen. Die Todesursachenstatistik 1995 weist aus, dass Tumore bei etwa $1/5$ aller Todesfälle die Ursache in der betreffenden Altersklasse (in der Gesamtbevölkerung) darstellen. Ist die beobachtete Häufung von tumorbedingten Todesfällen signifikant zum Niveau 5%?

Bezeichne X die Anzahl der Tumortoten unter $n = 13$ Todesfällen. Dann ist das statistische Modell gegeben durch $\mathcal{X} = \{0, \dots n\}, \mathscr{F} = \mathscr{P}(\mathcal{X})$ und $\mathbb{P}_p = Bin(13, p)$ mit Parameter $p \in [0, 1]$ und das Testproblem ist gegeben durch

$$H_0 : p \leqslant 1/5 \quad \text{versus} \quad H_1 : p > 1/5.$$

Ziel ist ein nicht-randomisierter Test zum Niveau $\alpha = 0,05$. Naheliegenderweise konstruieren wir $\varphi(x) = \mathbb{1}_{\{x>c\}}$ wobei der kritische Wert $c > 0$ so gewählt wird, dass $\sup_{p\leqslant 1/5} \mathbb{P}_p(X > c) \leqslant \alpha$. Um eine möglichst große Güte zu erreichen, sollte c unter dieser Nebenbedingung möglichst klein gewählt werden. Für $k \in \mathcal{X}$ gilt

$$\mathbb{P}_p(X \leqslant k) = \sum_{l=0}^{k} \binom{13}{l} p^l (1-p)^{13-l}.$$

Da $p \mapsto \mathbb{P}_p(X \leqslant k)$ für alle $k \in \mathcal{X}$ monoton fallend auf $[0, 1]$ ist (ableiten), folgt $\sup_{p\leqslant 1/5} \mathbb{P}_p(X > c) = \mathbb{P}_{1/5}(X > c)$. Wegen

$$\mathbb{P}_{1/5}(X \leqslant 4) \approx 0,901 \quad \text{und} \quad \mathbb{P}_{1/5}(X \leqslant 5) \approx 0,970,$$

wählen wir $c = 5$. Somit kann die Hypothese zum Niveau $0,05$ nicht verworfen werden. Die Gütefunktion von φ

$$\beta_\varphi(p) = \mathbb{P}_p(X > 5) = \sum_{l=6}^{13} \binom{13}{l} p^l (1-p)^{13-l}, \quad p \in [0, 1],$$

ist monoton wachsend und somit ist φ auch unverfälscht.

Dieses Beispiel führt uns auf ein allgemeines Konstruktionsprinzip von Tests einer Hypothese $H_0 : \vartheta \in \Theta_0$ vs. $H_1 : \vartheta \in \Theta_1$ mit $\Theta_0 \neq \emptyset$ und $\Theta_1 = \Theta \setminus \Theta_0$.

Methode 3: Teststatistiken. Für Ablehnbereiche $(\Gamma_\alpha)_{\alpha\in(0,1)} \subseteq \mathscr{B}(\mathbb{R})$ und eine Teststatistik $T\colon (\mathcal{X}, \mathscr{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ sei ein Test gegeben durch

$$\varphi(x) = \mathbb{1}_{\{T(x)\in\Gamma_\alpha\}}, \quad x \in \mathcal{X}. \tag{1.1}$$

Oft werden die Ablehnbereiche als Intervalle $\Gamma_\alpha = (c_\alpha, \infty)$ konstruiert für kritische Werte

$$c_\alpha = \inf\Big\{c \in \mathbb{R} : \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) > c) \leqslant \alpha\Big\}, \quad \alpha \in (0, 1). \tag{1.2}$$

Ist $\Theta_0 = \{\vartheta_0\}$ einelementig, dann sind die kritischen Werte genau das $(1-\alpha)$-*Quantil* der Verteilung von T unter $\mathbb{P}_{\vartheta_0}$. Ein wichtiges Konzept in der Testtheorie, insbesondere in Anwendungen, sind die p-Werte.

Definition 1.18. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und der Test φ der Hypothese $H_0 : \vartheta \in \Theta_0 \neq \emptyset$ gegeben durch (1.1). Dann ist der p-Wert einer Realisierung $x \in \mathcal{X}$ bezüglich φ definiert als

$$p_\varphi(x) = \inf_{\alpha : T(x)\in\Gamma_\alpha} \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha).$$

Statt nur zu prüfen, ob ein Test eine Hypothese akzeptiert oder ablehnt, gibt der p-Wert (die Signifikanzwahrscheinlichkeit) das kleinste Signifikanzniveau an, zu dem eine Hypothese abgelehnt würde. Damit gibt der p-Wert Aufschluss darüber "wie stark" die Daten der Hypothese widersprechen.

Satz 1.19. *Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und sei φ ein Test der Hypothese $H_0 : \vartheta \in \Theta_0 \neq \emptyset$ gegeben durch $\varphi = \mathbb{1}_{\{T>c_\alpha\}}$ für eine Teststatistik $T : \mathcal{X} \to \mathbb{R}$ und kritische Werten $(c_\alpha)_{\alpha\in(0,1)}$ aus (1.2). Dann ist der p-Wert einer Realisierung $x \in \mathcal{X}$ bezüglich φ gegeben durch*

$$p_\varphi(x) = \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta(T(X) \geqslant t^*) \quad \textit{mit} \quad t^* := T(x).$$

*Sei $\alpha \in (0,1)$ ein **fest vorgegebenes** Niveau. Ist die Verteilung $\mathbb{P}_\vartheta^T$ stetig für alle $\vartheta \in \Theta_0$, gilt*

$$\varphi(x) = 1 \quad \Longleftrightarrow \quad p_\varphi(x) < \alpha \quad \mathbb{P}_\vartheta - f.s. \text{ für alle } \vartheta \in \Theta_0.$$

Ist $\mathbb{P}_\vartheta^T$ (topologisch) diskret verteilt für alle $\vartheta \in \Theta_0$, gilt

$$\varphi(x) = 1 \quad \Longleftrightarrow \quad p_\varphi(x) \leqslant \alpha \quad \mathbb{P}_\vartheta - f.s. \text{ für alle } \vartheta \in \Theta_0.$$

Beweis. Definiere $\mathbb{P}_0 := \sup_{\vartheta\in\Theta_0} \mathbb{P}_\vartheta$. Da $c \mapsto \mathbb{P}_0(T > c)$ monoton fallend ist, gilt

$$p_\varphi(x) = \inf_{\alpha : t^* > c_\alpha} \mathbb{P}_0(T > c_\alpha) \geqslant \mathbb{P}_0(T \geqslant t^*).$$

Da $c_\alpha < t^*$ äquivalent zur Existenz eines $c < t^*$ mit $\mathbb{P}_0(T > c) \leqslant \alpha$ ist, folgt aus $\mathbb{P}_0(T > c_\alpha) \leqslant \alpha$ (Rechtsstetigkeit der Verteilungsfunktion), dass

$$p_\varphi(x) \leqslant \inf\{\alpha : c_\alpha < t^*\} \leqslant \inf\big\{\alpha : \mathbb{P}_0\big(\cap_{c<t^*} \{T > c\}\big) \leqslant \alpha\big\} = \mathbb{P}_0(T \geqslant t^*).$$

Zusammen erhalten wir $p_\varphi(x) = \mathbb{P}_0(T \geqslant t^*)$.

Sei nun α fest und T zunächst stetig verteilt. Aus $p_\varphi(x) = \mathbb{P}_0(T \geqslant t^*) < \alpha$ und $\lim_{c\uparrow t^*} \mathbb{P}_0(T \in (c, t^*)) = 0$ folgt $\mathbb{P}_0(T > c) \leqslant \alpha$ für ein $c < t^*$. Dann muss aber $\varphi(x) = 1$ gelten. Andersherum gilt

$$\varphi(x) = 1 \quad \Rightarrow \quad \exists c < t^* : \mathbb{P}_0(T \geqslant t^*) \leqslant \alpha - \mathbb{P}_0(T \in (c, t^*)).$$

Dabei gilt $\mathbb{P}_\vartheta(T \in (c, t^*)) = \mathbb{P}_\vartheta(T \in (c, T(x))) > 0$ für $\mathbb{P}_\vartheta$-f.a. $x \in \mathcal{X}$ und für alle $\vartheta \in \Theta_0$. Ist T diskret verteilt, bleibt zu bemerken, dass $p_\varphi(x) = \mathbb{P}_0(T \geqslant t^*) = \mathbb{P}_0(T > c)$ für ein $c < t^*$. □

Bemerkung 1.20.

(i) Der Vorteil von p-Werten ist, dass sie unabhängig von einem a priori festgesetzten Signifikanzniveau α berechnet werden können. Deshalb werden in allen gängigen Statistik-Softwaresystemen statistische Hypothesentests über die Berechnung von p-Werten implementiert.

(ii) **Warnung**: Alle Rahmenbedingungen des Experiments, insbesondere also das Signifikanzniveau, müssen vor dessen Durchführung festgelegt werden! Ein Signifikanzniveau darf nicht a posteriori aufgrund der erzielten p-Werte festgelegt werden. Dies widerspricht richtiger statistischer Praxis! Insbesondere wäre α eine Zufallsvariable (als Funktion in den Beobachtungen) und obiger Satz kann nicht angewendet werden.

(iii) Der p-Wert gibt eine Antwort auf die Frage: "Wie wahrscheinlich sind die gemessenen Daten, gegeben, dass die Nullhypothese stimmt?" (und **nicht** auf die Frage "Wie wahrscheinlich ist es, dass die Nullhypothese wahr ist, gegeben den gemessenen Daten?")

Beispiel 1.21. Geburten in Berlin:

(i) *Hypothese*: Es werden genauso viele Jungen wie Mädchen geboren.
Sind von $n \in \mathbb{N}$ Geburten $w \leqslant n$ Mädchen zur Welt gekommen, ist das statistische Modell gegeben durch den Stichprobenraum $\mathcal{X} = \{0, \dots, n\}$ und somit $(\mathcal{X}, \mathscr{P}(\mathcal{X}), (\mathbb{P}_\vartheta)_{\vartheta \in [0,1]})$ mit Binomialverteilungen $\mathbb{P}_\vartheta = Bin(n, \vartheta)$. Die Hypothese führt auf das zweiseitige Testproblem

$$H_0 : \vartheta = 1/2 \quad \text{versus} \quad H_1 : \vartheta \neq 1/2,$$

wobei $w \in \mathcal{X}$ beobachtet wird. Wir setzten das Niveau $\alpha = 0,05$. Die Teststatistik $T(w) = |\frac{w}{n} - \vartheta|$ führt auf einen *zweiseitigen Binomialtest.*

(ii) *Hypothese*: Höchstens die Hälfte der geborenen Kinder hat nicht verheiratete Eltern.
Von $n \in \mathbb{N}$ geboren Kindern haben $v \leqslant n$ verheiratete Eltern. Mit $(\mathcal{X}, \mathscr{P}(\mathcal{X}), (\mathbb{P}_\vartheta)_{\vartheta \in [0,1]})$ wie oben betrachten wir hier das einseitige Testproblem

$$H_0 : \vartheta \leqslant 1/2 \quad \text{versus} \quad H_1 : \vartheta > 1/2,$$

wobei $v \in \mathcal{X}$ beobachtet wird. Das Niveau $\alpha = 0,05$ zusammen mit der Teststatistik $T(w) = \frac{w}{n} - \vartheta$ führt auf einen *einseitigen Binomialtest.*

Bemerkung 1.22. Bei großen Stichprobenumfängen ist es sinnvoll, einen *Gauß-Test* für geeignet normalisierter Teststatistik zu verwenden, um Binomialtest zu approximieren: Für $\vartheta \in (0,1)$ normalisieren wir die Beobachtung $X \sim Bin(n, \vartheta)$ durch $Y := \frac{X - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$. Aus dem Zentralen Grenzwertsatz folgt dann für eine standardnormalverteilte Zufallsvariable $Z \sim \mathcal{N}(0,1)$, dass

$$\begin{aligned} \mathbb{P}_\vartheta(T(X) > c_\alpha) &= \mathbb{P}_\vartheta\Big(\frac{|X - n\vartheta|}{\sqrt{n\vartheta(1-\vartheta)}} > \sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big) \\ &\overset{n\to\infty}{\longrightarrow} \mathbb{P}\Big(|Z| > \sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big) \\ &= \frac{1}{2}\Big(1 - \Phi\Big(\sqrt{\frac{n}{\vartheta(1-\vartheta)}} c_\alpha\Big)\Big) \overset{!}{=} \alpha, \end{aligned}$$

Mit der Verteilungsfunktion $\Phi(x) = \mathbb{P}(Z \leqslant x)$. Folglich wählen wir $c_\alpha = \sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}} q_{1-2\alpha} = \sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}} \Phi^{-1}(1-2\alpha)$ mit $\vartheta = \vartheta_0$ unter H_0.

1.1.3 Konfidenzmengen (Bereichsschätzung)

Während ein (Punkt-)Schätzer einen einzelnen Wert angibt, möglichst in der Nähe des wahren Parameters, um Rückschlüsse auf das zugrunde liegende Modell zu ziehen, geben Konfidenzbereiche ein Intervall an, in dem der Parameter mit gegebener Wahrscheinlichkeit liegt.

Definition 1.23. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein statistisches Modell mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$. Eine mengenwertige Abbildung $C\colon \mathcal{X} \to \mathscr{P}(\mathbb{R}^d)$ heißt Konfidenzmenge zum Konfidenzniveau $1-\alpha$ (oder zum Irrtumsniveau α) für $\alpha \in (0,1)$, falls die Messbarkeitsbedingung $\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\} \in \mathscr{F}$ für alle $\vartheta \in \Theta$ erfüllt ist und es gilt

$$\mathbb{P}_\vartheta(\rho(\vartheta) \in C) = \mathbb{P}_\vartheta\big(\{x \in \mathcal{X} : \rho(\vartheta) \in C(x)\}\big) \geqslant 1 - \alpha \quad \text{für alle } \vartheta \in \Theta.$$

Im Fall $d = 1$ und falls $C(x)$ für jedes $x \in \mathcal{X}$ ein Intervall ist, heißt C Konfidenzintervall.

Beachte, dass $\rho(\vartheta)$ fix ist, während C zufällig ist. Man muss Konfidenzmengen also wie folgt *interpretieren*: Werden in m unabhängigen Experimenten für (verschiedene) Parameter Konfidenzmengen zum Niveau $0,95$ konstruiert, dann liegt der unbekannte Parameter in 95% der Fälle im der jeweiligen Konfidenzmenge (für m groß genug; starkes Gesetz der großen Zahlen).

Ein verbreitetes Konstruktionsprinzip für die Konfidenzintervalle ist die Verwendung eines Schätzers und dessen Verteilung, wie im nächsten Beispiel illustriert.

Beispiel 1.24. Im Bernoulli-Experiment von Beispiel 1.5 gilt für $C_n := [\widehat{p}_n - \varepsilon_n, \widehat{p}_n + \varepsilon_n]$

$$\mathbb{P}_p(p \in C_n) = \mathbb{P}_p(|\widehat{p}_n - p| < \varepsilon_n) = \mathbb{P}_p\Big(\big|\sum_{i=1}^n (X_i - p)\big| < n\varepsilon_n\Big) \overset{!}{\geqslant} 1 - \alpha.$$

Da $\sum_{i=1}^n X_i \sim Bin(n,p)$ können wir ε_n mithilfe der Quantile der Binomialverteilung bestimmen. Für große n könnte man wieder eine Normalapproximation verwenden. Das resultierende Konfidenzintervall besitzt dann aber nur asymptotisch das Niveau $1-\alpha$.

Eine alternative Konstruktion von Konfidenzmengen bietet folgender Korrespondenzsatz:

Satz 1.25. *Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Dann gilt:*

(i) Liegt für jedes $\vartheta_0 \in \Theta$ ein Test φ_{ϑ_0} der Hypothese $H_0 : \vartheta = \vartheta_0$ zum Signifikanzniveau α vor, so definiert $C(x) = \{\vartheta \in \Theta : \varphi_\vartheta(x) = 0\}$ eine Konfidenzmenge zum Konfidenzniveau $1-\alpha$.

(ii) Ist C eine Konfidenzmenge zum Niveau $1-\alpha$, dann ist $\varphi_{\vartheta_0}(x) = 1 - \mathbb{1}_{C(x)}(\vartheta_0)$ ein Niveau-α-Test der Hypothese $H_0 : \vartheta = \vartheta_0$.

Beweis. Nach Konstruktion erhält man in beiden Fällen,

$$\forall \vartheta \in \Theta : \forall x \in \mathcal{X} : \varphi_\vartheta(x) = 0 \quad \Longleftrightarrow \quad \vartheta \in C(x).$$

Damit ist φ_ϑ ein Test zum Niveau α für alle ϑ genau dann, wenn

$$1 - \alpha \leqslant \mathbb{P}_\vartheta(\varphi = 0) = \mathbb{P}_\vartheta(\{x : \vartheta \in C(x)\})$$

und somit ist C eine Konfidenzmenge zum Niveau α. □

Beispiel 1.26. Mit Hilfe des Korrespondenzsatzes können wir ein Konfidenzintervall zum Niveau $0,95$ für die Geburtswahrscheinlichkeit von Mädchen in Berlin berechnen. Im Modell aus Beispiel 1.21(i) ist das Konfidenzintervall gegeben durch

$$C(w) = \{\vartheta \in [0,1] : |\frac{w}{n} - \vartheta| \leqslant c_{0,05}\} = \{\vartheta \in [0,1], p_\varphi(w) > 0,05\},$$

wobei $p_\varphi(w)$ den zu φ gehörigen p-Wert der Realisierung w bezeichnet. Ist C sogar ein Konfidenzintervall? (Übung ☑).

1.2 Minimax- und Bayesansatz

Wir haben bereits verschiedene Schätzmethoden, wie den Maximum-Likelihood-Schätzer oder die Momentenmethode kennen gelernt. Natürlich gibt es noch viel mehr Konstruktionen. Wie sollte eine Methode anhand des gegeben Schätzproblems ausgewählt werden? Sei also $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$ und Verlustfunktion L. Als mögliches Vergleichskriterium käme die Risikofunktion $R(\vartheta, \widehat{\rho}) = \mathbb{E}_\vartheta[L(\vartheta, \widehat{\rho})]$ eines Schätzers $\widehat{\rho}$ in Frage. Beachte jedoch folgendes Beispiel:

Beispiel 1.27. Sei $X \sim \mathcal{N}(\mu, 1), \mu \in \mathbb{R}$, und $L(\mu, \widehat{\mu}) = (\widehat{\mu} - \mu)^2$. Betrachte die zwei Schätzer $\widehat{\mu}_1 = X$ und $\widehat{\mu}_2 = 5$. Die Risiken sind dann gegeben durch

$$R(\mu, \widehat{\mu}_1) = \mathbb{E}_\vartheta[(X-\mu)^2] = 1 \quad \text{und} \quad R(\mu, \widehat{\mu}_2) = (5-\mu)^2.$$

Damit hat $\widehat{\mu}_1$ kleineres Risiko als $\widehat{\mu}_2$ genau dann, wenn $\mu \notin [4,6]$.

Definition 1.28. Im statistischen Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ mit abgeleitetem Parameter $\rho\colon \Theta \to \mathbb{R}^d$ und Verlustfunktion L, heißt ein Schätzer $\widehat{\rho}$ minimax, falls

$$\sup_{\vartheta\in\Theta} R(\vartheta, \widehat{\rho}) = \inf_{\widetilde{\rho}} \sup_{\vartheta\in\Theta} R(\vartheta, \widetilde{\rho}),$$

wobei sich das Infimum über alle Schätzer (d.h. messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ erstreckt.

Definition 1.29. Der Parameterraum Θ trage eine σ-Algebra $\mathscr{F}_\Theta$, die Verlustfunktion L sei produktmessbar und $\vartheta \mapsto \mathbb{P}_\vartheta(B)$ sei messbar für alle $B \in \mathscr{F}$. Die a-priori-Verteilung π des Parameters ϑ ist gegeben durch ein Wahrscheinlichkeitsmaß auf $(\Theta, \mathscr{F}_\Theta)$. Das zu π assoziierte Bayesrisiko eines Schätzers $\widehat{\rho}$ ist

$$R_\pi(\widehat{\rho}) := \mathbb{E}_\pi[R(\vartheta, \widehat{\rho})] = \int_\Theta \int_{\mathcal{X}} L(\vartheta, \widehat{\rho}(x)) \mathbb{P}_\vartheta(\mathrm{d}x) \pi(\mathrm{d}\vartheta).$$

Der Schätzer $\widehat{\rho}$ heißt Bayesschätzer oder Bayes-optimal (bezüglich π), falls

$$R_\pi(\rho) = \inf_{\widetilde{\rho}} R_\pi(\widetilde{\rho}),$$

wobei sich das Infimum über alle Schätzer (d.h. messbaren Funktionen) $\widetilde{\rho}\colon \mathcal{X} \to \mathbb{R}^d$ erstreckt.

Während ein Minimaxschätzer den maximal zu erwartenden Verlust minimiert, kann das Bayesrisiko als ein (mittels π) gewichtetes Mittel der zu erwartenden Verluste angesehen werden. Alternativ wird π als die subjektive Einschätzung der Verteilung des zugrundeliegenden Parameters interpretiert.

