Methoden der Statistik

Bernhard Stankewitz
bernhard.stankewitz@posteo.de

11. Juni 2018

Zusammenfassung

Studentische Mitschrift der Vorlesung.

Inhaltsverzeichnis

1 Grundbegriffe

1.1 Schätzen und Konfidenz

Aus einer Zeitungsnotiz vom 22.9.2017 stammt die folgende Wahlprognose:

CDU	SPD	Grüne	FDP	Linke	AfD
36 %	22 %	8 %	10.5 %	9.5 %	10 %

Notiz: Die Unsicherheit beträgt 2 Prozentpunkte.

Tabelle 1: Wahlprognose vom 22.9.2017

Was ist grob geschehen? n = 1074 Wahlberechtigte wurden zufällig ausgewählt und davon hat ein entsprechender Anteil sich für die jeweilige Partei entschieden. Dies ergibt die entsprechenden Schätzwerte, was so keiner weitere Mathematik bedarf. Wichtig ist aber die angegebene Unsicherheit von 2 Prozentpunkten, um potentielle Koalitionen oder Scheitern an 5-Hürde einschätzen zu können. („Uncertainty quantification", statistische Inferenz). Um Aussagen über die Unsicherheit zu machen, benötigen wir ein Modell. Der Einfachheit halber betrachten wir nur die AfD.

Modell: Es gebe N Wahlberechtigte und davon N_{AfD} AfD-Wähler. In der Stichprobe von n Befragten erhalten wir X AfD-Wähler. Bevor die Befragung startet ist die Anzahl X zufällig mit $X \sim \text{Bin}(n, p)$, wobei $p = N_{\text{AfD}}/N$ unbekannt und zu schätzen ist. Unser Schätzer lautet $\widehat{p} = X/n$. Allgemeiner definieren wir:

Definition 1.1 (Statistisches Modell)**.** Sei $(\mathbb{P}_\theta)_{\theta\in\Theta}$ eine Familie von Wahrscheinlichkeitsmaßen auf einem Messraum $(\mathfrak{X},\mathscr{F})$. Dann heißt $(\mathfrak{X},\mathscr{F},(\mathbb{P})_{\theta\in\Theta})$ ein *statistisches Modell*. $\mathfrak{X}$ heißt *Stichprobenraum* und $\emptyset\neq\Theta$ *Parametermenge*. Ist $\mathscr{F}_\Theta$ eine σ-Algebra auf Θ, so heißt jede $(\mathscr{F},\mathscr{F}_\Theta)$-messbare Funktion $\widehat{\theta}:\mathfrak{X}\to\Theta$ *Schätzer* von Θ.

Im Fall der Wahlprognose erhalten wir

$$(\mathfrak{X},\mathscr{F})=(\{0,\dots,n\},\mathscr{P}(\{0,\dots n\})),\quad (\Theta,\mathscr{F}_\Theta)=([0,1],\mathscr{B}_{[0,1]})\quad\text{und}\quad \widehat{\theta}(k)=\widehat{p}=\frac{k}{n}$$

für $k\in\mathfrak{X}$. Wir sind jetzt interessiert an Eigenschaften von $\widehat{p}$.

Definition 1.2 (Erwartungstreue)**.** Sei $(\mathfrak{X},\mathscr{F},(\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Ist $\Theta\subset\mathbb{R}^p$ borelsch und $\mathscr{F}_\Theta=\mathscr{B}_\Theta$, so heißt ein Schätzer $\widehat{\theta}$ erwartungstreu, falls

$$\mathbb{E}_\theta\widehat{\theta}:=\int\widehat{\theta}\,d\mathbb{P}_\theta=\theta\qquad\text{für alle }\theta\in\Theta.$$

In unserem Modell ergibt sich

$$\mathbb{E}_p\widehat{p}=\sum_{k=0}^{n}\frac{k}{n}\binom{n}{k}p^k(1-p)^{n-k}=\frac{np}{n}=p\qquad\text{für alle }p\in[0,1],$$

d.h. $\widehat{\theta}$ ist erwartungstreu. Erwartungstreue Schätzer besitzen keinen systematischen Fehler, d.h. sie unter oder überschätzen den Parameter nicht im Mittel, anders als ein nichtgeeichtes Messinstrument.

Bei großem Stichprobenumfang n ist eine erstrebenswerte Eigenschaft die asymptotische Konsistenz.

Definition 1.3 (Konsistenz)**.** Ist $\Theta\subset\mathbb{R}^p$ borelsch, $\mathscr{F}_\Theta=\mathscr{B}_\Theta$ und ist $(\mathfrak{X}_n,\mathscr{F}_n,(\mathbb{P}_\theta^n)_{\theta\in\Theta})_{n\in\mathbb{N}}$ eine Folge von statistischen Modellen, so heißt eine Folge $(\widehat{\theta}_n)_{n\in\mathbb{N}}$ von Schätzern auf den jweiligen Modellen konsistent, falls

$$\widehat{\theta}_n\overset{\mathbb{P}_\theta^n}{\to}\theta:\Leftrightarrow\mathbb{P}_\theta^n\{|\widehat{\theta}_n-\theta|\geq\varepsilon\}\to 0\qquad\text{für alle }\theta\in\Theta.$$

Im Fall der Wahlprognose erhalten wir

$$\mathfrak{X}_n=\{0,\dots n\},\quad \mathscr{F}_n=\mathscr{P}(\{0,\dots,n\}),\quad \mathbb{P}_p^n=\mathrm{Bin}(n,p),\quad \widehat{\theta}_n=\widehat{p}_n=\frac{X}{n}.$$

Nach der Chebyshev-Ungleichung erhalten wir für unseren Schätzer dann wegen

$$\mathbb{P}^n\{|\widehat{p}-p|\geq\varepsilon\}\leq\frac{\mathrm{Var}_p(\widehat{p}_n)}{\varepsilon^2}=\frac{np(1-p)}{n^2\varepsilon}\overset{n\to\infty}{\to}0.$$

Konsistenz. Konsistenz ist eine notwendige Bedingung dafür, dass ein Schätzer überhaupt als sinnvoll zu betrachten ist. Wenn wir ein Gesetz der großen Zahl zum Nachweis der Konsistenz verwenden möchten, so müssen wir uns hier auf eine unendlich lange Bernoulli-schema stützen, d.h. das Modell

$$\big(\{0,1\}^{\mathbb{N}},\mathscr{P}(\{0,1\})^{\otimes\mathbb{N}},(\mathrm{Bin}(1,p)_p^{\otimes\mathbb{N}})\big)$$

Für $x\in\{0,1\}^{\mathbb{N}}$ betrachte dann $X_n(x)=\sum_{i=1}^n x_i$. Unter $\mathbb{P}_p:=\mathrm{Bin}(1,p)^{\otimes\mathbb{N}}$ ist X_n eine $\mathrm{Bin}(n,p)$-verteilte Zufallsvariable. Diese erzeugt ein eigenes Modell nämlich

$$\big(\{0,\dots n\},\mathscr{P}(\{0,\dots n\}),(\mathbb{P}_p^{X_n})_{p\in[0,1]}.\big)$$

Das führt auf die folgende Definition.

Definition 1.4. Ist $(\mathfrak{X},\mathscr{F},(\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell und ist $(\tilde{\mathfrak{X}},\tilde{\mathscr{F}})$ ein weiterer Messraum, so heißt jede $(\mathscr{F},\tilde{\mathscr{F}})$-messbare Funktion T *Statistik* . Das von der Statistik T generierte statistische Modell ist $(\tilde{\mathfrak{X}},\tilde{\mathscr{F}},(\mathbb{P}_\theta^T)_{\theta\in\Theta})$ Bei Angabe einer Statistik wird das Ausgangsmodell oft nicht explizit erwähnt.

Bei der Wahlprognose reicht es zu sagen, dass eine $\mathrm{Bin}(n,p)$-verteilte Statistik X mit $p\in[0,1]$ unbekannt beobachtet wird. Unter $\mathbb{P}_p$ gilt nach dem schwachen Gesetz der großen Zahl, dass $\widehat{p}_n\overset{\mathbb{P}_p}{\to}\mathbb{E}_p\widehat{p}_n=p$, was $\widehat{p}_n\overset{\mathbb{P}_p^{\widehat{p}_n}}{\to}p$ impliziert.

Wie können wir die Genauigkeit des Schätzers $\widehat{p}$ bzw. seine statistische Unsicherheit messen? Ein Standardmaß ist die Varianz bzw. Standardabweichung. Wir erhalten

$$\mathrm{Var}_p(\widehat{p})=\mathrm{Var}_p\Big(\frac{X}{n}\Big)=\frac{np(1-p)}{n^2}=\frac{p(1-p)}{n}.$$

Beachte: Die Varianz hängt von wahren Parameter ab. Um auf der sicheren Seite zu sein, kann man daher nur angeben, dass die Standardabweichung maximal $1/(2\sqrt{n})$ beträgt. Um eine Standardabweichung von maximal 2 Prozentpunkten zu gewährleisten benötigen wir mindestens den Stichprobenumfang n mit $1/(2\sqrt{n})=0.002$, dh $n=625$. Ein allgemeineres Gütemaß (auch für nicht erwartungstreue Schätzer) ist der mittlere quadratische Fehler (MSE):

Definition 1.5 (MSE)**.** Für $\Theta \subset \mathbb{R}^p$ borelsch sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Ist $\widehat{\theta}$ ein Schätzer von θ und gilt $\widehat{\theta} \in L^2(\mathbb{P}_\theta)$ für alle $\theta \in \Theta$, so ist der *mittlere quadratische Fehler* (MSE) gegeben durch

$$R(\theta, \widehat{\theta}) := \mathbb{E}_\theta|\widehat{\theta} - \theta|^2, \qquad \theta \in \Theta.$$

Satz 1.6 (Bias-Varianz-Zerlegung)**.** *In der Situation von Definition gilt die folgende Bias-Varianz-Zerlegung*

$$\mathbb{E}_\theta|\widehat{\theta} - \theta|^2 = |\mathbb{E}_\theta\widehat{\theta} - \theta|^2 + \mathbb{E}|\widehat{\theta} - \mathbb{E}_\theta\widehat{\theta}|^2, \qquad \textit{für alle } \theta \in \Theta.$$

Ist $\widehat{\theta}$ erwartungstreu, gilt also $R(\theta, \widehat{\theta}) = Var_\theta(\widehat{\theta})$.

Beweis. Nachrechnen. □

Es stellt sich die Frage, ob es im Fall unserer Wahlprognose einen anderen Schätzer als $\widehat{p}$ mit kleinerem MSE gibt. Wir betrachten den trivialer Schätzer $\tilde{p} := 0.1$ unabhängig von der Beobachtung. Dann gilt $\mathrm{Var}_p(\tilde{p}) = 0$ für alle $p \in [0, 1]$. Es folgt also

$$R(p, \tilde{p}) = \mathrm{Bias}^2 = (0.1 - p)^2 \qquad \text{für alle } p \in [0, 1]$$

Im Vergleicht dazu gilt

$$R(p, \widehat{p}) = \mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n} \qquad \text{für alle } p \in [0, 1]$$

Falls das wahre p in einer Umbegung von 0.1 liegt, ist also $\tilde{p}$ bezüglich des MSE besser als $\widehat{p}$. Beachte allerdings, dass $\tilde{p}$ weder erwartungstreu noch konsistent ist.

Es gibt zwei Standardansätze um Schätzer bezüglich des MSE zu vergleichen. Den worst-case und den average-case.

Definition 1.7 (Minimax- und Bayes-Schätzer)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell.

(a) Ein Schätzer $\widehat{\theta}$ von θ heißt *minimax* falls für jeden anderen Schätzer $\tilde{\theta}$ von θ gilt

$$\sup_{\theta\in\Theta} R(\theta, \widehat{\theta}) \le \sup_{\theta\in\Theta} R(\theta, \tilde{\theta}), \qquad \text{d.h.} \qquad \sup_{\theta\in\Theta} R(\theta, \widehat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} R(\theta, \tilde{\theta}),$$

wobei das Infimum über alle Schätzer $\tilde{\theta}$ gebildet wird.

(b) Die Parametermenge Θ sei mit einer σ-Algebra $\mathscr{F}_\Theta$ versehen. Dann heißt ein Wahrscheinlichkeitsmaß π auf $\mathscr{F}_\Theta$ a-priori Verteilung des Parameters. Ein Schätzer $\widehat{\theta}_\pi$ heißt dann Bayes-optimal (bezüglich π und MSE), falls für alle anderen Schätzer $\tilde{\theta}$ gilt

$$\int_\Theta R(\theta, \widehat{\theta})\,\pi(d\theta) \le \int_\Theta R(\theta, \tilde{\theta})\,\pi(d\theta), \qquad \text{d.h.} \qquad \int_\Theta R(\theta, \widehat{\theta})\,\pi(d\theta) = \inf_{\tilde{\theta}} \int_\Theta R(\theta, \tilde{\theta})\,\pi(d\theta),$$

wobei das Infimum wieder über alle Schätzer $\tilde{\theta}$ gebildet wird.