Beispiel 1.27 (fortgesetzt). Offensichtlich kann $\widehat{\mu}_2$ kein Minimaxschätzer sein. Zunächst ist es aber nicht klar, ob es einen besseren Schätzer als $\widehat{\mu}_2$ gibt. Tatsächlich werden wir später beweisen, dass $\widehat{\mu}_1$ minimax ist. Unter der a-priori-Verteilung $\mu \sim \pi = \mathcal{U}([4,6])$ hat jedoch $\widehat{\mu}_2$ das kleinere Bayesrisiko $R_\pi(\widehat{\mu}_2) = \frac{1}{3} < 1 = R_\pi(\widehat{\mu}_1)$.

Das Bayesrisiko kann auch als insgesamt zu erwartender Verlust in folgendem Sinne verstanden werden: Definiere $\Omega := \mathcal{X} \times \Theta$ und die gemeinsame Verteilung von Beobachtung und Parameter $\widetilde{\mathbb{P}}$ auf $(\mathcal{X} \times \Theta, \mathscr{F} \otimes \mathscr{F}_\Theta)$ gemäß $\widetilde{\mathbb{P}}(\mathrm{d}x, \mathrm{d}\vartheta) = \mathbb{P}_\vartheta(\mathrm{d}x)\pi(\mathrm{d}\vartheta)$. Bezeichnen X und T die Koordinatenprojektionen von Ω auf $\mathcal{X}$ bzw. Θ, dann gilt $R_\pi(\widehat{\rho}) = \mathbb{E}_{\widetilde{\mathbb{P}}}[L(T, \widehat{\rho}(X))]$.

Wiederholung: Auf einem Wahrscheinlichkeitsraum $(\Omega,\mathscr{F},\mathbb{P})$ ist die bedingte Wahrscheinlichkeit eines Ereignisses $A \in \mathscr{F}$ gegeben $B \in \mathscr{F}$ mit $\mathbb{P}(B) > 0$ definiert als $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$. Sei $\Omega = \bigcup_{i \in I} B_i$ eine abzählbare Zerlegung in paarweise disjunkte Ereignisse $B_i \in \mathscr{F}$, dann besagt die **Bayesformel** für jedes $A \in \mathscr{F}$ mit $\mathbb{P}(A) > 0$ und alle $k \in I$

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(B_k)\mathbb{P}(A|B_k)}{\sum_{i \in I} \mathbb{P}(B_i)\mathbb{P}(A|B_i)}.$$

Mittels bedingten Erwartungswerten (Stochastik II) kann diese Formel auf Dichten ausgedehnt werden.

Definition 1.30. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes statistisches Modell mit Dichten $f_{X|T=\vartheta} := \frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}$. Sei π eine a-priori-Verteilung auf $(\Theta, \mathscr{F}_\Theta)$ mit Dichte f_T bzgl. einem Maß ν. Ist $f_{X|T=\cdot}\colon \mathcal{X} \times \Theta \to \mathbb{R}_+$ $(\mathscr{F} \otimes \mathscr{F}_\Theta)$-messbar, dann ist die a-posteriori-Verteilung des Parameters gegeben der Beobachtung $X = x$ definiert durch die ν-Dichte

$$f_{T|X=x}(\vartheta) = \frac{f_{X|T=\vartheta}(x) f_T(\vartheta)}{\int_\Theta f_{X|T=t}(x) f_T(t) \nu(\mathrm{d}t)}, \quad \vartheta \in \Theta \quad (\widetilde{\mathbb{P}}^X\text{-f.ü.}). \tag{1.3}$$

Das a-posteriori-Risiko eines Schätzers $\widehat{\rho}$ gegeben $X = x$ ist definiert durch

$$R_\pi(\widehat{\rho}|x) = \int_\Theta L(\vartheta, \widehat{\rho}(x)) f_{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta).$$

Beachte, dass im Nenner in (1.3) die Randdichte $f_X = \int_\Theta f_{X|T=t}(\cdot) f_T(t) \nu(\mathrm{d}t)$ bzgl. μ von X in $(\mathcal{X} \times \Theta, \mathscr{F} \otimes \mathscr{F}_\Theta, \widetilde{\mathbb{P}})$ steht, so dass der Nenner in (1.3) für $\widetilde{\mathbb{P}}^X$-f.a. $x \in \mathcal{X}$ größer als null ist.

Beispiel 1.31. Setze $\Theta = \{0,1\}, L(\vartheta, r) = |\vartheta - r|$ (0-1-Verlust) und betrachte eine a-priori-Verteilung π mit $\pi(\{0\}) =: \pi_0$ und $\pi(\{1\}) =: \pi_1 = 1 - \pi_0$. Die Wahrscheinlichkeitsmaße $\mathbb{P}_0$ und $\mathbb{P}_1$ mögen Dichten p_0 und p_1 bzgl. einem Maß μ besitzen (z.B. $\mu = \mathbb{P}_0 + \mathbb{P}_1$). Dann ist die a-posteriori-Verteilung durch die Zähldichte

$$f_{T|X=x}(i) = \frac{\pi_i p_i(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}, \quad i = 0,1 \quad (\widetilde{\mathbb{P}}^X\text{-f.ü.})$$

gegeben. Damit ist das a-posteriori-Risiko eines Schätzers $\widehat{\vartheta}\colon \mathcal{X} \to \{0,1\}$ gegeben durch

$$R_\pi(\widehat{\vartheta}|x) = \frac{\widehat{\vartheta}(x)\pi_0 p_0(x) + (1 - \widehat{\vartheta}(x))\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}.$$

Satz 1.32. *Es gelten die Bedingungen der vorangegangenen Definition. Für das Bayesrisiko eines Schätzers $\widehat{\rho}$ gilt*

$$R_\pi(\widehat{\rho}) = \int R_\pi(\widehat{\rho}|x) f_X(x) \mu(\mathrm{d}x).$$

Minimiert $\widehat{\rho}(x)$ für $\widetilde{\mathbb{P}}^X$-f.a. das a-posteriori-Risiko $\min_{t \in \mathrm{ran}(\rho)} R_\pi(t|x)$, dann ist $\widehat{\rho}$ Bayesschätzer.

Beweis. Aus (1.3) folgt $f_{T|x=x}(\vartheta) f_X(x) = f_{X|T=\vartheta}(x) f_T(\vartheta)$. Der Satz von Fubini ergibt

$$\begin{aligned} R_\pi(\widehat{\rho}) &= \int_\Theta \int_{\mathcal{X}} L(\vartheta, \widehat{\rho}(x)) \mathbb{P}_\vartheta(\mathrm{d}x) \pi(\mathrm{d}\vartheta) \\ &= \int_\Theta \int_{\mathcal{X}} L(\vartheta, \widehat{\rho}(x)) f_{T|x=x}(\vartheta) f_X(x) \mu(\mathrm{d}x) \nu(\mathrm{d}\vartheta) = \int_{\mathcal{X}} R_\pi(\widehat{\rho}|x) \mu(\mathrm{d}x). \end{aligned}$$

□

Korollar 1.33. *Unter quadratischem Verlust ist der Bayesschätzer gegeben durch*

$$\widehat{\rho}(x) = \int_\Theta \rho(\vartheta) f_{T|X=x}(\vartheta) \nu(\mathrm{d}\vartheta) =: \mathbb{E}[\rho(\vartheta)|X = x].$$

Der Bayesschätzer bzgl. absolutem Verlust ist gegeben durch den Median der a-posteriori-Verteilung. Für den 0-1-Verlust ist der Bayesschätzer der Modus der a-posteriori-Verteilung.

Beweis. Übung ☑. □

Methode 4: Bayesschätzer. Durch die Wahl einer Verlustfunktion und einer a-priori-Verteilung im statistischen Modell erhalten wir nach Berechnung der a-posteriori-Verteilung und durch das vorangegangene Korollar einen expliziten Bayesschätzer.

Beispiel 1.34. Sei $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ eine mathematische Stichprobe mit bekanntem $\sigma^2 > 0$ und a-priori-Verteilung $\mu \sim \mathcal{N}(a, b^2)$. Mittels Bayesformel kann die a-posteriori-Verteilung für eine Realisierung $x = (x_1, \ldots, x_n)$ berechnet werden:

$$\begin{aligned} f_{T|X=x}(\mu) &\sim f_{X|T=\mu}(x) f_T(\mu) \\ &\sim \exp\Big(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\Big) \exp\Big(-\frac{(\mu - a)^2}{2b^2}\Big) \\ &\sim \exp\Big(-\frac{\mu^2 - 2\mu\overline{x}_n}{2\sigma^2/n} - \frac{\mu^2 - 2a\mu}{2b^2}\Big) \\ &\sim \exp\Big(-\frac{(b^2 + \sigma^2/n)\mu^2 - 2\mu(b^2\overline{x}_n + a\sigma^2/n)}{2b^2\sigma^2/n}\Big) \\ &\sim \exp\Big(-\frac{1}{2}\Big(\frac{n}{\sigma^2} + \frac{1}{b^2}\Big)\Big(\mu - \frac{b^2\overline{x}_n}{b^2 + \sigma^2/n} - \frac{a\sigma^2/n}{b^2 + \sigma^2/n}\Big)^2\Big). \end{aligned}$$

Gegeben der Beobachtung X ist ϑ also a-posteriori verteilt gemäß

$$\mathcal{N}\Big(\frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X}_n - \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a, \Big(\frac{n}{\sigma^2}+\frac{1}{b^2}\Big)^{-1}\Big).$$

Der Bayesschätzer bzgl. quadratischem Verlust, gegeben durch den a-posteriori Mittelwert, ist damit

$$\widehat{\vartheta}_n = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X}_n - \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a.$$

Bemerkung 1.35. Erhalten wir bei Wahl einer Klasse von a-priori-Verteilungen für ein statistisches Modell dieselbe Klasse (i.A. mit anderen Parametern) als a-posteriori-Verteilung zurück, so nennt man die entsprechenden Verteilungsklassen konjugiert. Im obigen Beispiel haben wir gesehen, dass die Normalverteilungen zur den Normalverteilungen konjugiert sind (genauer müsste man sagen, dass für unbekannten Mittelwert in der Normalverteilung a-priori Normalverteilungen konjugiert sind). Als weiteres Beispiel sind die Beta-Verteilungen zur Binomialverteilung konjugiert sind (siehe Übung ☑). In diesen (Einzel-)Fällen ist es besonders einfach, die Bayesschätzer zu konstruieren. Für komplexere Modelle werden häufig computer-intensive Methoden wie MCMC (Markov Chain Monte Carlo) verwendet, um die a-posteriori-Verteilung zu berechnen (Problem: i.A. hochdimensionale Integration).

Lemma 1.36. *Unter den Bedingungen der vorangegangen Definition gilt für jeden Schätzer* $\widehat{\rho}$

$$\sup_{\vartheta\in\Theta} R(\vartheta,\widehat{\rho}) = \sup_{\pi} R_\pi(\widehat{\rho}),$$

wobei sich das zweite Supremum über alle a-priori-Verteilungen π *erstreckt. Insbesondere ist das Risiko eines Bayesschätzers stets kleiner oder gleich dem Minimaxrisiko.*

Beweis. Natürlich gilt $R_\pi(\widehat{\rho}) = \int_\Theta R(\vartheta,\widehat{\rho})\pi(\mathrm{d}\vartheta) \leqslant \sup_{\vartheta\in\Theta} R(\vartheta,\widehat{\rho})$. Durch Betrachtung der a-priori-Verteilung δ_ϑ folgt daher die Behauptung. □

Durch dieses Lemma können wir untere Schranken für das Minimaxrisiko durch das Risiko von Bayesschätzern abschätzen. Mögliche Anwendungen illustriert folgender Satz.

Satz 1.37. *Sei* $X_1,\dots,X_n$ *eine* $\mathcal{N}(\mu,\sigma^2)$*-verteilte mathematische Stichprobe mit unbekanntem* $\mu\in\mathbb{R}$ *und bekanntem* $\sigma^2>0$*. Bezüglich quadratischem Risiko ist das arithmetische Mittel* $\overline{X}_n$ *ein Minimaxschätzer von* μ*.*

Beweis. Wir betrachten a-priori-Verteilungen $\mu\sim\pi=\mathcal{N}(0,b^2)$. Nach Beispiel 1.34 ist die a-posteriori-Verteilung

$$\mathcal{N}\Big(\frac{b^2\overline{X}_n}{b^2+\frac{\sigma^2}{n}}, \Big(\frac{n}{\sigma^2}+b^{-2}\Big)^{-1}\Big),$$

der Bayesschätzer bzgl. quadratischem Risiko ist gegeben durch den a-posteriori-Erwartungswert $\widehat{\mu}_n = b^2\overline{X}_n/(b^2+\sigma^2n^{-1})$ und dessen a-posteriori-Risiko ist gegeben durch die Varianz der a-posteriori-Verteilung. Ist f_X die Randdichte von X von $\widetilde{\mathbb{P}}$, folgt aus Satz 1.32

$$\begin{aligned} R_\pi(\widehat{\mu}_n) &= \int_{\mathbb{R}^n} \mathrm{Var}_{T|X=x}(\mu) f_X(x)\mathrm{d}x \\ &= \int_{\mathbb{R}^n} \big(n\sigma^{-2}+b^{-2}\big)^{-1} f_X(x)\mathrm{d}x = \big(n\sigma^{-2}+b^{-2}\big)^{-1}. \end{aligned}$$

Somit können wir das Minimaxrisiko nach unten abschätzen:

$$\begin{aligned} \inf_{\widetilde{\mu}}\sup_{\mu\in\mathbb{R}} R(\mu,\widetilde{\mu}) = \inf_{\widetilde{\mu}}\sup_{\pi} R_\pi(\widetilde{\mu}) &\geqslant \inf_{\widetilde{\mu}}\sup_{b>0} R_{\mathcal{N}(0,b^2)}(\widetilde{\mu}) \\ &\geqslant \sup_{b>0}\inf_{\widetilde{\mu}} R_{\mathcal{N}(0,b^2)}(\widetilde{\mu}) = \sup_{b>0}\big(n\sigma^2+b^{-2}\big)^{-1} = \frac{\sigma^2}{n}, \end{aligned}$$

wie behauptet, da $R(\mu,\overline{X}_n)=\sigma^2/n$. □

1.3 Ergänzungen: Quantile

Definition. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ mit Verteilungsfunktion $F(x) = \mathbb{P}((-\infty, x])$. Für $\alpha \in (0,1)$ ist das α-Quantil $q_\alpha \in \mathbb{R}$ von $\mathbb{P}$ definiert durch

$$\mathbb{P}((-\infty, q_\alpha)) \leqslant \alpha \leqslant \mathbb{P}((-\infty, q_\alpha]).$$

Die Quantilfunktion ist definiert als verallgemeinertes Inverses von F:

$$F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geqslant p\}, \quad p \in [0,1].$$

α-Quantile sind nicht eindeutig, falls F auf dem Niveau α irgendwo konstant ist. Es gilt aber

Lemma. *$F^{-1}(\alpha)$ ist ein α-Qunatil.*

Beweis. Aufgrund der Rechtsstetigkeit von F gilt $F(F^{-1}(\alpha)) \geqslant \alpha$. Für alle $x < F^{-1}(\alpha)$ gilt $F(x) < \alpha$ und wegen der linken Grenzwerte von F

$$\alpha \geqslant \lim_{r \uparrow F^{-1}(\alpha)} F(x) = \lim_{r \uparrow F^{-1}(\alpha)} \mathbb{P}((-\infty, r]) = \mathbb{P}((-\infty, r)). \qquad \square$$

Das verallgemeinerte Inverse hat folgende **Eigenschaften**:

(i) $F^{-1}(p) \leqslant x \Leftrightarrow p \leqslant F(x)$;

(ii) $F \circ F^{-1}(p) \geqslant p$ und Gleichheit gilt genau dann, wenn $p \in \operatorname{ran} F$. Die Gleichheit kann nur dann nicht gelten, wenn F unstetig bei $F^{-1}(p)$ ist;

(iii) $F^{-1} \circ F(x) \leqslant x$, wobei Gleichheit genau dann nicht gilt wenn x im Inneren oder am rechten Rand einer "Ebene" (kein Anstieg) von F liegt.

Damit gilt $F \circ F^{-1}(p) = p$ auf $(0,1)$ genau dann, wenn F stetig ist (d.h. $\operatorname{ran} F = [0,1]$) und $F^{-1} \circ F(x) = x$ gilt auf $\mathbb{R}$ genau dann, wenn F strikt monoton wachsend ist. Folglich ist F^{-1} ein echtes Inverses genau dann, wenn F stetig und streng monoton wachsend ist.

Satz. *Ist $U \sim Uni([0,1])$, dann besitzt die Zufallsvariable $F^{-1}(U)$ die Verteilungsfunktion F (Quantilstransformation). Besitzt X die Verteilungsfunktion F, dann gilt $F(X) \sim Uni([0,1])$ genau dann, wenn F stetig ist.*

Beweis. Aus (i) folgt $\mathbb{P}(F^{-1}(U) \leqslant x) = \mathbb{P}(U \leqslant F(x)) = F(x)$ für alle $x \in \mathbb{R}$. Anderseits gilt für $p \in (0,1)$ wegen (i) und (ii)

$$\mathbb{P}(F(X) \leqslant p) = \mathbb{P}(X \leqslant F^{-1}(p)) = F(F^{-1}(p)) = p \quad \Longleftrightarrow \quad p \in \operatorname{ran} F. \qquad \square$$

Schließlich wollen wir noch den **QQ-Plot** (Qunatil-Quantil-Plot) verstehen: Die empirische Verteilungsfunktion einer mathematischen Stichprobe $X_1, \ldots, X_n$ ist gegeben durch $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leqslant x\}}$. Die Verteilungsfunktion der Standardnormalverteilung ist $\Phi(x) = \int_{-\infty}^x (2\pi)^{-1/2} e^{-y^2/2} \mathrm{d}y$. Für große n approximiert F_n die wahre Verteilungsfunktion F, da nach dem starken Gesetz der großen Zahlen $F_n(x) \to \mathbb{E}[\mathbb{1}_{\{X_1 \leqslant x\}}] = F(x)$ $\mathbb{P}$-f.s. für alle $x \in \mathbb{R}$ gilt (tatsächlich gilt diese Konvergenz sogar gleichmäßig auf $\mathbb{R}$ nach dem Satz von Borel-Cantelli). Falls $X_i \sim \mathcal{N}(\mu, \sigma^2)$, so gilt $F(x) = \Phi(\frac{x-m}{\sigma})$. Für die Quantilfunktion gilt also

$$F^{-1}(\Phi(x)) = \Phi^{-1}(\Phi(x)) \cdot \sigma + m = \sigma \cdot x + m,$$

d.h. $F^{-1} \circ \Phi$ ist eine Gerade. Im QQ-Plot wird F_n^{-1} (die empirischen Quantile) gegen Φ^{-1} aufgetragen und unter einer $\mathcal{N}(\mu, \sigma^2)$-Annahme sollten die Werte in etwa auf einer Geraden liegen.

2 Lineares Modell

2.1 Regression und kleinste Quadrate

Regression ist eine Methode um den Zusammenhang zwischen einer *Zielgröße* (*Response-Variable*) Y und einem Vektor von erklärenden Variablen (*Kovariablen, Regressoren*) $X = (x_1, \dots, x_k)$ zu analysieren. Beginnen wir mit dem *einfachen linearen Modell*

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, \dots, n,$$

mit Zufallsvariablen $\varepsilon_1, \dots, \varepsilon_n$, die zentriert sind ($\mathbb{E}_i[\varepsilon_i] = 0$) und endliche Varianz $\mathrm{Var}(\varepsilon_i) = \sigma^2 > 0$ haben. Die Parameter $a, b \in \mathbb{R}, \sigma > 0$ sind unbekannt. Gesucht ist eine *Regressionsgerade* der Form $y = ax + b$, die die Beobachtungen möglichst gut erklärt. Der Parameter σ ist typischerweise nicht das Ziel der statistischen Inferenz und somit ein *Störparameter*.

Beispiel 2.1. Y_i ist das Wachstum von Deutschlands Bruttoinlandsproduktes im Jahr i. Die Kovariable x_i ist die Veränderung der Arbeitslosenquote im Vergleich zum Vorjahr. Unter Verwendung der Daten von 1992 bis 2012 aus den "World Development Indicators" der Weltbank erhalten als Regressionsgrade erhalten wir $y = -1,080 \cdot x + 1,338$. Betrachten wir alle sechs Gründungsmitglieder der EU im gleichen Zeitraum ergibt ganz ähnlich $y = -1,075 \cdot x + 1,819$ Der lineare Zusammenhang beider Größen ist als *Okuns Gesetz* bekannt.

Um die Situation weiter zu vereinfachen nehmen wir zunächst an, dass $\varepsilon_1, \dots, \varepsilon_n$ unabhängig und $\mathcal{N}(0, \sigma^2)$-verteilt sind. Nun können wir den Maximum-Likelihood-Schätzer bestimmen: Der Beobachtungsvektor ist verteilt gemäß der Lebesguedichte

$$\begin{aligned} L(a, b, \sigma; y) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big(- \frac{(y_i - ax_i - b)^2}{2\sigma^2} \Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 \Big), \quad y \in \mathbb{R}^n. \end{aligned}$$

Somit ist die Loglikelihoodfunktion

$$l(a, b, \sigma; y) := \log L(a, b, \sigma; y) = -\frac{n}{2}(\log \sigma^2 + \log(2\pi)) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Das Maximieren der Likelihood über a, b ist also äquivalent zum Minimieren der Summe der quadrierten Residuen (RSS: residual sum of squares). Auch wenn die Fehler nicht normalverteilt sind, kann diese Methode gute Ergebnisse erzielen.