Beispiel 1.8 (Minimax und Bayes-Ansatz für $\mathrm{Bin}(n, p)$).

(a) $\widehat{p} := X/n$ ist nicht minimax schätzer von p. Betrachte

$$R(p, \widehat{p}) = \mathbb{E}(\widehat{p} - p)^2 = \mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}$$

Dieser Ausdruck wird maximal für $p = 1/2$. Wenn wir versuchen in diesem „worst case“ besonders gut zu sein können wir einen Schätzer mit geringerem Minimaxrisko erreichen. Für

$$\tilde{p} := \frac{n}{n+2}\frac{X}{n} + \frac{2}{n+2}\frac{1}{2}$$

Für diesen Schätzer erhalten wir

$$\mathbb{E}(\tilde{p} - p)^2 = \mathrm{Bias}(\tilde{p})^2 + \mathrm{Var}(\tilde{p}) = \Big(\frac{-2p+1}{n+2}\Big)^2 + \Big(\frac{n}{n+2}\Big)^2\frac{p(1-p)}{n}$$

Ableiten und 0 setzen liefert, dass auch dieser Ausdruck sein Maximum in $p = 1/2$ annimmt. Dort besitzt es aber einen etwas kleineren Wert als der Schätzer $\widehat{p}$.

(b) Für $X \sim \mathrm{Bin}(n, p)$ mit $p \in [0, 1]$ a priori $\mathrm{Unif}([0, 1])$-verteilt bestimmen wir den Bayes-Schätzer $\widehat{p}_\pi$. Für jeden Schätzer, $\tilde{p}$ von p gilt

$$\begin{aligned}\int_0^1 \mathbb{E}_p(\tilde{p} - p)^2\,\pi(dp) &= \int_0^1 \sum_{k=0}^n (\tilde{p}(k) - p)^2 \binom{n}{k} p^k(1-p)^{n-k}\,dp \\ &= \sum_{k=0}^n \binom{n}{k} \int_0^1 p^k(1-p)^{n-k}(\tilde{p}(k) - p)^2\,dp.\end{aligned}$$

Minimiere dazu jedes Integral in $\tilde{p}(k)$. Als Bedingung erster Ordnung erhalten wir

$$\int_0^1 p^k(1-p)^{n-k} 2(\tilde{p}(k)-p)\,dp = 0,$$

was impliziert, dass

$$\tilde{p}(k) = \frac{\int_0^1 p^{k+1}(1-p)^{n-k}\,dp}{\int_0^1 p^k(1-p)^{n-k}\,dp} = \frac{B(k+2,n-k+1)}{B(k+1,n-k+1)} = \frac{k+1}{n+2}$$

Also ist $\tilde{p} = \frac{X+1}{n+2}$ Bayes-optimal bezüglich $\pi = \text{Unif}([0,1])$ und MSE.
Die Darstellung

$$\widehat{p}_\pi = \frac{n}{n+2}\frac{X}{n} + \frac{2}{n+2}\cdot\frac{1}{2}$$

des Bayes-Schätzers zeigt, dass er eine Konvexkombination des erwartungstreuen Schätzers $\frac{X}{n}$ und des Schätzers $1/2$ ohne Varianz ist. Der Schätzer X/n wird also in Richtung $1/2$ gezogen um die Varianz, die dort maximal ist, zu verringern. Bayes-Schätzer sind quasi nie erwartungstreu und habe oft die Eigenschaft durch Verzerrung die Varianz zu verringern (shrinkage estimator, ridge regression). Für große n ist die Korrektur zu vernachlässig.

Da π ein W-maß auf $(\Theta, \mathscr{F}_\Theta)$ ist, sowie jedes $\theta \in \Theta$ ein W-maß auf $(\mathfrak{X}, \mathscr{F})$ definiert, liegt es nahe folgendes Wahrscheinlichkeitsmaß auf dem Produktraum $(\Theta \times \mathfrak{X}, \mathscr{F}_\Theta \otimes \mathscr{F})$ die gemeinsame Verteilung von Parameter und Beobachtung zu definieren.

$$\mathbb{P}_\pi(A \times B) = \int_A \mathbb{P}_\theta(B)\,\pi(d\theta), \qquad A \in \mathscr{F}_\Theta, B \in \mathscr{F}.$$

Dazu wird benötigt, dass $\theta \mapsto \mathbb{P}_\theta(B)$ für alle B messbar ist, d.h. $\mathbb{P}_\theta(B)$ ist Markovkern.

Satz 1.9 (A posteriori Verteilung)**.** *Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell, wobei $(\mathfrak{X}, \mathscr{F})$ und $(\Theta, \mathscr{F}_\Theta)$ beides polnische Räume mit ihren Borel-Sigma-Algebren sind. Weiter sei π eine a priori Verteilung auf $(\Theta, \mathscr{F}_\Theta)$ und $(\theta, B) \mapsto \mathbb{P}_\theta(B)$ sei ein Markov-Kern, wobei*

$$\pi = f^\Theta \nu \qquad \textit{und} \qquad \mathbb{P}_\theta = f^{X|\Theta}(.|\theta)\mu \qquad \textit{für alle } \theta \in \Theta,$$

mit σ-endlichen Maßen μ und ν. Falls $f^{X|\Theta}(.|.) : \mathfrak{X} \times \Theta \to [0,\infty]$ messbar ist bezüglich $\mathscr{F} \otimes \mathscr{F}_\Theta$, dann ist die a posteriori Verteilung von Θ gegeben der Beobachtung X gegeben durch

$$\mathbb{P}^{\Theta|X=x} = f^{\Theta|X}(.|x)\nu \qquad \textit{mit} \qquad f^{\Theta|X}(\theta|x) = \frac{f^{X|\Theta}(x|\theta)f^\Theta(\theta)}{\int f^{X|\Theta}(x|\theta')f^\Theta(\theta')\,\nu(d\theta')}, \quad (x,\theta) \in \mathfrak{X}\times\Theta.$$

Beweis. **Schritt 1: Konsistenz.** Bemerke zunächst, dass die Notation $f^{X|\Theta}$ konsistent ist mit der Notation der regulären bedingten Erwartung, da

$$\begin{aligned}\mathbb{P}_\pi\{X \in B_X, \Theta \in B_\Theta\} = \mathbb{P}_\pi\{B_X \times B_\Theta\} &= \int \mathbf{1}_{B_\Theta}(\theta)\mathbb{P}_\theta(B_X)\,\pi(d\theta) \\ &= \int_{B_\Theta} \mathbb{P}_\theta(B_X)\,\mathbb{P}^\Theta(d\theta) = \int_{\{\Theta\in B_\Theta\}} \mathbb{P}_\Theta(B_X)\,d\mathbb{P}\end{aligned}$$

für beliebige $B_X \in \mathscr{F}$ und $B_\Theta \in \mathscr{F}_\Theta$.