Methode 5: Methode der kleinesten Quadrate. Im einfachen linearen Modell sind die Kleinste-Quadrate-Schätzer $\widehat{a}, \widehat{b}$ durch Minimierung der Summe quadratischen Abstände

$$(\widehat{a}, \widehat{b}) := \arg\min_{a,b} \sum_{i=1}^{n} (Y_i - ax_i - b)^2$$

gegeben.

Satz 2.2. *Im einfachen linearen Modell mit unabhängigen und $\mathcal{N}(0, \sigma^2)$-verteilten Fehlern, ist der Maximum-Likelihood-Schätzer gleich dem Kleinste-Quadrate-Schätzer und es gilt*

$$\widehat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \quad \textit{und} \quad \widehat{b} = \overline{Y}_n - \widehat{a}\overline{x}_n,$$

wobei $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ und $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Beweis. Es bleibt festzustellen, dass wir durch Differentiation folgende Normalgleichungen erhalten:

$$0 = \sum_{i=1}^{n} x_i(Y_i - ax_i - b) \quad \text{und} \quad 0 = \sum_{i=1}^{n} (Y_i - ax_i - b),$$

die leicht gelöst werden können. □

Bemerkung 2.3. Bei der Wahl anderer Fehlerverteilungen ergibt das Maximum-Likelihood-Prinzip andere (nicht weniger sinnvolle) Schätzer (Übung ☑), die aber im Allgemeinen nicht in geschlossener Form darstellbar sind. Populäre nicht gaußsche Fehlerverteilungen sind Laplace- und Exponential-Verteilungen.

Haben wir $k \geqslant 2$ Kovariablen und n Beobachtungen Y_i, führt das zur *multiplen linearen Regression*

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i, \quad i = 1, \ldots, n,$$

wobei die Fehlerterme (ε_i) iid. und zentriert sind mit $0 < \mathrm{Var}(\varepsilon_i) =: \sigma^2 < \infty$. In Vektorschreibweise erhalten wir

$$\begin{aligned}
Y &= (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n && \text{Response-Vektor,} \\
X &:= \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix} \in \mathbb{R}^{n \times (k+1)} && \text{Design-Matrix,} \\
\varepsilon &:= (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n && \text{Vektor der Fehlerterme,} \\
\beta &:= (\beta_0, \ldots, \beta_k)^\top \in \mathbb{R}^{k+1} && \text{Parametervektor,}
\end{aligned}$$

so dass das multiple Regressionsmodell in der Form

$$Y = X\beta + \varepsilon$$

geschrieben werden kann. Der kleinste-Quadrate-Schätzer löst folglich das Minimierungsproblem

$$\min_b |Xb - Y|^2.$$

Beispiel 2.4. Im "crime"-Datensatz von Agresti and Finlay (1997, Kap. 9) stehen für die 51 Staaten der USA die beiden Responsevariablen

- Anzahl der Gewaltverbrechen pro 100.000 Einwohnern (crime),
- Morde pro 1.000.000 Einwohner (murder),

und folgende Kovariablen zur Verfügung:

- Prozentualer Anteil der Bevölkerung die in Ballungs-/ Großstadtgebieten leben (pctmetro),
- Prozentualer Anteil der weißen Bevölkerung (pctwhite),
- Prozentualer Anteil der Bevölkerung mit einem High-School-Abschluss (pcths),
- Prozentualer Anteil der Bevölkerung der unter der Armutsgrenze leben (poverty) und
- Prozentualer Anteil der Bevölkerung mit alleinerziehenden Eltern (single).

Bemerkung 2.5. Wechselwirkungen zwischen zwei Kovariablen x_i und x_j werden durch Interaktionsterme $x_i \cdot x_j$ modelliert. Kategorielle Kovariablen sollten durch eine Menge von sogenannten "Dummy-Indikatoren" kodiert werden, um nicht implizit eine (inadäquate) Metrisierung auf dem diskreten Wertebereich solcher Kovariablen zu induzieren. Eine kategorielle Kovariable mit ℓ möglichen Ausprägungen wird dabei durch $(\ell - 1)$ Indikatoren (d.h. $\{0,1\}$-wertige Variablen) repräsentiert. Der j-te Dummy-Indikator kodiert dabei das Ereignis, dass die Kategorie $(j+1)$ bei der zugehörigen Kovariablen vorliegt, $j = 1, \ldots, \ell - 1$. Sind also alle $(\ell - 1)$ Indikatoren gleich Null, so entspricht dies der (Referenz-) Kategorie 1 der zugehörigen kategoriellen Kovariable (vgl. Varianzanalyse).

Dies führt uns zur allgemeinen Definition des *linearen Modells:*

Definition 2.6. Ein lineares Modell mit n reellwertigen Beobachtungen $Y = (Y_1, \ldots, Y_n)^\top$ und k-dimensionalem Parameter $\beta \in \mathbb{R}^k, k < n$, besteht aus einer reellen Matrix $X \in \mathbb{R}^{n \times k}$ von vollem Rang k, der Designmatrix, und einem Zufallsvektor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, den Fehler- oder Störgrößen, mit $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \Sigma_{i,j}$ für eine Kovarianzmatrix $\Sigma > 0$. Beobachtet wird eine Realisierung von

$$Y = X\beta + \varepsilon.$$

Der (gewichtete) Kleinste-Quadrate-Schätzer $\widehat{\beta}$ von β minimiert den gewichteten Euklidischen Abstand zwischen Beobachtungen und Modellvorhersage:

$$|\Sigma^{-1/2}(X\widehat{\beta} - Y)|^2 = \inf_{b \in \mathbb{R}^k} |\Sigma^{-1/2}(Xb - Y)|^2.$$

Im gewöhnlichen Fall $\Sigma = \sigma^2 E_n$ mit Fehlerniveau $\sigma > 0$ erhalten wir den gewöhnlichen Kleinste-Quadrate-Schätzer (OLS: ordinary least squares)

$$|X\widehat{\beta} - Y|^2 = \inf_{b \in \mathbb{R}^k} |Xb - Y|^2,$$

der unabhängig von der Kenntniss von σ^2 ist.

Bemerkung 2.7. Wir schreiben $\Sigma > 0$, falls Σ eine symmetrische, strikt positiv-definite Matrix ist. Dann ist Σ diagonalisierbar mit $\Sigma = TDT^\top, D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ Diagonalmatrix und T orthogonale Matrix, und wir setzen $\Sigma^{-1/2} := TD^{-1/2}T^\top$ mit $D^{1/2} := \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2})$. Wie erwartet, gilt $(\Sigma^{-1/2})^2 = \Sigma^{-1}$ und somit $|\Sigma^{-1/2}v|^2 = \langle \Sigma^{-1}v, v \rangle$.

Zusätzlich zur einfachen und multiplen Regression umfasst das lineare Modell weitere Beispiele.

Beispiel 2.8 (Polynomiale Regression)**.** Wir beobachten

$$Y_i = a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_{k-1} x_i^{k-1} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Damit ergibt sich als Parameter $\beta = (a_0, \ldots, a_{k-1})^\top$ und eine Designmatrix vom Vandermonde-Typ

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{k-1} \end{pmatrix}.$$

Die Matrix hat vollen Rang, sofern k der Designpunkte (x_i) verschieden sind.

Lemma 2.9. *Setze $X_\Sigma := \Sigma^{-1/2}X$. Mit Π_{X_Σ} werde die Orthogonalprojektion von $\mathbb{R}^n$ auf den Bildraum* $\mathrm{ran}(X_\Sigma)$ *bezeichnet. Dann gilt*

$$\Pi_\Sigma = X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$$

und für den Kleinste-Quadrate-Schätzer

$$\widehat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y.$$

Insbesondere existiert der Kleinste-Quadrate-Schätzer, ist eindeutig und erwartungstreu.

Beweis. Zunächst beachte, dass $X_\Sigma^\top X_\Sigma = X^\top \Sigma^{-1} X$ invertierbar ist wegen der Invertierbarkeit von Σ und der Rangbedingung an X:

$$X^\top \Sigma^{-1} X v = 0 \Rightarrow v^\top X^\top \Sigma^{-1} X v = 0 \Rightarrow |\Sigma^{-1/2} X v| = 0 \Rightarrow |Xv| = 0 \Rightarrow v = 0.$$

Setze $P_{X_\Sigma} := X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$ und $w = P_{X_\Sigma} v$ für ein $v \in \mathbb{R}^n$. Dann folgt $w \in \operatorname{ran}(X_\Sigma)$ und im Fall $v = X_\Sigma u$ durch Einsetzen $w = P_{X_\Sigma} X_\Sigma u = v$, so dass P_{X_Σ} eine Projektion auf $\operatorname{ran}(X_\Sigma)$ ist. Da P_{X_Σ} selbstadjungiert (symmertrisch) ist, handelt es sich um die Orthogonalprojektion Π_{X_Σ}:

$$\forall u \in \mathbb{R}^n, \forall w \in \operatorname{ran} X_\Sigma : \langle u - P_{X_\Sigma} u, w \rangle = \langle u, w \rangle - \langle u, P_{X_\Sigma} w \rangle = 0.$$

Aus der Eigenschaft $\widehat{\beta} = \arg\min_b |\Sigma^{-1/2}(Y - Xb)|^2$ folgt, dass $\widehat{\beta}$ die beste Approximation von $\Sigma^{-1/2} Y$ durch $X_\Sigma b$ liefert. Diese ist durch die Orthogonalprojektionseigenschaft $\Pi_{X_\Sigma} \Sigma^{-1/2} Y = X_\Sigma \widehat{\beta}$ bestimmt. Es folgt

$$X_\Sigma^\top \Pi_{X_\Sigma} \Sigma^{-1/2} Y = (X_\Sigma^\top X_\Sigma)\widehat{\beta} \Rightarrow (X_\Sigma^\top X_\Sigma)^{-1} X^\top \Sigma^{-1} Y = \widehat{\beta}.$$

Schließlich folgt aus der Linearität des Erwartungswertes und $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\widehat{\beta}] = \mathbb{E}[(X_\Sigma^\top X_\Sigma)^{-1} X^\top \Sigma^{-1} (X\beta + \varepsilon)] = \beta + 0 = \beta. \qquad \square$$

Bemerkung 2.10.

- Im gewöhnlichen linearen Modell bzw. der multiplen linearen Regression gilt $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$ und ist somit unabhängig vom unbekannten Parameter $\sigma > 0$.
- $X_\Sigma^\dagger := (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$ heißt auch Moore-Penrose-(Pseudo-)Inverse von X_Σ, so dass $\widehat{\beta} = X_\Sigma^\dagger \Sigma^{-1/2} Y$ bzw. $\widehat{\beta} = X^\dagger Y$ im gewöhnlichen linearen Modell gilt.

Wir kommen zum zentralen Satz in der Regressionsanalyse:

Satz 2.11 (Gauß-Markov)**.** *Ist der Parameter $\rho = \langle \beta, v \rangle$ für ein $v \in \mathbb{R}^k$ im linearen Modell zu schätzen, so ist $\widehat{\rho} = \langle \widehat{\beta}, v \rangle$ ein (in den Daten Y) linearer erwartungstreuer Schätzer, der unter allen linearen erwartungstreuen Schätzern minimale Varianz besitzt, nämlich* $\operatorname{Var}(\widehat{\rho}) = |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v|^2$.

Beweis. Die Linearität ist klar und aus dem vorangegangen Lemma folgt, dass $\widehat{\rho}$ erwartungstreu ist. Sei nun $\widetilde{\rho} = \langle Y, w \rangle$ ein beliebiger linearer erwartungstreuer Schätzer von ρ. Dies impliziert für alle $\beta \in \mathbb{R}^k$

$$\mathbb{E}[\langle Y, w \rangle] = \rho \Rightarrow \langle X\beta, w \rangle = \langle \beta, v \rangle \Rightarrow \langle X^\top w - v, \beta \rangle = 0$$

und somit $v = X^\top w = X_\Sigma^\top \Sigma^{1/2} w$. Nach Pythagoras erhalten wir

$$\begin{aligned} \operatorname{Var}(\widetilde{\rho}) &= \mathbb{E}[\langle \varepsilon, w \rangle^2] = \mathbb{E}[w^\top \varepsilon \varepsilon^\top w] \\ &= w^\top \Sigma w = |\Sigma^{1/2} w|^2 = |\Pi_{X_\Sigma} (\Sigma^{1/2} w)|^2 + |(E_n - \Pi_\Sigma)(\Sigma^{1/2} w)|^2. \end{aligned}$$

Damit gilt $\operatorname{Var}(\widetilde{\rho}) \geqslant |\Pi_{X_\Sigma} (\Sigma^{1/2} w)|^2 = |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X^\top w| = |X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} v| = \operatorname{Var}(\widehat{\rho})$. $\square$

Bemerkung 2.12. Man sagt, dass der Schätzer $\widehat{\rho}$ im Satz von Gauß-Markov bester linearer erwartungstreuer Schätzer (blue: best linear unbiased estimator) ist. Eingeschränkt auf lineare Schätzer ist der Kleinste-Quadrate-Schätzer damit minimax. Ob es einen besseren nichtlinearen Schätzer geben kann, werden wir in Kapitel 3 beantworten.

Im gewöhnlichen linearen Modell ist die optimale Varianz insbesondere $\sigma^2 |X(X^\top X)^{-1} v|^2$. In diesem Spezialfall ist es auch von Interesse das Rauschniveau σ^2 zu schätzen. Dies ermöglicht es insbesondere Tests und Konfidenzbereiche zu konstruieren.

Lemma 2.13. *Im gewöhnlichen linearen Modell mit $\sigma > 0$ und Kleinste-Quadrate-Schätzer $\widehat{\beta}$ gilt $X\widehat{\beta} = \Pi_X Y$ und $R := Y - X\widehat{\beta}$ bezeichne den Vektor der Residuen. Die geeignet normalisierte Stichprobenvarianz*

$$\widehat{\sigma}^2 := \frac{|R|^2}{n-k} = \frac{|Y - X\widehat{\beta}|^2}{n-k}$$

ist erwartungstreuer Schätzer von σ^2.

Beweis. $X\widehat{\beta} = \Pi_X Y$ folgt aus Lemma 2.9. Einsetzen zeigt $\mathbb{E}[|Y - X\widehat{\beta}|^2] = \mathbb{E}[|Y - \Pi_X Y|^2] = \mathbb{E}[|(E_n - \Pi_X)\varepsilon|^2]$. Ist nun $e_1, \dots, e_{n-k}$ eine Orthnormalbasis vom $(n-k)$-dimensionalen Bild $\mathrm{ran}(E_n - \Pi_X) \subseteq \mathbb{R}^n$, so folgt

$$\mathbb{E}[|(E_n - \Pi_X)\varepsilon|^2] = \sum_{i=1}^{n-k} \mathbb{E}[\langle \varepsilon, e_i \rangle^2] = \sigma^2(n-k),$$

was die Behauptung impliziert. □

Beachte, dass der Maximum-Likelihood-Schätzer von σ^2 gegeben ist durch $\widehat{\sigma}^2_{ML} = n^{-1}|R|^2 \neq \widehat{\sigma}^2$ (Übung ☑). Der erwartungstreue Schätzer $\widehat{\sigma}^2$ wird in der Praxis bevorzugt, hat jedoch größere Varianz als $\widehat{\sigma}^2_{ML}$.

Bevor wir uns mit statistischer Inferenz, also der Konstruktion von Tests und Konfidenzintervallen, im linearen Modell beschäftigen, soll der Baysianische Ansatz auf das Regressionsproblem angewendet werden.

Satz 2.14. *Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim \mathcal{N}(0, \sigma^2 E_n)$ und bekanntem $\sigma > 0$ genüge $\beta \in \mathbb{R}^k$ der a-priori-Verteilung*

$$\beta \sim \mathcal{N}(m, \sigma^2 M)$$

mit Parametern $m \in \mathbb{R}^k$ und symmetrisch positiv definiter Matrix $M \in \mathbb{R}^{k \times k}$. Dann ist die a-posteriori-Verteilung von β gegeben einer Realisierung $y \in \mathbb{R}^n$ gegeben durch

$$\beta | Y = y \sim \mathcal{N}(\mu_y, \Sigma_y) \quad \textit{mit} \quad \Sigma_y = \sigma^2 \big(X^\top X + M^{-1}\big)^{-1}, \mu_y = \Sigma_y(\sigma^{-2} X^\top y + \sigma^{-2} M^{-1} m).$$

Insbesondere ist der Bayesschätzer bzgl. quadratischem Verlust gegeben durch $\widehat{\beta}_{Bayes} = \big(X^\top X + M^{-1}\big)^{-1}(X^\top Y + M^{-1}m)$.

Beweis. Für die a-posteriori-Dichte an der Stelle $t \in \mathbb{R}^k$ gilt

$$\begin{aligned} f_{\beta|Y=y}(t) &\sim \exp\Big(-\frac{1}{2\sigma^2}(y - Xt)^\top (y - Xt)\Big) \exp\Big(-\frac{1}{2\sigma^2}(t-m)^\top M^{-1}(t-m)\Big) \\ &\sim \exp\Big(\frac{1}{\sigma^2} t^\top X^\top y - \frac{1}{2\sigma^2} t^\top X^\top X t - \frac{1}{2\sigma^2} t^\top M^{-1} t + \frac{1}{\sigma^2} t^\top M^{-1} m\Big) \\ &= \exp\Big(\frac{1}{\sigma^2} t^\top \big(X^\top y + M^{-1} m\big) - \frac{1}{2\sigma^2} t^\top \big(X^\top X + M^{-1}\big) t\Big). \end{aligned}$$

Daher ist β gegeben $Y = y$ normalverteilt mit Kovarianzmatrix $\Sigma_y = (\sigma^{-2} X^\top X + \sigma^{-2} M^{-1})^{-1}$ und Mittelwert $\mu_y = \Sigma_y(X^\top y + M^{-1} m)/\sigma^2$. □

Es ist erneut bemerkenswert, dass der Bayesschätzer $\widehat{\beta}_{Bayes}$ nicht von σ^2 abhängt.

Bemerkung 2.15. Indem wir auch den Parameter σ^2 mit einer a-priori-Verteilung versehen, erhalten wir ein (mehrstufiges) Bayesmodell. Da wir besonders an konjugierten Verteilungsklassen interessiert sind, wird hierzu oft die inverse Gamma-Verteilung verwendet: Ist $Z \sim \Gamma(a, b)$ so ist $1/Z \sim IG(a, b)$ invers Gamma-verteilt mit Parametern $a, b > 0$ und Lebesguedichte

$$f_{a,b}(x) = \frac{b^a}{\Gamma(a)} x^{-(a-1)} e^{-a/x} \mathbb{1}_{(0,\infty)}(x), \quad x \in \mathbb{R}.$$

Das Bayesmodell ist also gegeben durch

$$Y|\beta,\sigma^2 \sim \mathcal{N}(X\beta,\sigma^2 E_n), \quad \beta|\sigma^2 \sim \mathcal{N}(m,\sigma^2 M), \quad \sigma \sim IG(a,b).$$

Die gemeinsame Verteilung von $(\beta,\sigma^2) \sim NIG(m,M,a,b)$ wird Normal-inverse Gammaverteilung genannt und besitzt die Dichte

$$\begin{aligned} f(\beta,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{k/2}|M|^{1/2}} \exp\Big(-\frac{1}{2\sigma^2}(\beta-m)^\top M^{-1}(\beta-m)\Big)\frac{b^a}{\Gamma(a)(\sigma^2)^{a+1}}e^{-a/\sigma^2} \\ &\sim \frac{1}{(\sigma^2)^{k/2+a+1}} \exp\Big(\frac{1}{2\sigma^2}\big((\beta-m)^\top M^{-1}(\beta-m)+b\big)\Big), \qquad \beta\in\mathbb{R}^k, \sigma^2>0. \end{aligned}$$

In diesem Modell ist die a-posteriori-Verteilung von σ^2 gegeben β und Y gegeben durch $\sigma^2|\beta,Y \sim IG(a',b')$ mit $a' = a+\frac{n}{2}+\frac{k}{2}$ und

$$b' = b + \frac{1}{2}(Y-X\beta)^\top(Y-X\beta) + \frac{1}{2}(\beta-m)^\top M^{-1}(\beta-m).$$

Die a-posteriori-Verteilung von (β,σ^2) gegeben Y ist $(\beta,\sigma^2)|Y \sim NIG(\widetilde{m},\widetilde{M},\widetilde{a},\widetilde{b})$ mit Parametern

$$\begin{aligned} \widetilde{M} &= (X^\top X + M^{-1})^{-1}, \quad \widetilde{m} = \widetilde{M}(M^{-1}m + X^\top y), \\ \widetilde{a} &= a+\frac{n}{2}, \quad \widetilde{b} = b + \frac{1}{2}\Big(Y^\top Y + m^\top M^{-1} m - \widetilde{m}^\top \widetilde{M}^{-1}\widetilde{m}\Big), \end{aligned}$$

siehe Fahrmeir et al. (2009, Kap. 3.5).