Schritt 2: Definitorische Eigenschaft der bedingten Erwartung. Aus dem Satz von Tonelli folgt

$$\mathbb{P}_\pi(\mathfrak{X}\times\Theta) = \iint \mathbb{P}_\theta(dx)\,\pi(d\theta) = \iint f^{X|\Theta}(x|\theta)f^\Theta(\theta)\,\mu(dx)\nu(d\theta).$$

Deswegen ist der Nenner der a posteriori Dichte μ-f.ü. wohldefiniert, $\mathscr{F}$-messbar und endlich. Analog erhalten wir die Messbarkeit von $x \mapsto (f^{\Theta|X}(.|x)\nu)(B_\Theta)$ für $B_\Theta \in \mathscr{F}^\Theta$. Damit folgt

$$\begin{aligned}\int_{B_X} (f^{\Theta|X}(.|x)\nu)(B_\Theta)\,\mathbb{P}_\pi^X(dx) &= \int_{B_X}\int_{B_\Theta} f^{\Theta|X}(\theta|x)\,\nu(d\theta)\,\mathbb{P}_\pi^X(dx) \\ &= \int_{B_X\times\Theta}\int_{B_\Theta} f^{\Theta|X}(\theta|x)\,\nu(d\theta)\,\mathbb{P}_{\theta'}(dx)\pi(d\theta') \\ &= \int_{B_X\times\Theta}\int_{B_\Theta} f^{\Theta|X}(\theta|x)\,\nu(d\theta)f^{X|\Theta}(x|\theta')f^\Theta(\theta')\,\mu(dx)\nu(d\theta') \\ &= \int_{B_X}\int_{B_\Theta} f^{X|\Theta}(x|\theta)f^\Theta(\theta)\,\nu(d\theta)\mu(dx) = \mathbb{P}_\pi(B_X\times B_\Theta).\end{aligned}$$

□

Beispiel 1.10 (A posteriori Verteilung bei der Wahlprognose). Im Fall der Wahlprognose erhalten wir $\mathbb{P}_p = \mathrm{Bin}(n,p)$, $p \sim \mathrm{Unif}([0,1]) = \pi$. Daher folgt $\pi = \mathbf{1}_{[0,1]}\lambda$ und $\mathbb{P}_p$ hat die Dichte $\binom{n}{k}p^k(1-p)^{n-k}$ bezüglich des Zählmaßes μ auf $\{0,1,\dots n\}$. Die a posteriori Dichte ergibt sich also als

$$f(p|k) \propto \binom{n}{k} p^k(1-p)^{n-k}\mathbf{1}_{[0,1]}(p) \propto p^k(1-p)^{n-k}\mathbf{1}_{[0,1]}(p).$$

Daraus folgt $f(p|k)$ ist die Dichte der $B(k+1, n-k+1)$-Verteilung. Vergleiche dazu auch die Übung.

Es stellt sich die Frage, ob es auch eine andere Methode gibt, die ohne die subjektive Wahl einer a priori Verteilung auskommt und trotzdem sinnvoll Schätzer liefert? Im Beispiel war $X \sim \mathrm{Bin}(n,p)$ mit $p \in [0,1]$ unbekannt. Als Ergebnis erhalte ich einen Wert $k \in \{0,\dots,n\}$. Eine sinnvolle Möglichkeit ist $p \in [0,1]$ so zu bestimmen, dass $\mathbb{P}_p\{X = k\}$ maximal ist, d.h. die Zähldichte $\binom{n}{k}p^k(1-p)^{n-k}$ wird maximiert.

Definition 1.11 (ML-Schätzer)**.** Ist $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell und besitzt jedes $\mathbb{P}_\theta$ eine Dichte f_θ bzgl. eines σ-endlichen Maßes μ, so heißt die zufällige Funktion

$$L : \Theta \to \mathbb{R}, \qquad L(\theta) = L(\theta, x) = f_\theta(x)$$

Likelihoodfunktion. Ein Schätzer $\widehat{\theta} : X \to \Theta$ heißt *Maximum-Likelihood-Schätzer* (MLE), falls gilt

$$L(\widehat{\theta}(x), x) = \sup_{\theta\in\Theta} L(\theta, x) \qquad \text{für } \mu\text{-fast alle } x \in \mathfrak{X}.$$

Beispiel 1.12 (MLE für $\mathrm{Bin}(n,p)$). In dem eingehenden Beispiel ist $\mathbb{P}_p = \mathrm{Bin}(n,p)$. D.h. $f_p = \binom{n}{k}p^k(1-p)^{n-k}$ bezüglich des Zählmaßes μ. Als Likelihoodfunktion ergibt sich

$$L(p, X) = \binom{n}{X} p^X(1-p)^{n-X}, \qquad p \in [0,1].$$

Der MLE ergibt sich, indem wir die Log-Likelihoodfunktion

$$\ell(p, X) := \ln L(p, X) = \ln\binom{n}{X} + X\ln p + (n-X)\ln(1-p)$$

maximieren. Als Bedingung erster Ordnung erhalten wir

$$\partial_p \ell(p, X) = 0 + \frac{X}{p} - \frac{n-X}{1-p} = 0.$$

Damit ergibt sich X/n als MLE.

Bemerkung 1.13 (Konzeptionelles zur ML-Schätzung).

(a) In der Situation von Definition 1.11 sind die Dichten f_θ nur μ-f.ü. eindeutig und somit auch der MLE. Meist ist es z.B. kanonisch eine stetige Dichte zu nehmen, wenn das möglich ist. Beachte, dass der Schätzer $\widehat{\theta}$ vor der Beobachtung X festgelegt wird und jede μ-Nullmenge auch eine eine $\mathbb{P}_\theta$-Nullmenge ist und somit die Änderung von $\widehat{\theta}$ auf einer μ-Nullmenge $\mathbb{P}_\theta$-f.s. keine Auswirkungen hat.

(b) Insbesondere für kontinuierliche Verteilungen $\mathbb{P}_\theta$ kann der MLE schlecht sein. Oft liefert die ML-Methode aber sinnvolle und gute Schätzer, die automatisch aus dem Modell bestimmt werden können.

(c) Der MLE ist maximum a-posteriori (MAP) Schätzer im Bayes-modell mit uniformer a-priori-Verteilung.

(d) Der MLE hängt in folgendem Sinne nicht von der Wahl von μ ab: Seien L_1, L_2 Likelihoodfunktionen bezüglich μ_1, μ_2, sowie L_3 Likelyhoodfunktion bzgl. $\mu_3 = \mu_1 + \mu_2$. Nach dem Satz von Radon-Nikodym gilt dann

$$L_3(\theta, x) = \frac{d\mathbb{P}_\theta}{d\mu_3}(x) = \frac{d\mathbb{P}_\theta}{d\mu_1}(x)\frac{d\mu_1}{d\mu_3}(x) = L_1(\theta, x)\frac{d\mu_1}{d\mu_3}(x).$$

Also gilt für die ML-Schätzer, dass $\widehat{\theta}_3 = \widehat{\theta}_1$ auf $\{\frac{d\mu_1}{d\mu_3} > 0\}$. Insbesondere die Gleichheit $\mathbb{P}_\theta$-f.s. für alle $\theta \in \Theta$.

Kommen wir nun zurück zur Bestimmung der Unsicherheit eines Schätzers. Bislang habe wir nur das grobe Kriterium MSE bzw. Varianz betrachtet. Eine präzisere und natürlichere Beschreibung erfolgt durch Konfidenzmengen.

Definition 1.14 (Konfidenzmenge)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Zu gegebenen $\alpha \in (0,1)$ sei für jedes $x \in \mathfrak{X}$ durch $C_{1-\alpha}(x) \subset \Theta$ eine Teilmenge im Parameterraum definiert. Dann heißt die zufällige, d.h. datenabhängige Menge $C_{1-\alpha}$ eine $(1-\alpha)$-Konfidenzmenge, falls gilt

$$\forall\theta \in \Theta : \mathbb{P}_\theta\{\theta \in C_{1-\alpha}\} = \mathbb{P}\{x \in \mathfrak{X} : \theta \in C_{1-\alpha}(x)\} \geq 1-\alpha.$$

Bemerkung 1.15 (Konzeptionelles zu Konfidenzmengen).

(a) Implizit fordern wir $\{x \in \mathfrak{X} : \theta \in C_{1-\alpha}(x)\} \in \mathscr{F}$ für alle $\theta \in \Theta$.

(b) $C_{1-\alpha}(x) = \Theta$ ist stets eine $1-\alpha$-Konfidenzmenge. Wir streben aber möglichst kleine Konfidenzmengen an, oft unter geometrischen Zusatzvoraussetzungen, wie z.B. dass $C_{1-\alpha}(x)$ ein Intervall oder eine konvexe Menge ist.

(c) Vgl. Buch. Ist $C_{1-\alpha}(x)$ die Realisierung (nach Datenerhebung) einer $(1-\alpha)$-Konfidenzmenge, so heißt das nicht, dass θ mit Wahrscheinlichkeit $(1-\alpha)$ in $C_{1-\alpha}(x)$ liegt. Es gibt aber gar kein Wahrscheinlichkeitsmaß auf Θ in dieser Modellierung. Vielmehr ist $C_{1-\alpha}$ vor der Datenerhebung so konstruiert, dass es für jeden möglichen Parameterwert θ diesen mit Wahrscheinlichkeit $1-\alpha$ enthält.

(d) Das Bayes'sche Analogon ist ein "credible set"$C_{1-\alpha}$ wo gefordert wird

$$\int_\Theta \mathbb{P}_\theta\{\theta \in C_{1-\alpha}\}\, \pi(d\theta) \geq 1-\alpha.$$

Beispiel 1.16 (Approximatives Konfidenzintervall für $\mathrm{Bin}(n,p)$). Wir betrachten wieder die $\mathbb{P}_p = \mathrm{Bin}(n,p), p \in [0,1]$ und machen den Ansatz

$$I_{1-\alpha} = \left[\widehat{p} - a(\widehat{p}), \widehat{p} + a(\widehat{p})\right]$$

für ein $1-\alpha$-Konfidenzintervall mit $\widehat{p} = X/n$ und eventuell zufälliger Länge $2a(\widehat{p})$. Nun ist die Binomialverteilung schwer explizit zu berechnen und wir nähern sie gemäß ZGWS durch die Normalverteilung ("Normalapproximation") an. Es gilt

$$\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{d} N(0,1) \qquad \text{, d.h.} \qquad I_{1-\alpha} = \left[\widehat{p} \pm q_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]$$

ist ein asymptotische $1-\alpha$-Konfidenzintervall, wobei $q_{1-\alpha/2}$ das entsprechende Quantil der Normalverteilung ist. Da der wahre Parameter $p \in [0,1]$ nicht bekannt ist schätzen wir noch $p(1-p) \leq 1/4$ ab um ein berechenbares Intervall zu erhalten.

Auf Grund der Abschätzung ist das so erhaltene asymptotische Konfidenzintervall sehr konservativ. Statt der Abschätzung würden wir $\sqrt{p(1-p)/n}$ gerne durch $\sqrt{\widehat{p}(1-\widehat{p})/n}$ schätzen. Um das so entstehende Konfidenzintervall genau interpretieren zu können, notieren wir einige Resultate über Verteilungskonvergenz ohne Beweis.

Theorem 1.17 (Portmanteau-Theorem). *Sei (E,d) ein metrischer Raum und $\mu, (\mu_n)_{n\in\mathbb{N}} \in \mathscr{M}^b(E)$. Dann sind die folgenden Aussagen äquivalent.*

(i) $\mu_n \xrightarrow{w} \mu$;

(ii) $\int f\, d\mu_n \to \int f\, d\mu$ *für alle* $f \in Lip(E) \cap C^b(E)$;

(iii) $\int f\, d\mu_n \to \int f\, d\mu$ *für alle* $f : E \to \mathbb{R}$ *messbar und beschränkt mit* $\mu(U_f) = 0$, *wobei* U_f *die Menge der Unstetigkeitsstellen von* f *ist;*

(iv) $\mu(E) \leq \liminf \mu_n(E)$ *und* $\limsup \mu_n(F) \leq \mu(F)$ *für alle abgeschlossenen* $F \subset E$;

(v) $\limsup \mu_n(E) \leq \mu(E)$ *und* $\mu(U) \leq \liminf \mu_n(U)$ *für alle offenen* $U \subset E$;

(vi) $\mu_n(B) \to \mu(B)$ *für alle* $B \in \mathscr{B}_E$ *mit* $\mu(\partial B) = 0$;

Ist E auch lokalkompakt und polnisch, so ist dazu auch äquivalent

(vii) $\mu_n \xrightarrow{v} \mu$ *und* $\mu_n(E) \to \mu(E)$.