Korollar 2.16. *Unter den Voraussetzungen des vorangegangenen Satzes mit $m=0$ und $M=\tau^2 E_k, \tau>0$, gilt für den Bayesschätzer unter quadratischem Verlust*

$$\widehat{\beta}_{Bayes} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^k} |Y-X\beta|^2 + \frac{1}{\tau^2}|\beta|^2.$$

Beweis. Im Spezialfall $m=0$ und $M=\tau^2E_k$ folgt aus obigem Satz $\widehat{\beta}_{Bayes} = \big(X^\top X + \tau^{-2}E_k\big)^{-1}X^\top y$. Andererseits gilt

$$\begin{aligned} &\operatorname*{arg\,min}_\beta \Big((Y^\top - \beta^\top X^\top)(Y-X\beta) + \frac{1}{\tau^2}\beta^\top\beta\Big) \\ &\qquad = \operatorname*{arg\,min}_\beta \Big(-2Y^\top X\beta + \beta^\top(X^\top X + \frac{1}{\tau^2}E_k)\beta\Big). \end{aligned}$$

Null setzen des Differenzials der Funktion $\beta \mapsto -2Y^\top X\beta + \beta^\top(X^\top X + \frac{1}{\tau^2}E_k)\beta$ liefert $0 = -2Y^\top X + 2\beta^\top(X^\top X + \frac{1}{\tau^2}E_k)$, so dass aus der positiv Definitheit und Symmetrie von $X^\top X + \frac{1}{\tau^2}E_k$ die Behauptung folgt. □

Der Bayesansatz führt uns also zu einer neuen Schätzmethode im linearen Modell:

Methode 6: Ridge-Regression. Im linearen Modell $Y = X\beta + \varepsilon$ ist der Ridge-Regressionsschätzer oder Schrumpfungsschätzer (engl.: Shrinkage) mit Schrumpfungskoeffizient $\lambda \geqslant 0$ definiert als

$$\widehat{\beta}_{ridge} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^k} |Y-X\beta|^2 + \lambda|\beta|^2.$$

Durch Einführung des Strafterms (engl.: *penalty*) $\lambda|\beta|^2$ wird die Varianz auf Kosten eines Bias verringert. Dies ist insbesondere sinnvoll, wenn einige (wenige) Koeffizienten von β groß sind und die übrigen klein und liefert in diesen Fällen gute Schätzergebnisse auch wenn die Parameterdimension in einer ähnlichen Größenordnung liegt wie die Anzahl der Beobachtungen ($n \sim p$). Dies wird im nächsten Beispiel illustriert. Die richtige Wahl des Strumpfungsparameters λ ist allerdings ein schwieriges Problem.

Beispiel 2.17. Betrachten wir das Modell $Y_i = x_i^\top \beta + \varepsilon_i$ mit Kovariablenvektor $x_i \in \mathbb{R}^p$, Parameter $\beta \in \mathbb{R}^p$ und $\varepsilon_i \overset{iid.}{\sim} \mathcal{N}(0,1)$ mit $i = 1, \dots, n$. Wir wählen $n = 50$ und $p = 30$ wobei 10 Koeffizienten groß sind (zwischen 0,5 und 1) und 20 klein (zwischen 0 und 0,3) und bestimmen den mittleren Quadratischen Fehler aus 200 Simulationen für verschiedene Werte von $\lambda \in [0, 20]$ (Übung ☑).

2.2 Inferenz unter Normalverteilungsannahme

Im Folgenden werden wir das gewöhnliche lineare Modell unter der Normalverteilungsannahme $(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2 E_n)$ betrachten.

Beispiel 2.18. Sind die Messfehler $(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2 E_n)$ gemeinsam normalverteilt und $\rho = \langle v, \beta \rangle$ für $v \in \mathbb{R}^k$, so gilt

$$\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1}) \quad \text{und} \quad \widehat{\rho} = \langle v, \widehat{\beta} \rangle \sim \mathcal{N}(\gamma, \sigma^2 v^\top (X^\top X)^{-1} v).$$

Ist $\sigma > 0$ bekannt, so ist ein Konfidenzintervall zum Niveau 95% für ρ gegeben durch

$$I_{0,95}(\rho) := \big[\widehat{\rho} - 1,96\sigma\sqrt{v^\top (X^\top X)^\top v}, \widehat{\rho} + 1,96\sigma\sqrt{v^\top (X^\top X)^\top v}\big].$$

Dabei ist der Wert $1,96$ gerade das 0,975-Qunatil bzw. 0,025 Fraktil der Standardnormalverteilung. Analog (Korrespondenzsatz) wird der zweiseitige Gauß-Test der Hypothese $H_0 : \rho = \rho_0$ gegen $H_1 : \rho \neq \rho_0$ zum Niveau $\alpha \in (0,1)$ konstruiert: Wähle die Teststatistik $|\widehat{\rho} - \rho_0|$ und den kritischen Wert $q_{1-\alpha/2}\sigma\sqrt{v^\top (X^\top X)^\top v}$ mit dem $(1-\alpha/2)$-Quantil von $\mathcal{N}(0,1)$.

Ist σ unbekannt, so ist eine Idee, einfach σ durch den Schätzer $\widehat{\sigma}$ in obiger Formel zu ersetzen. Allerdings wird dann das vorgegebene Niveau nur noch asymptotisch erreicht für einen konsistenten Schätzer (Slutsky-Lemma). Im vorliegenden Fall können wir aber sogar die Verteilung für endliche Stichprobenumfänge exakt bestimmen.

Definition 2.19. Die t-Verteilung $t(n)$ (oder Student-t-Verteilung) mit $n \in \mathbb{N}$ Freiheitsgraden auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ ist gegeben durch die Lebesguedichte

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}} \Big(1 + \frac{x^2}{n}\Big)^{-(n+1)/2}, \quad x \in \mathbb{R}.$$

Die F-Verteilung $F(m,n)$ (oder Fisher-Verteilung) mit $(m,n) \in \mathbb{N}^2$ Freiheitsgraden auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ ist gegeben durch die Lebesguedichte

$$f_{m,n}(x) = \frac{m^{m/2} n^{n/2}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \mathbb{1}_{\mathbb{R}^+}(x), \quad x \in \mathbb{R}.$$

Dabei bezeichnet $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} \mathrm{d}t$ die Gamma-Funktion und $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ die Beta-Funktion.

Erinnerung: Für $X_1, \dots, X_m \sim \mathcal{N}(0,1)$ ist $X := \sum_{i=1}^m X_i^2 \sim \chi^2(n)$ verteilt mit Lebesguedichte $f_X(x) = (2^{m/2}\Gamma(\frac{m}{2}))^{-1} x^{m/2-1} e^{-x/2} \mathbb{1}_{\mathbb{R}^+}(x)$.

Lemma 2.20. *Es seien $X_1, \dots, X_m, Y_1, \dots, Y_n$ unabhängige $\mathcal{N}(0,1)$-verteilte Zufallsvariablen. Dann gilt*

$$T_n := \frac{X_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}} \sim t(n) \quad \textit{und} \quad F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2} \sim F(m,n).$$

Beweis. Es gilt $T_n^2 = F_{1,n}$, so dass mittels Dichtetransformation $f_{|T_n|}(x) = f_{F_{1,n}}(x^2) 2x, x \geqslant 0$, gilt. Da T_n symmetrisch (wie $-T_n$) verteilt ist, folgt $f_{T_n} = F_{F_{1,n}}(x^2)|x|, x \in \mathbb{R}$, und Einsetzen zeigt die Behauptung für T_n, sofern $F_{1,n}$ $F(1,n)$-verteilt ist.

Um die Behauptung für $F_{m,n}$ nachzuweisen, benutze, dass $X := \sum_{i=1}^m X_i^2$ $\chi^2(m)$-verteilt und $Y := \sum_{j=1}^n Y_j^2$ $\chi^2(n)$-verteilt sind. Wegen Unabhängigkeit von X und Y gilt für $z > 0$ (setze $w = x/y$)

$$\begin{aligned}\mathbb{P}(X/Y \leqslant z) &= \int\int \mathbb{1}_{\{x/y \leqslant z\}} f_x(x) f_Y(y) \mathrm{d}x\mathrm{d}y \\ &= \int \mathbb{1}_{\{w \leqslant z\}} \Big(\int f_X(wy) f_Y(y) y \mathrm{d}y \Big) \mathrm{d}w,\end{aligned}$$

so dass sich die Dichte wie folgt ergibt (setze $w = (z+1)y$)

$$\begin{aligned}f_{X/Y}(z) =& \int f_X(zy) f_Y(y) y \mathrm{d}y \\ =& \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zy)^{m/2-1} y^{n/2} e^{-(zy+y)/2} \mathrm{d}y \\ =& \frac{2^{-(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty (zw/(z+1))^{m/2-1} (w/(z+1))^{n/2} e^{-w/2} (z+1)^{-1} \mathrm{d}w \\ =& \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} z^{m/2-1} (z+1)^{-(m+n)/2}, \quad z > 0.\end{aligned}$$

Dichtetransformation ergibt damit für $F_{m,n} = \frac{n}{m}\frac{X}{Y}$ die Dichte $\frac{m}{n} f_{X/Y}(\frac{m}{n}x) = f_{m,n}(x)$. □

Bemerkung 2.21. Es gilt $T_n^2 = F_{1,n}$. Für $n = 1$ ist die $t(n)$-Verteilung gerade die Cauchy-Verteilung und für $n \to \infty$ konvergiert sie schwach gegen die Standardnormalverteilung. Für jedes $n \in \mathbb{N}$ besitzt $t(n)$ nur Momente bis zur Ordnung $p < n$ (sie ist *heavy-tailed*). Ähnliches gilt für die F-Verteilung, insbesondere konvergiert die Verteilung von $mF_{m,n}$ für $n \to \infty$ gegen die $\chi^2(m)$-Verteilung.

Aus diesem Lemma ergeben sich die Standardtests für die Parameter der Normalverteilung, siehe Witting (1985, S. 200-204).

Bevor wir zur Konstruktion von Tests und Konfidenzbändern im linearen Modell kommen noch ein weiteres nützliches Hilfsresultat zur Verteilung quadratischer Formen:

Lemma 2.22. *Seien $X \sim \mathcal{N}(0, E_n)$ und R eine symmetrische, idempotente $(n \times n)$-Matrix (d.h. $R = R^\top$ und $R^2 = R$) mit* $\operatorname{rank}(R) = r \leqslant n$. *Dann gilt*

(i) $X^\top R X \sim \chi^2(r)$,

(ii) $X^\top R X$ *ist unabhängig von* BX *für jede Matrix* $B \in \mathbb{R}^{p \times n}$ *mit* $p \leqslant n$ *und* $BR = 0$,

(iii) für jede weitere symmetrische, idempotente Matrix $S \in \mathbb{R}^{n \times n}$ *mit* $\operatorname{rank}(S) = s \leqslant n$ *und* $RS = 0$ *sind* $X^\top R X$ *und* $X^\top S X$ *unabhängig und*

$$\frac{s}{r} \frac{X^\top R X}{X^\top S X} \sim F(r, s).$$

Beweis. (i) Da R symmetrisch und idempotent ist, existiert eine Orthogonalmatrix P mit $R = P D_r P^\top$, wobei $D_r = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}$. Da P orthogonal ist und X standardnormalverteilt, folgt $W := P^\top X \sim \mathcal{N}(0, E_n)$. Wegen

$$X^\top R X = X^\top R^2 X = (RX)^\top (RX) = (P D_r W)^\top (P D_r W) = W^\top D_r W = \sum_{i=1}^r W_i^2$$

ist $X^\top R X$ $\chi^2(r)$-verteilt.

(ii) Wir setzen $Y := BX \sim \mathcal{N}(0, B^\top B)$ und $Z := RX \sim \mathcal{N}(0, R)$. Dann gilt

$$\operatorname{Cov}(Y, Z) = B \operatorname{Var}(X) R^\top = BR = 0.$$

Da (X, Y) als Lineartransformation von X gemeinsam normalverteilt ist, folgt aus der Unkorreliertheit bereits die Unabhängigkeit.

(iii) Genau wie in (ii) folgt die Unabhängigkeit von $Y := SX$ und $Z := RX$ und somit auch die Unabhängigkeit von $Y^\top Y = X^\top S X$ und $Z^\top Z = X^\top R X$. Zusammen mit (i) und dem vorangegangenen Lemma folgt die Behauptung. □

Als Korollar erhalten wir Konfidenzbereiche für die Schätzung von β und linearen Funktionalen im gewöhnlichen linearen Modell unter der Normalverteilungsannahme.

Satz 2.23. *Im gewöhnlichen linearen Modell unter der Normalverteilungsannahme* $(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2 E_n)$ *für* $\sigma > 0$ *gelten folgende Konfidenzaussagen für gegebenes Niveau* $\alpha \in (0, 1)$ *:*

(i) Ist $q_{F(k,n-k);1-\alpha}$ *das* $(1-\alpha)$*-Quantil der* $F(k, n-k)$*-Verteilung, so ist*

$$C := \left\{ \beta \in \mathbb{R}^k \middle| |X(\beta - \widehat{\beta})|^2 < k \widehat{\sigma}^2 q_{F(k,n-k);1-\alpha} \right\}$$

ein Konfidenzellipsoid zum Konfidenzniveau $1-\alpha$ *für* β*.*

(ii) Ist $q_{t(n-k);1-\alpha/2}$ *das* $(1-\frac{\alpha}{2})$*-Quantil der* $t(n-k)$*-Verteilung, so ist*

$$I := \left[\widehat{\rho} - \widehat{\sigma} \sqrt{v^\top (X^\top X)^{-1} v} q_{t(n-k);1-\alpha/2}, \widehat{\rho} + \widehat{\sigma} \sqrt{v^\top (X^\top X)^{-1} v} q_{t(n-k);1-\alpha/2} \right]$$

ein Konfidenzintervall zum Konfidenzniveau $1-\alpha$ *für* $\rho = \langle v, \beta \rangle$*.*

Beweis. (i) Nach Konstruktion gilt

$$X\widehat{\beta} = XX^\dagger Y = \Pi_X Y = X\beta + \Pi_X \varepsilon, \quad \widehat{\sigma}^2 = \frac{|(E_n - \Pi_X)\varepsilon|^2}{(n-k)}.$$

Da Π_X und $(E_n - \Pi_X)$ symmetrische, idempotente Matrizen mit Rang k bzw. $(n-k)$ sind (Projektionen auf $\operatorname{ran} X$ bzw. $(\operatorname{ran} X)^\perp$) und es gilt $(E_n + \Pi_X)\Pi_X = 0$, folgt aus Lemma 2.22:

$$\frac{|X(\beta - \widehat{\beta})|^2}{k\widehat{\sigma}^2} = \frac{(n-k)}{k} \frac{\varepsilon^\top (E_n - \Pi_X)\varepsilon}{\varepsilon^\top \Pi_X \varepsilon} \sim F(k, n-k).$$

Durch die Wahl des Quaniles folgt die Konfidenzaussage $\mathbb{P}_\beta(\beta \in C) = 1 - \alpha$.

(ii) Wegen $\widehat{\rho} \sim \mathcal{N}(\rho, \sigma^2 v^\top (X^\top X)^{-1} v)$ nach dem Satz von Gauß-Markov, ist

$$\frac{\rho - \widehat{\rho}}{\sigma \sqrt{v^\top (X^\top X)^{-1} v}} \sim \mathcal{N}(0, 1).$$

Andererseits sind $\widehat{\rho}$ und $\widehat{\sigma}^2$ unabhängig und es gilt $\widehat{\sigma}^2 = \sigma^2 Z/(n-k)$ für eine Zufallsvariable $Z \sim \chi^2(n-k)$. Damit ist

$$\frac{\rho - \widehat{\rho}}{\sqrt{\widehat{\sigma}^2 v^\top (X^\top X)^{-1} v}} \sim t(n-k). \qquad \square$$

Bemerkung 2.24. Ebenso kann man ein Konfidenzintervall für die Varianz konstruieren (Übung ☑).

Zusammen mit dem Korrespondenzsatz liefert dieses Resultat:

Methode 7: t-Test und F-Test. Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $(\varepsilon_i) \sim \mathcal{N}(0, \sigma^2 E_n)$ ist der (zweiseitige) t-Test der Hypothese $H_0 : \rho = \rho_0$ gegen die Alternative $H_1 : \rho \neq \rho_0$ für $\rho_0 = \langle v, \beta_0 \rangle$ zum Niveau $\alpha \in (0,1)$ gegeben durch

$$\varphi_{\rho_0}(Y) = \mathbb{1}_{\{|T_{n-k}(Y)| > q_{t(n-k);1-\alpha/2}\}} \quad \text{mit} \quad T_{n-k}(Y) := \frac{\rho_0 - \widehat{\rho}}{\widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}}.$$

Der F-Test der Hypothese $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$ zum Niveau $\alpha \in (0,1)$ ist gegeben durch

$$\varphi_{\beta_0}(Y) = \mathbb{1}_{\{F_{k,n-k}(Y) > q_{F(k,n-k);1-\alpha}\}} \quad \text{mit} \quad F_{k,n-k}(Y) := \frac{|X(\beta_0 - \widehat{\beta})|^2}{k\widehat{\sigma}^2}.$$

Schließlich wollen wir Hypothesentests noch für den allgemeineren Fall von linearen (bzw. affinen) Hypothesen konstruieren.

Definition 2.25. Im gewöhnlichen linearen Modell ist ein (zweiseitiges) lineares Testproblem gegeben durch

$$H_0 : K\beta = d \quad \text{versus} \quad H_1 : K\beta \neq d$$

für eine (deterministische) Matrix $K \in \mathbb{R}^{r \times k}$ mit vollem Rang $\operatorname{rank}(K) = r \leqslant k$ und einem Vektor $d \in \mathbb{R}^r$. K wird Kontrastmatrix genannt. Unter der Hypothese H_0 sind also insgesamt $r \leqslant k$ linear unabhängige Bedingungen an die Parameter des linearen Modells gestellt.

Beispiel 2.26. Test auf Gleichheit zweier Regressionskoeffizienten: Für $2 \leqslant j < l \leqslant k$ ist das Testproblem gegeben durch

$$H_0 : \beta_j = \beta_l \quad \text{versus} \quad H_1 : \beta_j \neq \beta_l.$$

Damit ist die Kontrastmatrix $K = (a_{1,i}) \in \mathbb{R}^{1 \times k}$ gegeben durch $a_{1,i} = \mathbb{1}_{\{i=j\}} - \mathbb{1}_{\{i=l\}}$ und $d = 0$.

Weitere Beispiele sind der Globaltest (Übung ☑):

$$H_0 : \forall j \in \{1, \dots, k\} : \beta_j = 0 \quad \text{versus} \quad H_1 : \exists j \in \{1, \dots, k\} : \beta_j \neq 0$$

sowie der Test eines Subvektors $\beta^* = (\beta_1^*, \dots, \beta_r^*)^\top$ mit $r \leqslant k$ (Übung ☑):

$$H_0 : \forall j \in \{1, \dots, r\} : \beta_j = \beta_j^* \quad \text{versus} \quad H_1 : \exists j \in \{1, \dots, r\} : \beta_j \neq \beta_j^*.$$

Die Grundidee für das Testen linearer Hypothesen ist, die Residuen $RSS = |Y - X\widehat{\beta}|^2$ des Kleinste-Quadrate-Schätzers mit den Residuen des auf $H_0 : K\beta = d$ eingeschränkten Kleinste-Quardate-Schätzers $\widehat{\beta}_{H_0}$, d.h.

$$RSS_{H_0} := |Y - X\widehat{\beta}_{H_0}|^2 \quad \text{mit} \quad |Y - X\widehat{\beta}_{H_0}|^2 = \min_{\beta \in \mathbb{R}^k : K\beta = d} |Y - X\beta|^2,$$

zu vergleichen. Ist die Abweichung (relativ zu RSS) zu groß, spricht dies gegen die Hypothese.

Satz 2.27. *Im gewöhnlichen linearen Modell unter Normalverteilungsannahme $(\varepsilon_j) \sim \mathcal{N}(0, \sigma^2 E_n)$ ist die lineare Hypothese*

$$H_0 : K\beta = d \quad \textit{versus} \quad H_1 : K\beta \neq d$$

mit Kontrastmatrix $K \in \mathbb{R}^{r \times k}$ und $d \in \mathbb{R}^r$ zu testen. Es gilt

(i) $\widehat{\beta}_{H_0} = \widehat{\beta} - (X^\top X)^{-1} K^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\widehat{\beta} - d)$,

(ii) $RSS_{H_0} - RSS = (K\widehat{\beta} - d)^\top (K(X^\top X)^{-1} K^\top)^{-1} (K\widehat{\beta} - d)$ *und* $(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2(r)$ *unter* H_0

(iii) die Fisher-Statistik $F := \frac{n-k}{r} \frac{RSS_{H_0} - RSS}{RSS}$ *ist unter* H_0 *gemäß* $F(r, n-k)$ *verteilt.*

Beweis. (i) Für jeden Vektor $\gamma \in \mathbb{R}^k$, der die Nebenbedingung $K\gamma$=d erfüllt, gilt

$$|Y - X\gamma|^2 = |Y - X\widehat{\beta} + X(\widehat{\beta} - \gamma)|^2 = |Y - X\widehat{\beta}|^2 + |X(\widehat{\beta} - \gamma)|^2$$

nach Pythagoras, da $Y - X\widehat{\beta} = (E_n - \Pi_X)Y \perp \operatorname{ran}(X)$. Außerdem ist

$$|X(\widehat{\beta} - \gamma)|^2 = |X(\widehat{\beta} - \widehat{\beta}_{H_0})|^2 + |X(\widehat{\beta}_{H_0} - \gamma)|^2 + 2\langle X(\widehat{\beta} - \widehat{\beta}_{H_0}), X(\widehat{\beta}_{H_0} - \gamma)\rangle.$$

Die Wahl von $\widehat{\beta}_{H_0}$ impliziert jedoch

$$\begin{aligned}\langle X(\widehat{\beta} - \widehat{\beta}_{H_0}), X(\widehat{\beta}_{H_0} - \gamma)\rangle =& \big((X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - d)\big)^\top X^\top X(\widehat{\beta}_{H_0} - \gamma)\\ =&(K\widehat{\beta} - d)^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta}_{H_0} - K\gamma) = 0,\end{aligned}$$

denn $\widehat{\beta}_{H_0}$ erfüllt die Nebenbedingung:

$$K\widehat{\beta}_{H_0} = K\widehat{\beta} - K(X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - d) = d.$$

Insgesamt erhalten wir also

$$|Y - X\gamma|^2 = |Y - X\widehat{\beta}|^2 + |X(\widehat{\beta} - \widehat{\beta}_{H_0})|^2 + |X(\widehat{\beta}_{H_0} - \gamma)|^2, \tag{2.1}$$

was offensichtlich für $\gamma = \widehat{\beta}_{H_0}$ minimal ist.