Corollar 1.18 (Konvergenz der Verteilungsfunktionen). *Seien $\mu, (\mu_n)_{n\in\mathbb{N}} \subset \mathscr{M}^1(\mathbb{R})$ mit zugehörigen Verteilungsfunktionen $F, (F_n)_{n\in\mathbb{N}}$. Dann sind äquivalent:*

(i) $\mu_n \xrightarrow{w} \mu$;

(ii) $F_n(x) \to F(x)$ *für alle Stetigkeitsstellen* $x \in \mathbb{R}$ *von* F.

Theorem 1.19 (Continuous mapping). *Seien $(E,d), (E',d')$ metrische Räume, $\varphi : E \to E'$ messbar und U_φ die Menge der Unstetigkeitsstellen von φ. Dann gilt für $\mu, (\mu_n)_{n\in\mathbb{N}} \subset \mathscr{M}^b(E)$ mit $\mu(U_\varphi) = 0$ und $\mu_n \xrightarrow{w} \mu$, dass $\varphi(\mu_n) \xrightarrow{w} \varphi(\mu)$.*

Lemma 1.20 (Slutzky). *Seien $X, (X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ Zufallsvariablen mit Werten in einem polnischen Raum (E,d). Falls gilt, dass $X_n \xrightarrow{d} X$ und $d(X_n, Y_n) \xrightarrow{\mathbb{P}} 0$, dann gilt auch $Y_n \xrightarrow{d} X$.*

Corollar 1.21 (Slutzky). *Sei (E,d) polnisch und $X, (X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$ Zufallsvariablen mit Werten in E. Dan gelten*

(i) $X_n \xrightarrow{\mathbb{P}} X$ *impliziert* $X_n \xrightarrow{d} X$.

(ii) Für $a \in E$ *gilt* $X_n \xrightarrow{d} a$ *impliziert* $X_n \xrightarrow{\mathbb{P}} a$.

(iii) Für $a \in E$ *gilt* $Y_n \xrightarrow{d} a, X_n \xrightarrow{d} X$ *impliziert dass* $(X_n, Y_n) \xrightarrow{d} (X, a)$.

(iv) Falls $E = \mathbb{R}^d$, *gilt in der Situation von (iii), dass* $\langle X_n, Y_n\rangle \xrightarrow{d} \langle X, a\rangle$ *und* $X_n + Y_n \xrightarrow{d} X + a$.

Beispiel 1.22 (Approximatives Konfidenzintervall für $\mathrm{Bin}(n,p)$). Mit Hilfe des Lemmas von Slutzky 1.20 und des entsprechenden Corollars erhalten wir jetzt

$$\frac{X-np}{\sqrt{n\widehat{p}(1-\widehat{p})}}\mathbf{1}_{\{\widehat{p}>0\}} = \frac{X-np}{\sqrt{np(1-p)}}\mathbf{1}_{\{\widehat{p}>0\}}\frac{p(1-p)}{\widehat{p}(1-\widehat{p})} \xrightarrow[n\to\infty]{d} N(0,1),$$

wodurch wir in dem Konfidenzintervall p durch $\widehat{p}$ ersetzen dürfen.

1.2 Hypothesentests

Motivation aus der Erkenntnistheorie: Aus grundsätzlichen Überlegungen heraus wird eine Theorie entwickelt, die empirisch überprüft werden soll. Theorien können nie empirisch verifiziert werden, nur falsifiziert. Die Wahrscheinlichkeit, dass die Theorie (die Nullhypothese) korrekt ist, aber falsifiziert wird (Fehler erster Art), sollte möglichst klein sein. Andererseits soll natürlich eine falsche Theorie nur mit geringer Wahrscheinlichkeit anhand der Daten nicht abgelehnt werden (Fehler zweiter Art).

Definition 1.23 (Test)**.** Sei $(\mathfrak{X},\mathscr{F},(\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit $\Theta=\Theta_0\uplus\Theta_1$.

(a) Eine messbar Funktion $\varphi:\mathfrak{X}\to\{0,1\}$ wird *nicht randomisierter Test* mit Niveau $\alpha\in[0,1]$ genannt, falls gilt, dass $\mathbb{P}_\theta\{\varphi=1\}\le\alpha$ für alle $\theta\in\Theta_0$.

(b) Eine messbare Funktion $\varphi:\mathfrak{X}\to[0,1]$ wird *randomisierter Test* mit Niveau $\alpha\in[0,1]$ genannt, falls gilt, dass $\mathbb{E}_\theta\varphi\le\alpha$ für alle $\theta\in\Theta_0$.

(c) Ein (randomisierter) Test φ der Hypothes $H_0:\theta\in\Theta_0$ vs. $H_1:\theta\in\Theta_1$ wird UMP (uniformly most powerful) für ein Niveau $\alpha\in[0,1]$ genannt, falls alle Tests $\tilde\varphi$ mit Niveau α erfüllen, dass $\mathbb{E}_\theta\tilde\varphi\le\mathbb{E}_\theta\varphi$ für alle $\theta\in\Theta_1$.

Bei diskreten Wahrscheinlichkeiten existiert oft kein nicht randomisierter Test, $\varphi:\mathfrak{X}\to\{0,1\}$,der das Niveau α ausschöpft, d.h. $\sup_{\theta\in\Theta_0}\mathbb{P}\{\varphi=1\}<\alpha$ für alle Tests φ mit $\sup_{\theta\in\Theta_0}\mathbb{P}\{\varphi=1\}\le\alpha$. Deswegen erlaubt man wie in Definition 1.23 (b), dass Tests Werte in $[0,1]$ annehmen. Die Interpretation ist die folgende: Für $\varphi(X)\in(0,1)$, wird ein zusätzliches unabhäniges Zufallsexperiment mit Verteilung $\mathrm{Ber}(\varphi(X))$ durchgeführt. Die Fehlerwahrscheinlichkeit erster Art ist dann gerade $\mathbb{E}_\theta\varphi$ für $\theta\in\Theta_0$. Weiterhin ist es angenehm, dass die ranomisierten Tests eine konvexe Menge bilden.

Im Fall einfacher Hypothese, d.h. $\Theta_0=\{\theta_0\}$ und $\Theta_1=\{\theta_1\}$ gibt es eine Testkonstruktion, die für gegebenes Niveau α die Fehlerwahrscheinlichkeit zweiter Art minimiert.

Definition 1.24 (Neyman-Pearson Test)**.** Sei $(\mathfrak{X},\mathscr{F},(\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit $\Theta=\{0,1\}$ und dominierendem σ-endlichem Maß μ. Weiter sei für $i=0,1$, p_i die μ-Dichte von $\mathbb{P}_i$. Jeder Test der Form

$$\varphi(x)=\begin{cases}1 & ,p_1(x)>cp_0(x),\\ \gamma(x) & ,p_1(x)=cp_0(x),\\ 0 & ,p_1(x)<cp_0(x)\end{cases}$$

mit einem *kritischen Wert* $c\in[0,\infty)$ und $\gamma(x)\in[0,1]$ heißt *Neyman-Pearson-Test* (NP-Test).

Lemma 1.25 (**Neyman-Pearson**)**.** *In der Situation von Definition 1.24 gilt*

(i) jeder NP-Test φ ist ein UMP Test für $H_0:\theta=\theta_0$ vs. $H_1:\theta=\theta_1$ zum Niveau $\alpha=\mathbb{E}_0\varphi$;

(ii) für jedes Niveau $\alpha\in(0,1)$ gibt es einen NP-Test zum Niveau α mit konstantem $\gamma\in[0,1]$;

(iii) falls φ ein UMP-Test ist zum Niveau $\alpha\in(0,1)$, ist φ μ-f.s. identisch zu einem NP-Test.

Beweis.

(i) Unter Verwendung der Konstruktion des NP-Tests erhalten wir für jeden anderen Test $\tilde\varphi$ zum Niveau $\alpha=\mathbb{E}_0\varphi$, dass

$$\mathbb{E}_1(\varphi-\tilde\varphi)=\int_{\{\varphi>\tilde\varphi\}}(\varphi-\tilde\varphi)p_1\,d\mu+\int_{\{\varphi\le\tilde\varphi\}}(\varphi-\tilde\varphi)p_1\,d\mu \tag{1.1}$$

$$\ge\int_{\{\varphi>\tilde\varphi\}}(\varphi-\tilde\varphi)cp_0\,d\mu+\int_{\{\varphi\le\tilde\varphi\}}(\varphi-\tilde\varphi)cp_0\,d\mu=c(\mathbb{E}_0\varphi-\mathbb{E}_0\tilde\varphi)\ge0. \tag{1.2}$$

(ii) Für festes $\alpha\in(0,1)$ wollen wir die Gleichung

$$\mathbb{E}_0\varphi=\mathbb{P}_0\{p_1>cp_0\}+\gamma\mathbb{P}_0\{p_1=cp_0\}=\alpha$$

nach $c\in[0,\infty)$ und $\gamma\in[0,1]$ lösen. Da $\mathbb{P}_0$-fast sicher gilt, dass $p_0>0$, können wir die Zufallsvariable p_1/p_0 betrachten. Wir definieren dann $c\in[0,\infty)$ als das $(1-\alpha)$-quantil von p_1/p_0 unter $\mathbb{P}_0$ und erhalten

$$\mathbb{E}_0\varphi=\mathbb{P}_0\{p_1>cp_0\}+\gamma\mathbb{P}_0\{p_1=cp_0\}=1-\mathbb{P}_0\{p_1\le cp_0\}+\gamma\mathbb{P}_0\{p_1=cp_0\}.$$

Die Behauptung folgt jetzt also, wenn wir

$$\gamma=\begin{cases}\frac{\mathbb{P}_0\{p_1\le cp_0\}-(1-\alpha)}{\mathbb{P}_0\{p_1=cp_0\}} & ,\mathbb{P}_0\{p_1=cp_0\}>0,\\ 0 & ,\text{sonst}\end{cases}$$

setzen.

(iii) Sei φ^* ein UMP-Test zum Niveau $\alpha \in (0,1)$ und φ der NP-Test aus (ii). Es reicht jetzt zu zeigen, dass $\varphi = \varphi^*$ auf $\{p_1 \neq cp_0\}$. Dies zeigen wir per Widerspruch. Nehem an, dass für $S := \{\varphi^* \neq \varphi\} \cap \{p_1 \neq cp_0\}$ gilt, dass $\mu(S) > 0$. Dann folgt

$$\int (\varphi - \varphi^*)(p_1 - cp_0)\, d\mu = \int_S (\varphi - \varphi^*)(p_1 - cp_0)\, d\mu > 0,$$

was impliziert, dass $\mathbb{E}_1\varphi - \mathbb{E}_1\varphi^* > c(\alpha - \mathbb{E}_0\varphi^*) \geq 0$. Widerspruch.

□

Beispiel 1.26 (Einseitiger Gauß-Test). Seien $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d. mit $\mu \in \mathbb{R}$ unbekannt und $\sigma^2 > 0$ bekannt. Wir wollen die Hypothesen $H_0 : \mu = \mu_0$ und $H_1 : \mu = \mu_1$ für gegebene Werte $\mu_0 < \mu_1$ gegeneinander testen. Wir erhalten

$$\mathfrak{X} = \mathbb{R}^n, \qquad \mathscr{F} = \mathscr{B}_{\mathbb{R}}^{\otimes n}, \qquad \mathbb{P}_0 = N(\mu_0, \sigma^2)^{\otimes n}, \qquad \mathbb{P}_1 = N(\mu_1, \sigma^2)^{\otimes n}$$

und als Dichtequotienten bezüglich des Lebesguemaßes

$$\begin{aligned}\frac{p_1}{p_0}(X) &= \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n \big((X_i - \mu_1)^2 - (X_i - \mu_0)^2\big)\Big) \\ &= \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n (\mu_1^2 - 2X_i(\mu_1 - \mu_0) - \mu_0^2)\Big).\end{aligned}$$

Auf Grund der Monotonie der Exponentialfunktion können gilt

$$\frac{p_1}{p_0} > c_\alpha \Leftrightarrow \sum_{i=1}^n X_i > c_\alpha \Leftrightarrow \overline{X}_n > c_\alpha,$$

wobei wir das geeignete $\tilde{c}_\alpha$ von Schritt zu schritt anpassen. Als NP-Test erhalten wir also

$$\varphi_\alpha(X) = \mathbf{1}_{\{\overline{X}_n > c_\alpha\}} + \gamma_\alpha \mathbf{1}_{\{\overline{X}_n = c_\alpha\}} = \mathbf{1}_{\{\overline{X}_n > c_\alpha\}}.$$

Unter der Hypothese H_0 gilt $\overline{X}_n \sim N(\mu_0, \sigma^2/n)$. Den kritischen Wert für ein Niveau α kann man dementsprechend einer Normalverteilungstabelle entnehmen.