(ii) Aus (2.1) mit $\gamma = \widehat{\beta}_{H_0}$ folgt durch Einsetzen von $\widehat{\beta}_{H_0}$

$$\begin{aligned}RSS_{H_0} - RSS =&|Y - X\widehat{\beta}_{H_0}|^2 - |Y - X\widehat{\beta}|^2 = |X(\widehat{\beta} - \widehat{\beta}_{H_0})|^2\\ =&(\widehat{\beta} - \widehat{\beta}_{H_0})^\top X^\top X(\widehat{\beta} - \widehat{\beta}_{H_0})\\ =&(K\widehat{\beta} - d)^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - d).\end{aligned}$$

Unter H_0 gilt für die Zufallsvariable $Z := K\widehat{\beta}$, dass $\mathbb{E}[Z] = d$ und $\operatorname{Var}(Z) = \sigma^2 K(X^\top X)^{-1}K^\top$. Aus der Normalverteilung von $\widehat{\beta}$ folgt daher $(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2(r)$.

(iii) Da $RSS_{H_0} - RSS$ eine Funktion von $\widehat{\beta}$ ist und somit unabhängig von RSS ist (Lemma 2.22), folgt die Verteilungsaussage für F aus der Charakterisierung der $F(r, n-p)$-Verteilung. □

Bemerkung 2.28. $W := rF$ heißt auch *Wald-Statistik*. Im Fall $d = 0$ ist $L := \{X\beta | \beta \in \mathbb{R}^k, K\beta = 0\}$ ein linearer Unterraum von $\operatorname{ran} X$ und $X\widehat{\beta}_{H_0} = \Pi_L Y$ die Orthogonalprojektion der Beobachtungen Y auf L. In diesem Fall gilt nach Pythagoras

$$RSS_{H_0} = |Y - \Pi_L Y|^2 = |Y - \Pi_X Y + (\Pi_X - \Pi_L)Y|^2 = |Y - \Pi_X Y|^2 + |X\widehat{\beta} - X\widehat{\beta}_{H_0}|^2,$$

so dass die Fisher-Statistik auch als

$$F = \frac{|X\widehat{\beta} - X\widehat{\beta}_{H_0}|^2}{r\widehat{\sigma}^2}$$

geschrieben werden kann.

Beispiel 2.26 (fortgesetzt). *Einsetzen von K und d liefert*

$$F = \frac{n-k}{RSS}\frac{(\widehat{\beta}_j - \widehat{\beta}_l)^2}{K(X^\top X)^{-1}K^\top}.$$

Wegen $\operatorname{Var}(\widehat{\beta}_j - \widehat{\beta}_l) = \operatorname{Var}(K\widehat{\beta}) = \sigma^2 K(X^\top X)^{-1}K^\top$ *ist* $\widehat{\operatorname{Var}}(\widehat{\beta}_j\text{-}\widehat{\beta}_l) = K(X^\top X)^{-1}K^\top\widehat{\sigma}^2$ *mit* $\widehat{\sigma}^2 = RSS/(n-k)$ *der natürliche (plug-in) Varianzschätzer. Damit können wir die Test-Statistik F als*

$$F = \frac{(\widehat{\beta}_j - \widehat{\beta}_l)^2}{\widehat{\operatorname{Var}}(\widehat{\beta}_j - \widehat{\beta}_l)} \overset{H_0}{\sim} F(1, n-k)$$

schreiben. Dieser F-Test ist äquivalent zum (zweiseitigen) t-Test mit der Teststatistik

$$T = \frac{\widehat{\beta}_j - \widehat{\beta}_l}{(\widehat{\mathrm{Var}}(\widehat{\beta}_j - \widehat{\beta}_l))^{1/2}} \sim t(n-k).$$

Beispiel 2.29 (Klimaentwicklung)**.** Wir folgen Beispiel 12.24 von Georgii (2007) und betrachten die mittleren Augusttemperaturen von 1799 bis 2008 in Karlsruhe (Quelle: `http://www.klimadiagramme.de/Europa/special01.htm`). Für die Jahre 1854 und 1945 liegen keine Daten vor, so dass wir $n = 208$ Beobachtungen haben. Eine polynomielle Regression in der Zeit t (in Jahrhunderten beginnend bei 1799) mit Graden $d = 1, \dots, 4$ liefert

$$\begin{aligned}
p_1(t) &= 18,7 + 0,1t,\\
p_2(t) &= 20,0 - 3,5t + 1,7t^2,\\
p_3(t) &= 19,5 - 0,6t - 1,7t^2 + 1,1t^3,\\
p_4(t) &= 19,4 + 0,5t - 4,1t^2 + 2,9t^3 - 0,4t^4.
\end{aligned}$$

Zunächst ist es plausibel, dass die zufälligen Schwankungen unabhängig von einander sind und als näherungsweise normalverteilt angenommen werden können (QQ-Plot). Um statistisch verwertbare Aussagen zu treffen, setzen wir noch das Niveau $\alpha = 0,05$ fest. Der Parametervektor ist $\beta = (\beta_0, \dots, \beta_d)^\top$. Welcher Grad des Regressionspolynoms ist sinnvoll?

Frage 1: Ist der positive Trend von p_1 signifikant? $H_0 : \beta_1 \leqslant 0$ vs. $H_1 : \beta_1 > 0$. Die zugehörige t-Statistik $T = \frac{\widehat{\beta}_1}{\widehat{\sigma}\sqrt{v^\top (X^\top X)^{-1} v}} \approx 0,62$ liegt deutlich unter dem kritischen Wert $q_{t(n-2),1-\alpha} \approx 1,65$ (einseitiger T-Test), so dass die Hypothese nicht verworfen werden kann.

Frage 2: Liegt den Beobachtungen ein linearer Zusammenhang zugrunde (im Modell mit $d = 4$)? $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Mittels Bemerkung 2.28 berechnen wir die Fisher-Statistik

$$F = \frac{\sum_{k=1}^n (p_4(t_k) - p_1(t_k))^2}{3\widehat{\sigma}^2} \approx 13,68 > 2,65 \approx q_{F(3,n-5),1-\alpha}.$$

Folglich kann die Hypothese abgelehnt werden und wir schlussfolgern, dass eine Regressionsgerade unzureichend ist.

Frage 3: Benötigen wir ein Polynom vierten Grades? $H_0 : \beta_4 = 0$. Die zugehörige t-Statistik hat den Wert $-0,41$ dessen Absolutbetrag kleiner als das Quantil $q_{t(n-5),0.975} \approx 1,97$ ist (zweiseitiger t-Test). Diese Nullhypothese kann also akzeptiert werden.

Frage 4: Benötigen wir ein Polynom dritten Grades ? $H_0 : \beta_3 = 0$ (im Modell mit $d = 3$). Die zugehörige t-Statistik hat den Wert $2,05$ dessen Absolutbetrag größer als das Quantil $q_{t(n-4),0.975} \approx 1,97$ ist. Die Hypothese kann also abgelehnt werden und der kubische Anteil im Regressionspolynom ist signifikant, d.h. p_3 ist signifikant besser geeignet die Beobachtungen zu beschreiben als p_2.

p_3 zeigt einen deutlichen Anstieg der Temperaturen im 19. Jahrhundert. Es sei bemerkt, dass wir hier nur eine Zeitreihe betrachtet haben und somit nicht auf einen allgemeinen Zusammenhang schließen können (Aufgabe der Klimatologen).

2.3 Varianzanalyse

Beispiel 2.30. Um den Einfluss von $k \in \mathbb{N}$ verschiedenen Düngemitteln auf den Ernteertrag zu vergleichen wird jedes Düngemittel $i \in \{1, \dots, k\}$ auf n_i verschiedenen Agrarflächen ausgebracht. Der durch Witterungseinflüsse etc. zufällige Ernteertrag kann mittels $Y_{ij} = \mu_i + \varepsilon_{ij}$ für $j = 1, \dots, n_i$ und $i = 1, \dots, k$ modelliert werden, wobei μ_i der mittlere Ernteertrag von Düngemittel i ist und ε_{ij} unabhängige, zentrierte Störgrößen sind. Wir fragen uns also ob $\mu_1 = \cdots = \mu_k$ gilt oder nicht.

Definition 2.31. Das Modell der einfaktoriellen Varianzanalyse (ANOVA1: (one-way) analysis of variance) ist gegeben durch Beobachtungen

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, k, j = 1, \dots, n_i,$$

mit iid.-verteilten Störgrößen $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Wir bezeichnen die erste Dimension als den Faktor und den Wert $i = 1, \dots, k$ als die Faktorstufe. Folglich geben $(n_i)_{i=1,\dots,k}$ die Anzahl der unabhängigen Versuchswiederholungen pro Faktor an und $n := \sum_{i=1}^{k} n_i$ ist der Gesamtstichprobenumfang. Gilt $n_1 = \cdots = n_k$, so sprechen wir von balanciertem Design.

Damit ist das ANOVA1-Modell ein Spezialfall des gewöhnlichen linearen Modells der Form

$$\mathbb{R}^n \ni Y := \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}_{=:X \in \mathbb{R}^{n \times k}} \cdot \underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}}_{=:\mu \in \mathbb{R}^k} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_k} \end{pmatrix}.$$

Beachte, dass $\operatorname{rank} X = k$. Die klassische Fragestellung der Varianzanalyse lautet: "Existieren Unterschiede in den Faktorstufen-spezifischen Mittelwerten μ_i?" oder anders formuliert "Hat der Faktor einen Einfluss auf die Response oder nicht?". Dies führt auf das *Testproblem*

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \exists i, l \in \{1, \dots, k\} : \mu_i \neq \mu_l.$$

Satz 2.32 (Streuungszerlegung)**.** *Im ANOVA1-Modell definieren wir das i-te Gruppenmittel, $i = 1, \dots, k$, bzw. das Gesamtmittel als*

$$\overline{Y}_{i\bullet} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad \textit{bzw.} \quad \overline{Y}_{\bullet\bullet} := \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

sowie

$$SSB := \sum_{i=1}^{k} n_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \quad \textit{und} \quad SSW := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\bullet})^2$$

(SSB: sum of squares between groups*; SSW:* sum of squares within groups *). Dann gilt*

$$SST := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = SSB + SSW.$$

Beweis. Es gilt

$$\begin{aligned} SST &= \sum_i \sum_j (Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i\bullet} + \overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \\ &= \sum_i \sum_j \big((Y_{ij} - \overline{Y}_{i\bullet})^2 + 2(Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) + (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \big), \end{aligned}$$

wobei

$$\begin{aligned} \sum_i \sum_j (Y_{ij} - \overline{Y}_{i\bullet})(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) &= \sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) \sum_j (Y_{ij} - \overline{Y}_{i\bullet}) \\ &= \sum_i (\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})(n_i \overline{Y}_{i\bullet} - n_i \overline{Y}_{i\bullet}) = 0. \end{aligned}$$

□

Offenbar spricht es gegen die Nullhypothese, wenn die Streuung zwischen den Gruppen größer ist als die Streuung innerhalb der Gruppen. Dies motiviert sowohl den Namen ANOVA als auch folgende Methode:

	Fg	Quadratsummen	Quadratmittel	F-Statistik
zwischen	$k-1$	$SSB = \sum_{i=1}^{k} n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$	$SSB/(k-1)$	$\frac{n-k}{k-1}\frac{SSB}{SSW}$
innerhalb	$n-k$	$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2$	$SSW/(n-k)$	
total	$n-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$	$SST/(n-1)$	

Tabelle 1: ANOVA-Tafel

Methode 8: Einfaktorielle Varianzanalyse (ANOVA1). Im Modell der einfaktoriellen Varianzanalyse testen wir

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{versus} \quad H_1 : \exists i, l \in \{1, \dots, k\} : \mu_i \neq \mu_l$$

zum Niveau $\alpha \in (0,1)$ durch den F-Test

$$\varphi_\mu(Y) = \mathbb{1}_{\{F(Y) > q_{F(k-1,n-k);1-\alpha}\}} \quad \text{mit} \quad F(Y) := \frac{n-k}{k-1}\frac{SSB}{SSW},$$

wobei $q_{F(k-1,n-k);1-\alpha}$ das $(1-\alpha)$-Quantil der $F(k-1, n-k)$-Verteilung ist.

Satz 2.33. *Im einfaktoriellen Varianzanalysemodell gilt:*

(i) Der Kleinste-Quadrate-Schätzer von $\mu = (\mu_1, \dots, \mu_k)^\top$ ist gegeben durch $\widehat{\mu} = (\overline{Y}_{1\bullet}, \dots, \overline{Y}_{k,\bullet})^\top$.

(ii) $SSW/\sigma^2 \sim \chi^2(n-k)$ und unter H_0 gilt $SSB/\sigma^2 \sim \chi^2(k-1)$

(iii) SSW und SSB sind unabhängig und somit $F := \frac{n-k}{k-1}\frac{SSB}{SSW} \overset{H_0}{\sim} F(k-1, n-k)$.

Beweis. (i) Nachrechnen zeigt

$$\widehat{\mu} = (X^\top X)^{-1} X^\top Y = \begin{pmatrix} 1/n_1 & & 0 \\ & \ddots & \\ 0 & & 1/n_k \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \vdots \\ \sum_{j=1}^{n_k} Y_{kj} \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\bullet} \\ \vdots \\ \overline{Y}_{k\bullet} \end{pmatrix}.$$

(ii)+(iii) Wegen $RSS = |Y - X\widehat{\mu}|^2 = SSW$ folgt $SSW/\sigma^2 \sim \chi^2(n-k)$ und die Unabhängigkeit von SSW und $\widehat{\mu}$ aus Lemma 2.22. Nach dem vorangegangen Satz gilt weiterhin $SSB = SST - SSW$. Somit folgt die Behauptung aus Satz 2.27, falls $SST = RSS_{H_0}$. Nun gilt

$$RSS_{H_0} = \min_{\mu \in \mathbb{R}} \Big| Y - X \underbrace{\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}}_{\in \mathbb{R}^k} \Big|^2 = \min_{\mu \in \mathbb{R}} \Big| Y - \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{=: X_0 \in \mathbb{R}^{n_\bullet \times 1}} \mu \Big|^2.$$

Dieses Minimierungsproblem wird gelöst durch $\widehat{\mu}_{H_0} = (X_0^\top X_0)^{-1} X_0^\top Y = n^{-1} \sum_{i,j} Y_{ij} = \overline{Y}_{\bullet\bullet}$. Damit folgt $RRS_{H_0} = SST$. □

Bemerkung 2.34. In der Effektdarstellung wird das einfaktorielle Varianzanalysemodell als

$$Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, k, j = 1, \dots, n_i,$$

geschrieben mit *"Intercept"* $\mu_0 := \frac{1}{n}\sum_{i=1}^{k} n_i \mu_i = \mathbb{E}[\overline{Y}_{\bullet\bullet}]$ und $\alpha_i := \mu_i - \mu_0$, den Effekt der Faktorstufe $i = 1, \dots, k$. Insbesondere muss in dieser Darstellung die Nebenbedingung $0 = \sum_{i=1}^{k} n_i \alpha_i$ oder äquivalent $n_k \alpha_k = -\sum_{i=1}^{k-1} n_i \alpha_i$ beachtet werden, damit die Designmatrix weiter vollen Rang hat. Der Parametervektor ist also gegeben durch $(\mu_0, \alpha_1, \dots, \alpha_{k-1})^\top$. Die F-Statistik um die Globalhypothese $H_0 : \alpha_1 = \cdots = \alpha_{k-1} = 0$ zu überprüfen, ist identisch zur Statistik aus Satz 2.33.

Beispiel 2.35 (Zweistichproben t-Test)**.** Soll die Gleichwertigkeit von bspw. zwei Düngemitteln getestet werden, ist $k = 2$ und das Testproblem $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. Wegen $n\overline{Y}_{\bullet\bullet} = n_1\overline{Y}_{1\bullet} + n_2\overline{Y}_{2\bullet}$ gilt

$$\begin{aligned} SSB &= n_1(\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet})^2 + n_2(\overline{Y}_{2\bullet} - \overline{Y}_{\bullet\bullet})^2 \\ &= n_1\overline{Y}_{1\bullet}^2 + n_2\overline{Y}_{2\bullet}^2 + n\overline{Y}_{\bullet\bullet}^2 - 2(n_1\overline{Y}_{1\bullet} + n_2\overline{Y}_{2\bullet})\overline{Y}_{\bullet\bullet} \\ &= n_1\overline{Y}_{1\bullet}^2 + n_2\overline{Y}_{2\bullet}^2 - \frac{1}{n}\big(n_1\overline{Y}_{1\bullet}^2 + n_2\overline{Y}_{2\bullet}^2\big)^2 = \frac{n_1 n_2}{n}\big(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}\big)^2. \end{aligned}$$

Somit ist

$$\varphi = \mathbb{1}_{\{|T| > q_{t(n-2),1-\alpha/2}\}} \quad \text{mit} \quad T := \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})SSW/(n-2)}}$$

mit dem $(1-\alpha/2)$-Quantil der $t(n-2)$-Verteilung $q_{t(n-2),1-\alpha/2}$ ein Test der Hypothese H_0 zum Niveau $\alpha \in (0,1)$.

Definition 2.36. Das Modell der zweifaktoriellen Varianzanalyse mit balanciertem Design (ANOVA2) ist gegeben durch Beobachtungen

$$\begin{aligned} Y_{ijk} =& \mu_{ij} + \varepsilon_{ijk}, \quad i = 1,\dots,I, j = 1,\dots,J, k = 1,\dots,K \\ =& \mu_0 + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \end{aligned}$$

mit $I, J, K \geqslant 2$, iid.-verteilten Störgrößen $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ und Nebenbedingungen (der Effektdarstellung)

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0.$$

Wir haben also zwei Faktoren mit Faktorstufen $i = 1,\dots,I$ und $j = 1,\dots,J$. (α_i) bzw. (β_j) heißen Haupteffekte des ersten bzw. zweiten Faktors. (γ_{ij}) heißen Interaktions- bzw. Wechselwirkungseffekte.

Das ANOVA2-Modell ist also ein lineares Modell mit zwei kategoriellen Kovariablen. Die Gesamtanzahl an Beobachtungen ist gegeben durch $n = I \cdot J \cdot K$. Die typische Testprobleme sind

$$H_0 : \forall i : \alpha_i = 0 \quad \text{versus} \quad H_1 : \exists i \in \{1,\dots,I\} : \alpha_i \neq 0, \tag{2.2}$$

$$H_0 : \forall j : \beta_j = 0 \quad \text{versus} \quad H_1 : \exists j \in \{1,\dots,J\} : \beta_j \neq 0, \tag{2.3}$$

$$H_0 : \forall i,j : \gamma_{ij} = 0 \quad \text{versus} \quad H_1 : \exists i \in \{1,\dots,I\}, j \in \{1,\dots,J\} : \gamma_{ij} \neq 0. \tag{2.4}$$

Satz 2.37. *Im zweifaktoriellen Varianzanalysemodell mit balanciertem Design gilt:*

(i) Die Kleinsten-Quadrate-Schätzer für μ_0, α_i, β_j und γ_{ij}, $i = 1,\dots,I-1, j = 1,\dots,J-1$, sind gegeben durch ($\bullet$ heißt, dass über die jeweilige Koordinate gemittelt wird)

$$\widehat{\mu}_0 = \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\alpha}_i = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}, \quad \widehat{\beta}_j = \overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet},$$
$$\widehat{\gamma}_{ij} = (\overline{Y}_{ij\bullet} - \overline{Y}_{\bullet\bullet\bullet}) - \widehat{\alpha}_i - \widehat{\beta}_j = \overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}.$$

(ii) Definieren wir

$$SSW := \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij\bullet})^2,$$

$$SSB_1 := JK\sum_{i=1}^{I}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2, \quad SSB_2 := IK\sum_{j=1}^{J}(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2,$$

$$SSB_{12} := K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet})^2,$$

dann können die Hypothesen (2.2), (2.3) bzw. (2.4) mit den F-Statistiken

$$\frac{IJ(K-1)}{I-1}\frac{SSB_1}{SSW} \sim F(I-1, IJ(K-1)), \quad \frac{IJ(K-1)}{J-1}\frac{SSB_2}{SSW} \sim F(J-1, IJ(K-1)) \quad \textit{bzw.}$$
$$\frac{IJ(K-1)}{(I-1)(J-1)}\frac{SSB_{12}}{SSW} \sim F\big((I-1)(J-1), IJ(K-1)\big)$$

getestet werden.

Beweis. Übung ☑. □

Bemerkung 2.38. Selbstverständlich erhält man analoge Resultate, wenn wir für jede "Zelle" $(i,j) \in \{1,\dots,I\} \times \{1,\dots,J\}$ verschiedene Stichprobenumfänge $n_{ij} \geqslant 2$ beobachten.

Beispiel 2.39. Ein Bauer möchte wissen ob die Größe seiner geernteten Kohlköpf sich für zwei verschiedene Kultursorten unterscheidet. Auch der Pflanztag könnte eine Rolle spielen.

3 Exponentialfamilien and verallgemeinerte lineare Modelle

3.1 Die Informationsungleichung

Der Satz von Gauß-Markov hat uns bereits ein Optimalitätsresultat geliefert, dass allerdings auf lineare Schätzer im linearen Modell eingeschränkt ist. Wir suchen nun allgemeiner nach unverzerrten Schätzern deren Schätzwerte möglichst wenig um den korrekten Wert streuen.

Definition 3.1. Sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell. Ein erwartungstreuer Schätzer T eines abgeleiteten Parameters $\rho(\vartheta)$ heißt varianzminimierend bzw. (gleichmäßig) bester Schätzer (UMVUE: uniformly minimum variance unbiased estimator), wenn für jeden weiteren erwartungstreuen Schätzer S gilt:

$$\mathrm{Var}_\vartheta(T) \leqslant \mathrm{Var}_\vartheta(S) \quad \text{für alle } \vartheta \in \Theta.$$

Wir werden zunächst eine untere Schranke für die Varianz beweisen und anschließend untersuchen, für welche Schätzer diese erreicht wird.

Definition 3.2. Ein vom Maß μ dominiertes, statistisches Modell $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ heißt regulär, wenn die folgenden Eigenschaften erfüllt sind:

(i) Θ ist eine offene Menge in $\mathbb{R}^d, d \geqslant 1$.

(ii) Die Likelihood-Funktion $L(\vartheta, x)$ ist auf $\Theta\times\mathcal{X}$ strikt positiv und nach ϑ stetig differenzierbar. Bezeichnen wir den Gradienten in ϑ mit $\nabla_\vartheta = (\frac{\partial}{\partial\vartheta_1}, \dots, \frac{\partial}{\partial\vartheta_d})^\top$, existiert insbesondere die Scorefunktion

$$U_\vartheta(x) := \nabla_\vartheta \log L(\vartheta, x) = \frac{\nabla_\vartheta L(\vartheta, x)}{L(\vartheta, x)}.$$

(iii) Für jedes $\vartheta \in \Theta$ existiert die Fisher-Information

$$I(\vartheta) := \mathbb{E}_\vartheta\Big[U_\vartheta(X)U_\vartheta(X)^\top\Big]$$

und ist positiv definit.

(iv) Es gilt die Vertauschungsrelation

$$\int h(x)\nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta \int h(x)L(\vartheta, x)\mu(\mathrm{d}x) \tag{3.1}$$

für $h(x) = 1$.

Ein Schätzer $T\colon \mathcal{X} \to \mathbb{R}$ heißt regulär, falls $\mathbb{E}[|T(X)|^2] < \infty$ und (3.1) auch für $h(x) = T(x)$ gilt.

Bemerkung 3.3.

(i) Der Satz von Lebesgue liefert eine hinreichende Bedingung für die Vertauschungsrelation (3.1): Sie gilt falls für jedes $\vartheta_0 \in \Theta$ eine Umgebung $V_{\vartheta_0} \subseteq \Theta$ existiert, so dass

$$\int_{\mathcal{X}} \sup_{\vartheta \in V_{\vartheta_0}} \Big|\nabla_\vartheta L(\vartheta, x)\Big| \mu(\mathrm{d}x) < \infty.$$

Außerdem kann man (3.1) für jedes gegebene Modell (und jeden Schätzer) explizit nachprüfen.

(ii) Als Konsequenz von (3.1) ergibt sich

$$\mathbb{E}_\vartheta[U_\vartheta] = \int \nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta \int L(\vartheta, x)\mu(\mathrm{d}x) = \nabla_\vartheta 1 = 0$$

und damit $\mathrm{Var}_\vartheta(U_\vartheta) = I(\vartheta)$.

(iii) Ist $L(\vartheta, x)$ in ϑ zweimal stetig differenzierbar und gilt (3.1) mit $h(x) = 1$ und L ersetzt mit $\frac{\partial L}{\partial \vartheta_i}$ für alle $i \in \{1, \dots, d\}$, dann gilt $I(\vartheta) = -\mathbb{E}_\vartheta[H_{U_\vartheta(X)}(\vartheta)]$ für die Hesse-Matrix $H_{U_\cdot(x)}$ der Scorefunktion $\vartheta \mapsto U_\vartheta(x)$ (Übung ☑).

(iv) Warum heißt $I(\vartheta)$ Information? Erstens: $I(\vartheta) = 0$ gilt auf einer Umgebung $\Theta_0 \subseteq \Theta$ genau dann, wenn $U_\vartheta(x) = 0$ für alle $\vartheta \in \Theta_0$ und μ-f.a. $x \in \mathcal{X}$, also wenn $L(\vartheta, x)$ für μ-f.s. konstant ist und somit keine Beobachtung die Parameter in Θ_0 unterscheiden kann (dieser Fall ist daher in der Definition ausgeschlossen). Zweitens, verhält sich die Fisher-Information bei unabhängigen Beobachtungen additiv: Ist $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ ein reguläres Modell mit Fisher-Information I, so hat das Produktmodell $(\mathcal{X}^n, \mathcal{F}^{\otimes n}, (\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta})$ die Fisher-Information $I^{\otimes n} = nI$ (Beweis als Übung ☑).

Satz 3.4 (Cramér-Rao-Ungleichung, Informationsschranke)**.** *Gegeben seien ein reguläres statistisches Modell $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$, eine zu schätzende stetig differenzierbare Funktion $\rho\colon \Theta \to \mathbb{R}$, und ein regulärer erwartungstreuer Schätzer T von ρ. Dann gilt*

$$\mathrm{Var}_\vartheta(T) \geqslant \Big(\nabla\rho(\vartheta)\Big)^\top I(\vartheta)^{-1}\nabla\rho(\vartheta) \quad \textit{für alle } \vartheta \in \Theta. \tag{3.2}$$

Beweis. Aus der Zentriertheit von U_ϑ und der Regularität und Erwartungstreue von T erhalten wir

$$\begin{aligned}\mathrm{Cov}_\vartheta(U_\vartheta, T) = \mathbb{E}_\vartheta[TU_\vartheta] &= \int_{\mathcal{X}} T(x)\nabla_\vartheta L(\vartheta, x)\mu(\mathrm{d}x)\\ &= \nabla \int_{\mathcal{X}} T(x)L(\vartheta, x)\mu(\mathrm{d}x) = \nabla\mathbb{E}_\vartheta[T] = \nabla\rho\end{aligned}$$

für alle $\vartheta \in \Theta$. Für jeden Vektor $e \in \mathbb{R}^d$ ergibt die Cauchy-Schwarz-Ungleichung somit

$$(\langle e, \nabla\rho\rangle)^2 = \mathrm{Cov}_\vartheta(\langle e, U_\vartheta\rangle, T)^2 \leqslant \mathrm{Var}_\vartheta(\langle e, U_\vartheta\rangle)\,\mathrm{Var}_\vartheta(T) = \langle I(\vartheta)e, e\rangle\,\mathrm{Var}_\vartheta(T),$$

also

$$\mathrm{Var}_\vartheta(T) \geqslant \frac{(\langle\nabla\rho, e\rangle)^2}{\langle I(\vartheta)e, e\rangle}.$$

Maximieren über $e \in \mathbb{R}^d$ ergibt mit $e = I(\vartheta)^{-1}\nabla\rho(\vartheta)$ die Behauptung. □

Definition 3.5. Ein regulärer erwartungstreuer Schätzer für den Gleichheit in (3.2) gilt, heißt Cramér-Rao-effizient.

Im Folgenden beschränken wir uns auf einparametrische ($d = 1$) Modelle.

Satz 3.6. *Unter den Bedingungen von Satz 3.4 mit $\Theta \subseteq \mathbb{R}$ erreicht der Schätzer T die untere Schranke für alle $\vartheta \in \Theta$ genau dann, wenn μ-f.ü. gilt*

$$T - \rho(\vartheta) = \rho'(\vartheta) I(\vartheta)^{-1} U_\vartheta \quad \text{für alle } \vartheta \in \Theta.$$

Falls $\rho' \neq 0$ ist dies äquivalent zu

$$L(\vartheta, x) = \exp\Big(\eta(\vartheta) T(x) - \zeta(\vartheta)\Big) c(x),$$

wobei $\eta\colon \Theta \to \mathbb{R}$ eine Stammfunktion von I/ρ', $c\colon \mathcal{X} \to (0, \infty)$ messbar und $\zeta(\vartheta) = \log \int c(x) \exp(\eta(\vartheta) T(x)) \mu(\mathrm{d}x)$ eine Normierungsfunktion sind.

Beweis. Definieren wir $v(\vartheta) := \rho'(\vartheta) I^{-1}(\vartheta)$ (konstant in x) erhalten wir wegen $\mathrm{Cov}_\vartheta(U_\vartheta, T) = \rho'(\vartheta)$

$$\begin{aligned} 0 &\leqslant \mathrm{Var}_\vartheta\left(T - v(\vartheta) U_\vartheta\right) \\ &= \mathrm{Var}_\vartheta(T) + v(\vartheta)^2 \mathrm{Var}_\vartheta(U_\vartheta) - 2 v(\vartheta) \mathrm{Cov}_\vartheta(U_\vartheta, T) = \mathrm{Var}_\vartheta(T) - \rho'(\vartheta)^2 I^{-1}(\vartheta), \end{aligned}$$

also wieder die Informationsungleichung. Gleichheit gilt genau dann, wenn $T - v(\vartheta) U_\vartheta$ $\mathbb{P}_\vartheta$-f.s. konstant also gleich seinem Erwartungswert $\rho(\vartheta)$ ist. Da $\mathbb{P}_\vartheta$ eine strikt positive μ-Dichte hat gilt $\mu(T - \rho(\vartheta) \neq v(\vartheta) U_\vartheta) = 0$. Wenn dies nun für alle $\vartheta \in \Theta$ gilt, so folgt sogar

$$\mu\big(T - \rho(\vartheta) \neq v(\vartheta) U_\vartheta \text{ für ein } \vartheta \in \Theta\big) = 0,$$

denn aus Stetigkeitsgründen kann man sich auf rationale ϑ beschränken und die abzählbare Vereinigung von Nullmengen ist wieder eine Nullmenge. Die explizite Form der Likelihood-Funktion folgt durch unbestimmte Integration bzgl. ϑ. □

Dieser Satz führt uns in natürlicher Weise auf eine wichtige Klasse von statistischen Modellen:

Definition 3.7. Es sei $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ ein von μ dominiertes statistisches Modell mit $\Theta \subseteq \mathbb{R}$ offen. Dann heißt $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$ (einparametrische) Exponentialfamilie in $\eta(\vartheta)$ und T, wenn messbare Funktionen $\eta\colon \Theta \to \mathbb{R}, T\colon \mathcal{X} \to \mathbb{R}$ und $c\colon \mathcal{X} \to (0, \infty)$ existieren, so dass

$$\frac{\mathrm{d}\mathbb{P}_\vartheta}{\mathrm{d}\mu}(x) = c(x) \exp\big(\eta(\vartheta) T(x) - \zeta(\vartheta)\big), \quad x \in \mathcal{X}, \vartheta \in \Theta,$$

wobei $\zeta(\vartheta) := \log \int c(x) \exp(\eta(\vartheta) T(x)) \mu(\mathrm{d}x)$. Dabei wird angenommen, dass T nicht μ-f.s. konstant ist. $\eta(\vartheta)$ heißt natürlicher Parameter der Exponentialfamilie und

$$\Xi := \Big\{\eta \in \mathbb{R} : \int_{\mathcal{X}} c(x) e^{\eta T(x)} \mu(\mathrm{d}x) \in (0, \infty)\Big\}$$

heißt natürlicher Parameterraum. Ist die Exponentialfamilie durch $\eta \in \Xi$ parametrisiert, dann wird sie als natürliche Exponentialfamilie bezeichnet.

Bemerkung 3.8.

(i) Die Darstellung ist nicht eindeutig, mit $a \neq 0$ erhält man beispielsweise eine Exponentialfamilie in $\widetilde{\eta}(\vartheta) = a\eta(\vartheta)$ und $\widetilde{T}(x) = T(x)/a$. Außerdem kann die Funktion c in das dominierenden Maß absorbiert werden: $\widetilde{\mu}(\mathrm{d}x) := c(x)\mu(\mathrm{d}x)$.

(ii) Die Identifizierbarkeitsfordung $\mathbb{P}_\vartheta \neq \mathbb{P}_{\vartheta'}$ für alle $\vartheta \neq \vartheta'$ ist äquivalent zur Injektivität von η.

Beispiel 3.9.

(i) $(\mathcal{N}(\mu,\sigma^2))_{\mu\in\mathbb{R}}$ mit $\sigma>0$ bekannt ist eine Exponentialfamilie in $\eta(\mu)=\mu/\sigma^2$ und $T(x)=x$:

$$L(\vartheta,x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x^2-2\mu x+\mu^2)/(2\sigma^2)},\quad x\in\mathbb{R}.$$

(ii) Die Familie der Poissonverteilungen $(Poiss(\lambda))_{\lambda>0}$ mit Intensitätsparameter λ bildet eine Exponentialfamilie mit natürlichem Parameter $\eta(\lambda)=\log\lambda$ und $T(x)=x$:

$$L(\lambda,x)=\frac{\lambda^x e^{-\lambda}}{x!}=\frac{1}{x!}e^{x\log\lambda-\lambda},\quad x\in\mathbb{Z}_+.$$

Lemma 3.10. *Ist ein statistisches Modell durch eine Exponentialfamilie in $\eta\colon\Theta\to\mathbb{R}$ und $T\colon\mathcal{X}\to\mathbb{R}$ mit differenzierbarem η gegeben, so ist dieses regulär. Ferner gilt*

(i) Jede Statistik $S\colon\mathcal{X}\to\mathbb{R}$ mit existierendem Erwartungswert ist regulär. $\rho(\vartheta):=\mathbb{E}_\vartheta[T]$ ist stetig differenzierbar mit $\rho'(\vartheta)=\eta'(\vartheta)\operatorname{Var}_\vartheta(T)\neq 0,\vartheta\in\Theta$.

(ii) Die Normierungsfunktion ζ ist auf $\Theta\subseteq\mathbb{R}$ stetig differenzierbar mit $\zeta'(\vartheta)=\eta'(\vartheta)\mathbb{E}_\vartheta[T]$ für $\vartheta\in\Theta$. Die Scorefunktion ist $U_\vartheta=\eta'(\vartheta)T-\zeta'(\vartheta)$.

(iii) Für die Fisher-Information gilt $I(\vartheta)=\eta'(\vartheta)\zeta''(\vartheta)=\eta'(\vartheta)\rho'(\vartheta)$ für alle $\vartheta\in\Theta$.

Beweis. O.B.d.A. ist $\eta(\vartheta)=\vartheta$ und somit $\eta'=1$ für alle $\vartheta\in\Theta$. Der allgemeine Fall ergibt sich durch Reparametrisierung und Anwendung der Kettenregel.

Schritt 1: Sei S eine beliebige reelle Statistik mit $S\in\mathcal{L}^1(\mathbb{P}_\vartheta)$ für alle $\vartheta\in\Theta$. Dann ist die Funktion

$$u_S(\vartheta):=e^{\zeta(\vartheta)}\mathbb{E}_\vartheta[S]=\int_{\mathcal{X}}S(x)e^{\vartheta T(x)}c(x)\mu(\mathrm{d}x)$$

auf Θ wohl definiert. Wir zeigen nun, dass u_s beliebig oft differenzierbar ist.

Ist $\vartheta\in\Theta$ und $t\in\mathbb{R}$ so klein, dass auch $\vartheta\pm t\in\Theta$, so gilt mittels monotoner Konvergenz

$$\begin{aligned}\sum_{k\geqslant 0}\frac{|t|^k}{k!}\int_{\mathcal{X}}|S(x)||T(x)|^k e^{\vartheta T(x)}c(x)\mu(\mathrm{d}x)&=\int_{\mathcal{X}}|S(x)|e^{\vartheta T(x)+|tT(x)|}c(x)\mathrm{d}x\\&\leqslant\int_{\mathcal{X}}|S(x)|\big(e^{(\vartheta+t)T(x)}+e^{(\vartheta-t)T(x)}\big)c(x)\mathrm{d}x<\infty.\end{aligned}$$

Also ist $ST^k\in\mathcal{L}^1(\mathbb{P}_\vartheta)$ für alle $\vartheta\in\Theta$ und insbesondere $T\in\mathcal{L}^2(\mathbb{P}_\vartheta)$ für alle ϑ. Ferner ist die Reihe

$$\sum_{k\geqslant 0}\frac{t^k}{k!}\int_{\mathcal{X}}S(x)T(x)^k e^{\vartheta T(x)}c(x)\mu(\mathrm{d}x)$$

absolut konvergent und Summation und Integration können vertauscht werden. Die Reihe nimmt daher den Wert $u_S(\vartheta+t)$ an. Damit ist u_S sogar analytisch.

Schritt 2: Es folgt $u_S'(\vartheta)=e^{\zeta(\vartheta)}\mathbb{E}_\vartheta[ST]$ und insbesondere $u_1'(\vartheta)=u_1(\vartheta)\mathbb{E}_\vartheta[T]$ sowie $u_1''(\vartheta)=u_1(\vartheta)\mathbb{E}_\vartheta[T^2]$. Für $\zeta(\vartheta)=\log u_1(\vartheta)$ bekommen wir also $\zeta'(\vartheta)=\mathbb{E}_\vartheta[T]=:\rho(\vartheta)$ und

$$\rho'(\vartheta)=\zeta''(\vartheta)=u_1''(\vartheta)/u_1(\vartheta)-(u_1'(\vartheta)/u_1(\vartheta))^2=\operatorname{Var}_\vartheta(T).$$

Aus der Differenzierbarkeit von ζ folgt

$$U_\vartheta=\frac{\partial}{\partial\vartheta}\log L(\vartheta,x)=T-\zeta'(\vartheta),\quad\vartheta\in\Theta$$

und somit $I(\vartheta)=\operatorname{Var}_\vartheta(U_\vartheta)=\operatorname{Var}_\vartheta(T)>0$. Weiter können wir schreiben

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_\vartheta[S]=&\big(u_S(\vartheta)e^{-\zeta(\vartheta)}\big)'=\big(u_S'(\vartheta)-u_S(\vartheta)\zeta'(\vartheta)\big)e^{-\zeta(\vartheta)}\\=&\mathbb{E}_\vartheta[ST]-\mathbb{E}_\vartheta[S]\zeta'(\vartheta)=\mathbb{E}_\vartheta[SU_\vartheta]\\=&\int_{\mathcal{X}}S(x)\frac{\partial}{\partial\vartheta}L(\vartheta,x)\mu(\mathrm{d}x).\end{aligned}$$

Daher gilt einerseits (3.1) für alle $h\in\mathcal{L}^1(\mathbb{P}_\vartheta)$ und anderseits folgt die Regularität des Modells. □

Korollar 3.11 (Existenz von besten Schätzern)**.** *Für jedes statistische Modell gegeben durch eine Exponentialfamlie mit differenzierbarem η und $\eta' \neq 0$ ist die zugrunde liegende Statistik T ein bester und Cramér-Rao-effizienter Schätzer für $\rho(\vartheta) := \mathbb{E}_\vartheta[T] = \zeta'(\vartheta)/\eta'(\vartheta)$. In dem Fall gilt*

$$\mathrm{Var}_\vartheta(T) = \rho'(\vartheta)/\eta'(\vartheta) \quad \textit{und} \quad I(\vartheta) = \eta'(\vartheta)\rho'(\vartheta) \quad \textit{für alle } \vartheta \in \Theta.$$

Für natürliche Exponentialfamilien gilt also insbesondere $\mathrm{Var}_\eta(T) = I(\eta)$.

Beweis. Folgt unmittelbar aus Satz 3.4 und Lemma 3.10. Für natürliche Exponentialfamilien gilt also $\mathrm{Var}_\eta(T) = \rho'(\eta) = I(\eta)$ und die Informationsschranke ist gegeben durch $\rho'(\eta)^2/I(\eta) = I(\eta)$. □

Beispiel 3.9 (fortgesetzt).

(i) *$(\mathcal{N}(\mu, \sigma^2))_{\mu\in\mathbb{R}}$ und bekanntem $\sigma > 0$ ist wie oben gesehen eine Exponentialfamilie in $\eta(\mu) = \mu/\sigma^2$, $T(x) = x$ und mit $\zeta(\mu) = \mu^2/(2\sigma^2)$. Somit ist $\rho(\mu) = \mathbb{E}_\mu[T] = \mu$ und $\mathrm{Var}_\mu(T) = \sigma^2$. Da T nicht von $\sigma > 0$ abhängt, ist T sogar bester Schätzer für den Erwartungswert für alle Normalverteilungen.*

(ii) *Für die Exponentialfamilie $(Poiss(\lambda))_{\lambda>0}$ in $\eta(\lambda) = \log\lambda$ und $T(x) = x$ gilt $\zeta(\lambda) = \lambda$. Wegen $\rho(\lambda) = \mathbb{E}_\lambda[T] = \lambda$ und $\mathrm{Var}_\lambda(T) = \lambda$ ist T bester Schätzer für λ.*

Lemma 3.12. *Ist $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ auf $(\mathcal{X}, \mathscr{F})$ eine Exponentialfamilie in $\eta\colon \Theta \to \mathbb{R}$ und $T\colon \mathcal{X} \to \mathbb{R}$ so ist $(\mathbb{P}_\vartheta^{\otimes n})_{\vartheta\in\Theta}$ eine Exponentialfamilie auf $(\mathcal{X}^n, \mathscr{F}^{\otimes n})$ mit zugrundeliegender Statistik $T_n = \frac{1}{n}\sum_{i=1}^n T\circ X_i$. Ist η differenzierbar mit $\eta' \neq 0$, folgt insbesondere, dass T_n ein bester Schätzer für $\rho(\vartheta) = \mathbb{E}_\vartheta[T]$ ist.*

Beweis. Übung ☑. □

Abschließend klären wir noch die Frage was das Maximum-Likelihood-Prinzip für natürliche Exponentialfamilien ergibt.

Lemma 3.13. *Ist $(\mathbb{P}_\vartheta)_{\vartheta\in\Theta}$ auf $(\mathcal{X}, \mathscr{F})$ eine natürliche Exponentialfamilie in $\eta \in \Xi$ und $T\colon \mathcal{X} \to \mathbb{R}$, dann ist T auf dem Ereignis $\{T(X) \in \mathrm{ran}(\zeta')\}$ der eindeutige Maximum-Likelihood-Schätzer des Parameters $\rho(\eta) := \mathbb{E}_\eta[T]$. Ferner ist $\zeta'\colon \Theta \to \mathbb{R}$ invertierbar und der eindeutige Maximum-Likelihood-Schätzer des natürlichen Parameters η ist gegeben durch*

$$\widehat{\eta} = (\zeta')^{-1}(T(X)).$$

Beweis. Um die Maximalstelle der Likelihood-Funktion zu finden, setzen wir die Scorefunktion gleich null. Auf $\{T(X) \in \mathrm{ran}(\zeta')\}$ gilt

$$\partial_\eta \log L(\eta, x) = U_\eta(x) = 0 \quad \Leftrightarrow \quad T(x) = \zeta'(\eta).$$

Da $\partial_\eta^2 \log L(\eta, x) = -\zeta''(\eta) = -\mathrm{Var}_\eta(T) < 0$, ist $\eta \mapsto -\log L(\eta, x)$ konvex und somit T der eindeutige Maximum-Likelihood-Schätzer des Parameters $\rho(\eta) = \zeta'(\eta)$. Aus $\zeta'' > 0$ folgt außerdem, dass ζ' invertierbar ist, so dass der Maximum-Likelihood-Schätzer des natürlichen Parameters gegeben ist durch $(\zeta')^{-1} \circ T$. □

3.2 Verallgemeinerte Lineare Modelle

Mit Hilfe von Exponentialfamilien wollen wir nun lineare Modelle verallgemeinern. Wie in Beispiel 3.9 gesehen bildet $(\mathcal{N}(\mu, \sigma^2))_{\mu\in\mathbb{R}}$ eine Exponentialfamilie mit natürlichem Parameter $\eta(\mu) = \mu/\sigma^2$ und Statistik $T(x) = x$, die ein effizienter Schätzer des Parameters $\rho(\mu) = \mathbb{E}_\mu[T] = \mu$ ist. Im gewöhnlichen linearen Modell sind nun die Beobachtungen gegeben durch

$$\mathbb{R}^n \ni Y = X\beta + \varepsilon,$$

mit Parametervektor $\beta \in \mathbb{R}^k$, Designmatrix $X \in \mathbb{R}^{n\times k}$ und iid. Fehlervariablen $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ mit Varianz $\sigma > 0$. Schreiben wir die Designmatrix als

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{mit Zeilenvektoren } x_1,\dots,x_n \in \mathbb{R}^k,$$

ist Beobachtung Y_i gemäß $\mathcal{N}(x_i\beta,\sigma^2)$ verteilt, folgt also einer Exponentialfamilie mit $\eta_i(\beta) = x_i\beta/\sigma^2$ und $\rho_i(\beta) = x_i\beta$, $i = 1,\dots,n$. Lassen wir nun andere Exponentialfamilien zu, können wir sowohl Situationen modellieren in den der Zusammenhang zwischen $\mathbb{E}[Y_i]$ und den Kovariablen (codiert in der Designmatrix X) nichtlinear ist als auch diskrete Beobachtungen Y_i zulassen.

Definition 3.14. Auf einem Produktmodell $(\mathcal{X}^n, \mathscr{F}^{\otimes n})$ liegt ein verallgemeinertes lineares Modell (GLM: generalized linear model) mit n unabhängigen Beobachtungen $Y_1,\dots,Y_n$ vor, falls die Randverteilungen von Y_i durch natürliche Exponentialfamilien gegeben sind mit Dichten

$$\frac{\mathrm{d}\mathbb{P}^{Y_i}_{\eta_i}}{\mathrm{d}\mu}(y_i) = \exp\Big(\frac{\eta_i y_i - \zeta(\eta_i)}{\varphi}\Big)c(y_i,\varphi), \quad i = 1,\dots,n,$$

bzgl. einem dominierenden Maß μ, mit unbekanntem Dispersionsparameter $\varphi > 0$,

$$\eta_i \in \Xi = \Big\{\eta \in \mathbb{R}\colon \int_{\mathcal{X}} e^{\eta y/\varphi} c(y,\varphi)\mu(\mathrm{d}y) \in (0,\infty)\Big\} \subseteq \mathbb{R}$$

für alle i und bekannten Funktionen $\zeta\colon \Xi \to \mathbb{R}$ und $c\colon \mathcal{X} \to \mathbb{R}_+$ mit $\zeta''(\eta) > 0$ für alle inneren Punkte $\eta \in \Xi^\circ$. Setze $\rho(\eta_i) := \mathbb{E}_\eta[Y_i]$. Für einen unbekannten Parametervektor $\beta \in \mathbb{R}^k$, eine Designmatrix $X \in \mathbb{R}^{n\times k}$ und eine bijektive, stetig differenzierbare Funktion $g\colon \mathbb{R} \to \mathbb{R}$ gelte weiter

$$\begin{pmatrix} g(\rho(\eta_1)) \\ \vdots \\ g(\rho(\eta_n)) \end{pmatrix} = X\beta.$$

g heißt Linkfunktion. Falls $\rho = g^{-1}$, gilt $(\eta_1,\dots,\eta_n)^\top = X\beta$ und g heißt kanonische Linkfunktion (oder kanonischer Link).

Während β der interessierende Parameter ist, wird φ als Störparameter angesehen. Für fixiertes φ ist Y_i also gemäß einer natürlichen Exponentialfamilie in $T(y) = y/\varphi$ verteilt. Aus den Eigenschaften natürlicher Exponentialfamilien folgt

$$\mathbb{E}_{\beta,\varphi}[Y_i] = \zeta'(\eta_i) \quad \text{und} \quad \mathrm{Var}_{\beta,\varphi}(Y_i) = \varphi\zeta''(\eta_i), \quad i = 1,\dots,n.$$

Beispiel 3.15. Das gewöhnliche lineare Modell ist ein GLM mit kanonischer Linkfunktion $g(x) = x$, $\zeta(\eta) = \eta^2/2$ und Dispersionsparameter $\varphi = \sigma^2$. Lassen wir allgemeinere Linkfunktionen zu erhalten wir *nicht-lineare Regressionsmodelle* (mit normalverteilten Fehlern) gegeben durch Beobachtungen $Y_i \sim \mathcal{N}(g^{-1}((X\beta)_i),\varphi)$.

Der Dispersionsparameter wird dazu verwendet eine Unterschätzung der (empirisch beobachteten) Varianz durch das Modell auszugleichen (siehe Übung ☑).

Um den unbekannten Parametervektor β in einem verallgemeinerten linearen Modell zu schätzen, verwenden wir den Maximum-Likelihood-Ansatz. Da ζ' streng monoton wachsend und die Linkfunktion g invertierbar sind, existiert die Funktion $\psi := (g\circ\rho)^{-1}$. Ist $x_i \in \mathbb{R}^k$ wieder die i-te Zeile von X, kann Loglikelihood-Funktion geschrieben werden als

$$\log L(\beta,\varphi;y) = \sum_{i=1}^n \Big(\frac{\psi(x_i\beta)y_i - \zeta(\psi(x_i\beta))}{\varphi} + \log(c(y_i,\varphi))\Big).$$

Als notwendige Bedingung an einen Maximum-Likelihood-Schätzer $\widehat{\beta}$ erhalten wir durch Ableiten

$$\nabla_\beta \log L(\widehat{\beta},\varphi;y) = \frac{1}{\varphi}\sum_{i=1}^n \big(y_i - \rho(\psi(x_i\widehat{\beta}))\big)\psi'(x_i\widehat{\beta})x_i^\top = 0. \tag{3.3}$$

Lemma 3.16. *In einem verallgemeinerten linearen Modell mit kanonischer Linkfunktion ist die Fisher-Information gegeben durch*

$$I(\beta) = \frac{1}{\varphi} \sum_{i=1}^{n} \zeta''(x_i\beta) x_i^\top x_i \in \mathbb{R}^{k\times k}.$$

Ist $I(\beta)$ positiv definit für alle β und existiert eine Lösung $\widehat{\beta}$ von (3.3), so ist $\widehat{\beta}$ der eindeutige Maximum-Likelihood-Schätzer von β.

Beweis. Aus Lemma 3.10 folgt, dass die Fisher-Information im natürlichen Parameter $(\eta_1, \dots, \eta_n)^\top$ gegeben ist durch $\frac{1}{\varphi}\sum_{i=1}^n \zeta''(\eta_i)$. Die Reparametrisierung $\eta_i = x_i\beta$ zusammen mit der Kettenregel ergibt die Darstellung von $I(\beta)$.

Der kanonische Link ist gegeben durch $g = \rho^{-1}$, so dass ψ in (3.3) die Identität ist. Wegen $\rho = \zeta'$, gilt also

$$\frac{\partial^2 \log L(\beta, \varphi; y)}{\partial\beta\partial\beta^\top} = -\frac{1}{\varphi}\sum_{i=1}^{n} \zeta''(x_i\beta) x_i^\top x_i = -I(\beta).$$

Da $I(\beta) > 0$, ist $\beta \mapsto -\log L(\beta, \varphi; y)$ streng konvex und somit $\widehat{\beta}$ der eindeutige Maximum-Likelihood-Schätzer. □

Bemerkung 3.17.

(i) Typischerweise besitzt $\widehat{\beta}$ keine geschlossene Form mehr und muss durch numerische Verfahren bestimmt werden. *Fishers Scoring-Methode* verwendet hierfür das iterative Verfahren

$$\widehat{\beta}^{(t+1)} = \widehat{\beta}^{(t)} + I(\beta)^{-1}\nabla_\beta \log L(\widehat{\beta}^{(t)}, \varphi; y), \quad t = 0, 1, \dots$$

(Beachte, dass sich der unbekannte Dispersionsparameter φ gerade rauskürzt). Für den kanonischen Link ist dieses Verfahren äquivalent zur *Newton-Raphson-Methode.*

(ii) Ist g nicht der kanonische Link ist eine Lösung von (3.3) nicht notwendigerweise ein Maximum-Likelihood-Schätzer.

Zwei wichtige Beispielklassen für verallgemeinerte lineare Modelle sind die *Poisson-Regression* und die *logistische Regression*, die abschließend eingeführt werden.

Die Poisson-Regression modelliert unabhängige Poisson-verteilte Beobachtungen, deren Intensitätsparameter von Kovariablen abhängen. Sie eignet sich also für Beobachtungen die Zähldatenstruktur haben. Wir hatten bereits gesehen dass die Familie $(Poiss(\lambda))_{\lambda>0}$ eine Exponentialfamilie in $\eta(\lambda) = \log\lambda$ und $T(x) = x$ ist: Bezüglich des Zählmaßes ist die Likelihood-Funktion gegeben durch

$$L(\lambda, x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{x\log\lambda - \lambda}, \quad x \in \mathbb{Z}_+,$$

und es gilt $\rho(\lambda) = \mathbb{E}_\lambda[T] = \lambda$.

Definition 3.18. Ein verallgemeinertes lineares Modell auf $(\mathbb{Z}_+^n, \mathscr{P}(\mathbb{Z}_+^n))$ heißt Poisson-Regression, falls die unabhängigen Beobachtungen Y_i $Poiss(\lambda_i)$-verteilt sind, wobei $\lambda_i = e^{\eta_i}$ mit natürlichen Parametern $\eta_i \in \mathbb{R}, i = 1, \dots, n$, und

$$\eta = (\eta_1, \dots, \eta_n)^\top = X\beta$$

mit unbekanntem $\beta \in \mathbb{R}^k$ und Designmatrix $X \in \mathbb{R}^{n\times k}$.

Bemerkung 3.19. Wir verwenden hier also den kanonischen Link $g(\lambda) = \log\lambda$. In der Praxis wird oft das erweiterte Modell $Y_i \sim Poiss(\lambda_i \cdot s_i)$ verwendet für einen so genannten *Zählrahmen* $s_i > 0$, $i = 1, \dots, n$. Dann gilt $\mathbb{E}_\beta[Y_i] = \exp(x_i\beta + \log(s_i))$ mit den Zeilen x_i von X. Der Term $\log(s_i)$ wird als *Offset* bezeichnet, da er jeder Beobachtung einen individuellen "Intercept" zuweist.

Beispiel 3.20. In einem großen Krankenhaus wird die Anzahl der Beschwerden über $n = 44$ Notfallärzten untersucht (Daten aus Le (2003)). Der Zählrahmen pro Arzt ist die Anzahl an Patientenbesuchen, die vier zu berücksichtigenden Kovariablen sind Vergütung (in \$/h), Erfahrung (in h), Geschlecht und Facharztausbildung (ja/nein).

Lemma 3.21. *Die Familie der Bernoulli-Verteilungen* $(Bernoulli(p))_{p\in(0,1)}$ *bildet eine Exponentialfamilie in* $\eta(p) = \log \frac{p}{1-p}$ *und* $T(x) = x$.

Beweis. Die vom Zählmaß dominierte Familie besitzt die Likelihood-Funktion

$$L(p,x) = p^x(1-p)^{1-x} = (1-p)\Big(\frac{p}{1-p}\Big)^x = \exp\Big(x\log\Big(\frac{p}{1-p}\Big) + \log(1-p)\Big), \quad x \in \{0,1\}. \quad \square$$

Definition 3.22. Ein verallgemeinertes lineares Modell auf $(\{0,1\}^n, \mathscr{P}(\{0,1\}^n))$ heißt logistische Regression, falls die unabhängigen Beobachtungen Y_i $Bernoulli(p_i)$-verteilt sind, $i = 1,\dots,n$, mit natürlichem Parameterraum $\mathbb{R}$, der kanonischen Link-Funktion $g\colon (0,1) \to \mathbb{R}, g(p) = \log(p/(1-p))$ und

$$\eta = (g(p_1),\dots,g(p_n))^\top = X\beta$$

mit unbekanntem $\beta \in \mathbb{R}^k$ und Designmatrix $X \in \mathbb{R}^{n\times k}$. Die Funktion g heißt Logit-Funktion und ihre Umkehrfunktion $g^{-1}\colon \mathbb{R} \to (0,1), g^{-1}(x) = (1+e^{-x})^{-1}$ heißt logistische Funktion.

Bemerkung 3.23. Es gilt also $\mathbb{E}[Y_i] = g^{-1}(\eta_i) = e^{\eta_i}/(1+e^{\eta_i})$, wobei die Funktion g^{-1} gerade die Verteilungsfunktion der standardisierten logistischen Verteilung ist (welche im Allgmeinen einen Mittelwerts- und einen Streuungsparameter besitzt). Das motiviert ein populäres Beispiel für eine nicht kanonische Linkfunktion: die Probit-Funktion $g(\lambda) = \Phi^{-1}(\lambda)$ mit der Verteilungsfunktion der Standardnormalverteilung Φ.

Da wir hier ein Modell gefunden haben um $\{0,1\}$-wertige Zufallsvariablen durch Kovariablen zu erklären, werden wir die logistische Regression im nächsten Kapitel zur Klassifikation verwenden.

3.3 Ergänzung: Numerische Bestimmung des Maximum-Likelihood-Schätzers

Das vermutlich grundlegendste numerische Verfahren zur Bestimmung von Nullstellen ist das *Newton-Verfahren* oder *Newton-Raphson-Verfahren:*

- Ziel: Finde $x^* \in \mathbb{R}$: $f(x^*) = 0$ für eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$.
- Verfahren:
  (i) Wähle einen Startpunkt $x_0 \in \mathbb{R}$ (der möglichst nahe an x^* liegen sollte).
  (ii) Approximiere x^* mit der rekursiven Vorschrift

$$x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{falls} \quad f'(x_n) \neq 0$$

- Abbruchkriterien: $|f(x_n)| < \varepsilon$ oder $|x_{n+1} - n_n| < \varepsilon$ für ein $\varepsilon > 0$.

Geometrisch ist x_{n+1} genau die Nullstelle der Tangente $y = f(x_n) + f'(x_n)(x - x_n)$ an f im Punkt $(x_n, f(x_n))$. Im allgemeineren Fall $f\colon \mathbb{R}^k \to \mathbb{R}^k$ erhalten wir die Rekursionsvorschrift

$$J_f(x_n)(x_{n+1} - x_n) = -f(x_n) \quad \Longleftrightarrow \quad x_{n+1} = x_n - J_f(x_n)^{-1} f(x_n)$$

mit der Jacobi-Matrix $J_f(x) = (\frac{\partial f_i}{\partial x_j})_{i,j=1,\dots,k} \in \mathbb{R}^{k\times k}$ falls diese positiv definit ist.

Das Newton-Verfahren soll nun verwendet werden um den Maximum-Likelihood-Schätzer $\widehat{\beta}$ in einem verallgemeinerten linearen Modell $(\mathcal{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_{\beta,\varphi}^{\otimes n})_{\beta\in\mathbb{R}^k,\varphi>0})$ mit kanonischem Link zu bestimmen. Zur Erinnerung ist in diesem Fall die Likelihood-Funktion gegeben durch

$$L(\beta,\varphi;y) = \prod_{i=1}^n \exp\Big(\frac{(x_i\beta)y_i - \zeta(x_i\beta)}{\varphi}\Big) c(y_i,\varphi)$$

mit n Zeilenvektoren $x_i \in \mathbb{R}^k$. Setzen wir also

$$f(\beta) = \nabla_\beta \log L(\beta, \varphi; y) = \frac{1}{\varphi} \sum_{i=1}^{n} \big(y_i - \zeta'(x_i\beta)\big) x_i^\top,$$

dann ist die Jacobi-Matrix gleich der Hesse-Matrix der Loglikelihood-Funktion

$$\begin{aligned} J_f(\beta) = \Big(\frac{\partial_{\beta_l} \log L(\beta, \varphi; y)}{\partial \beta_j}\Big)_{l,j=1,\dots,k} &= -\frac{1}{\varphi}\Big(\sum_{i=1}^{n} \zeta''(x_i\beta) x_{i,l} x_{i,j}\Big)_{l,j=1,\dots,k} \\ &= -\frac{1}{\varphi} \sum_{i=1}^{n} \zeta''(x_i\beta) x_i^\top x_i. \end{aligned}$$

Da diese nicht mehr von y abhängt, erhalten wir

$$J_f(\beta) = \mathbb{E}_\beta\big[H_{\log L(\cdot,\varphi,Y)}(\beta)\big] = -I(\beta).$$

Einsetzen in obige Iterationsvorschrift liefert *Fishers Scoring-Methode:*

$$\begin{aligned} \widehat{\beta}^{(t+1)} &= \widehat{\beta}^{(t)} + I(\beta)^{-1} \nabla_\beta \log L(\widehat{\beta}^{(t)}, \varphi; y) \\ &= \widehat{\beta}^{(t)} - \Big(\sum_{i=1}^{n} \zeta''(x_i\widehat{\beta}^{(t)}) x_i^\top x_i\Big)^{-1} \sum_{i=1}^{n} \big(Y_i - \zeta'(x_i\widehat{\beta}^{(t)})\big) x_i^\top, \quad t \in \mathbb{N}, \end{aligned}$$

wobei wir $\widehat{\beta}^{(0)} = 0$ setzen.

4 Klassifikation

Während im linearen Modell die Zielvariable quantitativ ist, gibt es viele Situationen in denen die Daten *qualitativ* bzw. *kategoriell* sind. Das Grundprinzip der *Klassifikation* ist anhand einer sogenannten *Trainingsmenge* $(x_1, Y_1), \dots, (x_n, Y_n)$ zu lernen, die Klassen zu unterscheiden, um anschließend vorherzusagen, zu welcher Klasse Beobachtungen zu neuen $x_{n+1}, \dots, x_{n+m}$ gehören (*statistisches Lernen*). Anders ausgedrückt, sollen $x_{n+1}, \dots, x_{n+m}$ *klassifiziert* werden.

Beispiel 4.1. Auf Grundlage vom monatlichen Kontostand der Kreditkarte und dem Jahreseinkommen soll vorhergesagt werden ob jemand zahlungsunfähig wird oder nicht. Als Trainingsdatensatz haben wir Daten $(x_{i,1}, x_{i,2}, Y_i)$ für $i = 1, \dots, n$ Personen gegeben, wobei $x_{i,1}$ bzw. $x_{i,2}$ der monatliche Kontostand der Kreditkarte bzw. das Jahreseinkommen von Person i sind und Y_i die Frage "Zahlungsunfähig?" mit "Ja" oder "Nein" beantwortet (simulierter `default`-Datensatz aus James et al. (2013)). Etwa 3% der Personen sind zahlungsunfähig. Beachte, dass in realen Anwendungen die Beziehung zwischen Ko- und Zielvariablen typischerweise nicht so eindeutig sind.

4.1 Logistische Regression

Stammen die Zielvariablen nur aus zwei verschiedenen Klassen (die mit 0 und 1 codiert werden), bietet sich die logistische Regression aus Kapitel 3 als Modell an. Zur Erinnerung heißt ein statisches Experiment $(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), (\mathbb{P}_\beta{}^{\otimes n})_{\beta\in\mathbb{R}^k})$ *multiple logistische Regression* mit Kovariablen $x_i = (1, x_{i,1}, \dots, x_{i,k-1}) \in \mathbb{R}^k$ (Zeilenvektor mit Absolutglied), falls Y_i *Bernoulli*(p_i)-verteilt ist, wobei $p_i = p(x_i, \beta)$ gegeben ist durch

$$\log\Big(\frac{p_i}{1-p_i}\Big) = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{i,j} \quad \text{für } i = 1, \dots, n.$$

Äquivalent gilt

$$p(x_i, \beta) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}.$$

Die Wahrscheinlichkeit, dass Y_i zur Klasse 1 gehört, wird also durch die $k-1$ Kovariablen erklärt.

Methode 9: Klassifikation mittels logistischer Regression. Nach Schätzung des Parametervektors $\widehat{\beta}$ auf der Trainingsmenge $(x_i, Y_i)_{i=1,\dots,n}$ können wir für eine jede neue Kovariablenrealisierung $x_{n+1} = (1, x_{n+1,1}, \dots, x_{n+1,k-1})$ (Zeilenvektor) einen zugehörigen Wert

$$\widehat{p}_{n+1} = p(x_{n+1}, \widehat{\beta}) = \frac{e^{x_{n+1}\widehat{\beta}}}{1 + e^{x_{n+1}\widehat{\beta}}}$$

"vorhersagen" und x_{n+1} der Klasse 1 zuordnen, falls $\widehat{p}_{n+1} \geqslant \tau$ für einen Schwellenwert $\tau \in [0,1]$. Andernfalls klassifizieren wir x_{n+1} mit 0.

Der Maximum-Likelihood-Ansatz führt auf die Maximierung der Loglikelihood-Funktion

$$\begin{aligned}\ell(\beta, y) := \log L(\beta, y) &= \sum_{i=1}^{n} \big(y_i \log p(x_i, \beta) + (1 - y_i)(1 - p(x_i, \beta)\big) \\ &= \sum_{i=1}^{n} \Big(y_i(x_i\beta) - \log\big(1 + e^{x_i\beta}\big)\Big).\end{aligned}$$

Nullsetzen des Gradienten führt auf k Gleichungen, die nichtlinear in β sind. Um den Maximum-Likelihood-Schätzer numerisch zu bestimmen, verwenden wir wieder das Newton-Verfahren. Dieses führt uns auf die *iterativ neugewichteten Kleinste-Quadrate-Methode* (IRLS: iteratively reweighted least squares):

Lemma 4.2. *In der logistischen Regression mit Designmatrix X von vollem Rang ist der $(t+1)$ste Iterationsschritt von Fishers Scoring-Methode gegeben durch*

$$\widehat{\beta}^{(t+1)} = (X^\top W_{\widehat{\beta}^{(t)}} X)^{-1} X^\top W_{\widehat{\beta}^{(t)}} Z_{\widehat{\beta}^{(t)}} = \arg\min_b \left| W_{\widehat{\beta}^{(t)}}^{1/2} (Z_{\widehat{\beta}^{(t)}} - Xb) \right|^2$$

mit adjustiertem Responsevektor $Z_\beta = X\beta + W_\beta^{-1}(Y - p_\beta)$, wobei

$$\begin{aligned} p_\beta &= \big(p(x_1, \beta), \dots, p(x_n, \beta)\big)^\top \in \mathbb{R}^n \quad \textit{und} \\ W_\beta &= \operatorname{diag}\big(p(x_1, \beta)(1 - p(x_1, \beta), \dots, p(x_n, \beta)(1 - p(x_n, \beta))\big) \in \mathbb{R}^{n \times n}. \end{aligned}$$

Beweis. Für die logistische Funktion $g(x) = e^x/(1 + e^x)$ gilt $g'(x) = g(x)(1 - g(x))$. Aus der expliziten Form der Loglikelihood-Funktion $\ell(\beta)$ folgt damit, dass Scorefunction und Hesse-Matrix gegeben sind durch

$$\nabla_\beta \ell(\beta, y) = X^\top (y - p_\beta) \quad \text{bzw.} \quad H_{\ell(\cdot,y)}(\beta) = -X^\top W_\beta X.$$

Somit ist der Iterationsschritt von Fishers Scoring-Methode

$$\begin{aligned}\widehat{\beta}^{(t+1)} &= \widehat{\beta}^{(t)} + \big(X^\top W_{\widehat{\beta}^{(t)}} X\big)^{-1} X^\top (Y - p_{\widehat{\beta}^{(t)}}) \\ &= \big(X^\top W_{\widehat{\beta}^{(t)}} X\big)^{-1} X^\top W_{\widehat{\beta}^{(t)}} \big(X\widehat{\beta}^{(t)} + W^{-1}(Y - p_{\widehat{\beta}^{(t)}})\big) \\ &= \big(X^\top W_{\widehat{\beta}^{(t)}} X\big)^{-1} X^\top W_{\widehat{\beta}^{(t)}} Z_{\widehat{\beta}^{(t)}}.\end{aligned}$$

Wie in Kapitel 2 gesehen, ist dies gerade die Lösung des gewichteten Kleinste-Quadrate-Problems. □

Bemerkung 4.3.

(i) Der Maximum-Likelihood-Schätzer $\widehat{\beta}$ erfüllt $\nabla_\beta \ell(\widehat{\beta}, y) = \sum_{i=1}^n x_i^\top (y - p(x_i, \widehat{\beta})) = 0$. Da der erste Koeffizient von x_i gleich 1 ist folgt $\sum_i y_i = \sum_i p(x_i, \widehat{\beta})$, d.h. die erwartete Anzahl der Beobachtungen in Klasse eins stimmt mit der beobachteten Anzahl überein.

(ii) Insbesondere zeigt dieses Lemma, dass der Maximum-Likelihood-Schätzer $\widehat{\beta}$ die Lösung eines gewichteten Kleinste-Quadrate-Problems mit Responsevektor $Z_{\widehat{\beta}} = X\widehat{\beta} + W_{\widehat{\beta}}^{-1}(Y - p_{\widehat{\beta}})$ und Gewichten $w_i = \widehat{p}_i(1-\widehat{p}_i)$, wobei beides wieder von $\widehat{\beta}$ abhängt. Die gewichteten Quadratsummen der Residuen sind dann

$$\sum_{i=1}^{n} \frac{(Y_i - \widehat{p}_i)^2}{\widehat{p}_i(1-\widehat{p}_i)}$$

und messen die Abweichung der Daten von der Modellvorhersage.

Beispiel 4.4. Wir betrachten wieder den Datensatz aus Beispiel 4.1, wobei wir eine zusätzliche Kovariable "Student" mit Werten "Ja" oder "Nein" zur Verfügung haben. Eine Logistische Regression, die nur "Student" und einen Intercept verwendet führt zu einem positiven Koeffizienten der Dummy-Variable, d.h. die Zahlungsunfähigkeitswahrscheinlichkeit ist für Studenten höher als für Nicht-Studenten. Verwenden wir alle drei Kovariablen erhalten wir jedoch einen negativen Zusammenhang! Wie kann man diesen so genannten *Konfundierungseffekt* erklären?

Die logistische Regression kann auch auf mehr als zwei Klassen ausgeweitet werden, indem wir statt der Bernoulli-Verteilung die Multinomialverteilung verwenden. Häufig wird jeden die Methode des nächsten Abschnittes dieser Variante vorgezogen. Inferenz für die logistische Regression beruht auf asymptotischen Überlegungen auf die wir in dieser Vorlesung nicht weiter eingehen werden.

4.2 Bayesklassifikation[1]

Die logistische Regression modelliert die Wahrscheinlichkeit $\mathbb{P}(Y = 1)$ unter Verwendung des Regressorvektors x für zwei Klassen 0 und 1. Verstehen wir den Kovariablenvektor als Zufallsvariable X, wird also die bedingte Wahrscheinlichkeit $\mathbb{P}(Y = 1|X = x)$ der Klasse 1 gegeben einer Kovariablenrealisierung $X = x$ modelliert. Stattdessen wird nun ein Bayesianischer Ansatz verfolgt.

Gegeben sei das zufällige Paar $\big(X, Y\big)$, welches Werte in $\mathbb{R}^d \times \big\{1, \dots, K\big\}$ annimmt. Hierbei bezeichnet Y die *Klassifizierung* von X. Das heißt, dass die Verteilung $\mathbb{P}_X$ von X durch die bedingte Verteilung $\mathbb{P}_{X|Y}$ festgelegt wird. Dieser Zusammenhang wird später durch die Bayesformel genauer erläutert. Wie in der logistischen Klassifizierung möchten wir einen *Klassifikator* $\mathcal{C}$ konstruieren, der einer Realisierung $X = x$ eine Klasse $\mathcal{C}(x) \in \big\{1, \dots, K\big\}$ zuordnet. Formal haben wir es also mit einer *deterministischen* Abbildung

$$\mathcal{C} \colon \mathbb{R}^d \mapsto \big\{1, \dots, K\big\}$$

zu tun. In der Praxis wird $\mathcal{C}$ normalerweise über ein *Trainingsample* $\mathcal{X}_n = \big\{(x_1, y_1), \dots, (x_n, y_n)\big\}$ konstruiert, worauf wir später genauer eingehen werden. Ein wesentliche Frage bezüglich eines Klassifikators ist seine Qualität hinsichtlich einer korrekten Klassifizierung. Diese können wir über die Wahrscheinlichkeit einer fehlerhaften Klassifizierung $\mathbb{P}\big(\mathcal{C}(X) \neq Y\big)$ beschreiben, und bezeichnen sie allgemein mit

$$\mathcal{R}\big(\mathcal{C}\big) = \mathbb{P}\big(\mathcal{C}(X) \neq Y\big).$$

$\mathcal{R}(C)$ entspricht also dem 0-1-Risiko. Ein optimaler Klassifizierer $\mathcal{C}^{\text{opt}}$ wäre daher (gegeben Messbarkeit)

$$\mathcal{C}^{\text{opt}} = \arg\min_{\mathcal{C}} \mathcal{R}\big(\mathcal{C}\big).$$

[1] Vielen Dank an Moritz Jirak für die Ausarbeitung dieses Abschnitts

Wie können wir diesen konstruieren? Zunächst ist es günstig das Risiko $\mathcal{R}(\mathcal{C})$ umzuformen:

$$\begin{aligned}\mathcal{R}(\mathcal{C}) &= \mathbb{P}\big(\mathcal{C}(X) \neq Y\big) = \int \mathbb{P}\big(\mathcal{C}(x) \neq Y \big| X = x\big)\mathbb{P}_X(\mathrm{d}x) \\ &= \int \big(1 - \mathbb{P}(\mathcal{C}(x) = Y \big| X = x)\big)\mathbb{P}_X(\mathrm{d}x) \\ &= 1 - \int \mathbb{P}\big(\mathcal{C}(x) = Y \big| X = x\big)\mathbb{P}_X(\mathrm{d}x).\end{aligned}$$

Wir sehen also, dass $\mathcal{R}(\mathcal{C})$ klein ist, wenn die bedingte Wahrscheinlichkeit $\mathbb{P}\big(\mathcal{C}(x) = Y \big| X = x\big)$ möglichst groß ist. Bedingen auf Y liefert weiter

$$\mathcal{R}(\mathcal{C}) = 1 - \int \Big(\sum_{k=1}^{K} \mathbb{P}\big(\mathcal{C}(x) = k \big| Y = k,\, X = x\big)\mathbb{P}\big(Y = k \big| X = x\big)\Big)\mathbb{P}_X(\mathrm{d}x).$$

Nun benutzen wir die Tatsache, dass $\mathcal{C}$ deterministisch ist (eine leichte Verallgemeinerung ist Unabhängigkeit von Y). Dadurch erhalten wir

$$\mathbb{P}\big(\mathcal{C}(x) = k \big| Y = k,\, X = x\big) = \mathbb{P}\big(\mathcal{C}(x) = k\big) = \mathbb{1}_{\{\mathcal{C}(x)=k\}},$$

und somit

$$\mathcal{R}(\mathcal{C}) = 1 - \int \Big(\sum_{k=1}^{K} \mathbb{1}_{\{\mathcal{C}(x)=k\}}\mathbb{P}\big(Y = k \big| X = x\big)\Big)\mathbb{P}_X(\mathrm{d}x).$$

Wir haben es nun mit einer überraschend einfachen Optimierung zu tun. Um $\mathcal{R}(\mathcal{C})$ zu minimieren, genügt es den Ausdruck

$$\mathcal{A}(x) := \sum_{k=1}^{K} \alpha(k,x)\mathbb{P}\big(Y = k \big| X = x\big) \quad \text{mit} \quad \alpha(k,x) = \mathbb{1}_{\{\mathcal{C}(x)=k\}},$$

für jedes $x \in \mathbb{R}^d$ zu maximieren. Es gilt nun allerdings $\alpha(k,x) \in \{0,1\}$. Die Größe $\mathcal{A}(x)$ ist folglich genau dann maximal, wenn wir das meiste Gewicht (und somit $\alpha(k,x) = 1$) auf

$$\max_{1 \leq k \leq K} \mathbb{P}\big(Y = k \big| X = x\big)$$

legen (dies kann leicht bewiesen werden). $\mathcal{A}(x)$ ist also genau dann maximal, wenn wir für jedes $x \in \mathbb{R}^d$

$$\alpha(k,x) = \begin{cases} 1, & \text{für } k = k^*, \\ 0, & \text{sonst,} \end{cases} \qquad \text{wobei} \qquad k^* = \arg\max_{k=1,\dots,K} \mathbb{P}\big(Y = k \big| X = x\big).$$

Dieses k^* liefert uns automatisch die optimale Klassifikation:

$$\mathcal{C}^{\mathrm{opt}}\big(x\big) = \arg\max_{k=1,\dots,K} \mathbb{P}\big(Y = k \big| X = x\big). \tag{4.1}$$

Theorem 4.5. *Der deterministische Klassifikator, welcher das Risiko $\mathcal{R}(\mathcal{C})$ minimiert, ist durch die Klassifikation in (4.1) gegeben und wird Bayesklassifikator genannt.*

Bemerkung 4.6.

(i) Die Optimalität der Bayesklassifikation hängt essentiell mit der Definition des Risikos $\mathcal{R}(\mathcal{C})$ zusammen.

(ii) Falls $K = 2$, erhalten wir eine sehr einfache Klassifizierungsregel: Wenn $\mathbb{P}\big(Y = 1 \big| X = x\big) \geq 1/2$, dann wählen wir Klasse $k = 1$, ansonsten Klasse $k = 2$.

Viele Klassifikationsalgorithmen versuchen die Bayesklassifikation zu imitieren. Ein allgemeiner Zugang ist die bedingte Wahrscheinlichkeit $\mathbb{P}\big(Y = k \big| X = x\big)$ anhand eines Trainingsamples zu schätzen. Ein bekannter Repräsentant dieser Art ist das KNN-Verfahren (K-nearest neighbour, Übung ☑). Ein anderes ist die lineare Diskrimimanzanalyse, die im nächsten Abschnitt behandelt wird.

4.3 Lineare Diskriminanzanalyse[2]

Das Problem der Bayesklassifikation ist, dass es nicht so einfach ist, gute Schätzer für die bedingten Wahrscheinlichkeiten $\mathbb{P}\big(Y = k\big|X = x\big)$ zu konstruieren. Allerdings kann die Bayesformel hier helfen. Wir modellieren die Verteilung von X für jede Klasse $k \in \{1, \dots, K\}$ mit $K \geqslant 2$ (also gegeben Y) durch eine Dichte

$$f_k(x) = \mathbb{P}(X = \mathrm{d}x|Y = k)$$

und wählen a-priori-Wahrscheinlichkeiten der Klassen $\pi_k = \mathbb{P}(Y = k) \in [0, 1]$ für $k = 1, \dots, K$ mit $\sum_k \pi_k = 1$. Die Bayesformel liefert dann die a-posteriori-Zähldichte von Y

$$p_k(x) = \mathbb{P}\big(Y = k\big|X = \mathrm{d}x\big) = \frac{\mathbb{P}\big(X = \mathrm{d}x\big|Y = k\big)\mathbb{P}\big(Y = k\big)}{\mathbb{P}\big(X = x\big)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}.$$

Die Idee der *linearen Diskrimimanzanalyse (LDA)* ist nun, $f_k(x)$ als Gaußdichte (univariat) zu modellieren, also

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\!\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right),$$

wobei μ_k und σ_k der Mittelwert und die Varianz der k-ten Klasse sind. Der Einfachheit halber sei $\sigma_1^2 = \ldots = \sigma_K^2$ in der folgenden Diskussion. Dann erhalten wir

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\big(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\big)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\big(-\frac{(x-\mu_l)^2}{2\sigma_l^2}\big)}. \tag{4.2}$$

Durch umformen erhalten wir, dass $p_k(x)$ genau dann maximal ist, wenn $\delta_k(x)$ maximal ist, gegeben durch

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log\big(\pi_k\big). \tag{4.3}$$

Tatsächlich sind die Werte π_k, μ_k und σ^2 aber unbekannt, und müssen geschätzt werden.

Methode 10: Lineare Diskriminanzanalyse. Wir definieren

$$\widehat{\pi}_k = \frac{n_k}{n}, \quad \widehat{\mu}_k = \frac{1}{|n_k|}\sum_{j:\, y_j = k} x_j \quad \text{und} \quad \widehat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^K \sum_{j:y_j=k} \big(x_j - \widehat{\mu}_k\big)^2,$$

wobei n die Gesamtanzahl des Trainingssamples $\mathcal{X}_n$ und n_k die Anzahl des Trainingssamples in der k-ten Klasse sind. Dann ist der Klassifizierer gegeben durch

$$\mathcal{C}(x) = \arg\max_{k=1,\dots,K} \widehat{\delta}_k(x) \quad \text{mit} \quad \widehat{\delta}_k(x) = x\frac{\widehat{\mu}_k}{\widehat{\sigma}^2} - \frac{\widehat{\mu}_k^2}{2\widehat{\sigma}^2} + \log\big(\widehat{\pi}_k\big).$$

Im multivariaten Fall erhalten wir analog die Klassifizierungsregel

$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log\big(\pi_k\big), \tag{4.4}$$

wobei Σ die d-dimensionale Kovarianzmatrix von $X \in \mathbb{R}^d$ ist, und $\mu_k \in \mathbb{R}^d$ der Vektor der komponentenweisen Erwartungswerte. Dabei können Σ, $\mu_1, \dots, \mu_K$ sowie $\pi_1, \dots, \pi_K$ wieder über Plug-in bzw. relative Häufigkeiten geschätzt werden.

Bemerkung 4.7. Eine weitere Verallgemeinerung stellt die *quadratische Diskriminanzanalyse* (QDA) dar, wo jede Klasse k eigene, im allgemeinen unterschiedliche Kovarianzmatrizen Σ_k besitzen. Dies führt zu einer quadratischen Klassifizierungsregel.

[2] Vielen Dank an Moritz Jirak für die Ausarbeitung dieses Abschnitts

Obwohl die Motivation für die logistische Klassifikation und LDA unterschiedlich ist, gibt es einen engen Zusammenhang. Betrachten wir den Fall $K = 2$. Dann gilt $p_2(x) = 1 - p_1(x)$ und eine kurze Rechnung ergibt für die LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_1 + c_2 x,$$

wobei die Konstanten c_1, c_2 von μ_1, μ_2 und σ^2 abhängen. Im Fall der logistischen Klassifizierung haben wir:

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = \beta_1 + \beta_2 x.$$

Der Unterschied liegt also nur in der Art und Weise, wie die Konstanten geschätzt werden! In der Praxis führt das oft zu sehr ähnlichen Ergebnissen, aber nicht immer.

5 Ausblick

Im letzten Teil der Vorlesung werden (voraussichtlich) noch folgende Themen behandelt:

(i) Modellwahl und statistisches Lernen

  (a) Variablenselektion (C_p, AIC, BIC, R^2)
  (b) Lasso
  (c) Dimensionsreduktion / Hauptkompontenanalyse

(ii) Resampling

  (a) Bootstrap
  (b) Kreuzvalidierung

Literatur

Agresti, A. and Finlay, B. (1997). *Statistical Methods for Social Sciences.* Prentice Hall.

Fahrmeir, L., Kneib, T., and Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen.* Springer, Berlin.

Georgii, H.-O. (2007). *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik.* de Gruyter, Berlin.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning (with Applications in R).* Springer, New York.

Le, C. T. (2003). *Introductory biostatistics.* John Wiley & Sons.

Witting, H. (1985). *Mathematische Statistik I.* Teubner.