Existiert in der Situation von Definition eine Likelihoodfunktion $(\theta, x) \mapsto L(\theta, x)$, so sind mir mit Hilfe des NP-Lemmas 1.25 in der Lage einfache Hypothesen gegeneinander zu testen. Oft wollen wir aber zusammengesetzte Hypothesen betrachten. Das führt auf die Idee, die jeweils „besten“ Parameter gegeneinander zu testen.

Definition 1.27 (LR-Test)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell mit Likelihoodfuntkion L und $\Theta = \Theta_0 \uplus \Theta_1$ eine nicht triviale Partition. Ein Test auf $H_0 : \theta \in \Theta_0$ gegen $H_1 : \theta \in \Theta_1$ der Form

$$\varphi_\alpha = \begin{cases} 1 & , \sup_{\theta_1\in\Theta_1} L(\theta_1) > c_\alpha \sup_{\theta_0\in\Theta_0} L(\theta_0), \\ \gamma_\alpha & , \sup_{\theta_1\in\Theta_1} L(\theta_1) = c_\alpha \sup_{\theta_0\in\Theta_0} L(\theta_0), \\ 0 & , \sup_{\theta_1\in\Theta_1} L(\theta_1) < c_\alpha \sup_{\theta_0\in\Theta_0} L(\theta_0) \end{cases}$$

mit geeigneten $c_\alpha \geq 0$, $\gamma_\alpha \in [0,1]$ heißt *Likelihood-Quotienten-Test*

Bemerkung 1.28 (Eigenschaften von LR-Tests).

(a) Unter Regularitätsvoraussetzungen an L und für asymptotisch große Stichprobenumfänge kann man optimale Eigenschaften von LR-Tests beweisen. Hier sehen wir es als eine konkrete Konstruktion an, deren Eigenschaften im Einzelfall nachgewiesen werden können.

(b) Ist $\widehat{\theta}_1$ ein MLE für θ_1 und $\widehat{\theta}_0$ ein MLE für $\theta \in \Theta_0$, so gilt in Definition 1.27

$$\varphi_\alpha = \begin{cases} 1 & , L(\widehat{\theta}_1) > c_\alpha L(\widehat{\theta}_0), \\ \gamma_\alpha & , L(\widehat{\theta}_1) = c_\alpha L(\widehat{\theta}_0), \\ 0 & , L(\widehat{\theta}_1) < c_\alpha L(\widehat{\theta}_0). \end{cases}$$

Beispiel 1.29 (t-Test). Betrachte das statistische Modell $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}}^{\otimes n}, (N_{\mu,\sigma^2})_{\mu\in\mathbb{R},\sigma^2>0})$, d.h. wir beobachten $X_1, \ldots, X_n \sim N_{\mu,\sigma^2}$ i.i.d., und das Testproblem $H_0 : \mu = \mu_0$ (σ beliebig) vs. $H_1 : \mu \neq \mu_0$. Welche Form hat der LR-Test? Es gilt

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2\Big)$$

Unter $H_0 : \mu = \mu_0$ ist der ML-Schätzer für (μ, σ^2) gegeben durch

$$\widehat{\mu}_0 := \mu_0 \qquad \text{und} \qquad \widehat{\sigma}_0^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2$$

Unter $H_1 : \mu \neq \mu_0$ ist der ML-Schätzer für (μ, σ^2) gegeben durch

$$\widehat{\mu}_1 := \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \qquad \text{und} \qquad \widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

(Falls $\overline{X} = \mu_0$ kein Problem durch Stetigkeit von L). Als Likelihoodquotient erhält man:

$$\begin{aligned}\frac{\sup_{\theta_1\in\Theta_1} L(\theta_1)}{\sup_{\theta_0\in\Theta_0} L(\theta_0)} &= \Big(\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_0^2}\Big)^{-n/2} \exp\Big(-\frac{1}{2\widehat{\sigma}_1^2}\sum_{i=1}^n (X_i - \widehat{\mu}_1)^2 + \frac{1}{2\widehat{\sigma}_0^2}\sum_{i=1}^n (X_i - \mu_0)^2\Big) \\ &= \Big(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_1^2}\Big)^{n/2} \exp\Big(-\frac{n}{2} + \frac{n}{2}\Big) = \Big(\frac{\frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2}{\widehat{\sigma}_1^2}\Big)^{n/2} \\ &= \Big(\frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X} + \overline{X} - \mu_0)^2}{\widehat{\sigma}_1^2}\Big)^{n/2} = \Big(1 + \frac{(\overline{X} - \mu_0)^2}{\widehat{\sigma}_1^2}\Big).\end{aligned}$$

Deswegen hat der LR-Test die Form $\varphi_\alpha = \mathbf{1}\{\frac{|\overline{X}-\mu_0|}{\widehat{\sigma}_1} \geq c_\alpha\}$ ohne Randomisierung. Dies ist der wichtige t-Test. Wir werden später sehen, dass die kritischen Werte der Tabelle der t-Verteilung entnommen werden können. Für große n lässt sich der Wert aber auch asymptotisch aus der Normalverteilungstabelle ablesen.

In der Praxis (Statistik-Software) wird bei gegebenen Daten $x \in \mathfrak{X}$ meist der sogenannte p-Wert angegeben mit der Interpretation, dass für jedes Niveau $\alpha > p(x)$ der Test H_0 abgelehnt hätte und für jedes Niveau $\alpha < p(x)$ H_0 akzeptiert hätte.

Definition 1.30 (p-Wert)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell. Für eine Statistik $T : \mathfrak{X} \to \mathbb{R}$ sei eine Familie von Tests der Hypothese $H_0 : \theta \in \Theta_0$ vom Niveau $\alpha \in (0,1)$ definiert durch

$$\varphi_\alpha(x) := \mathbf{1}_{\{T(x)\in\Gamma_\alpha\}}, \qquad x \in \mathfrak{X}$$

und Ablehnungsbereiche $\Gamma_\alpha \in \mathscr{B}_\mathbb{R}, \alpha \in (0,1)$. Das ist der p-Wert einer Realisierung $x \in \mathfrak{X}$ definiert als

$$p(x) := \inf_{\alpha : T(x)\in\Gamma_\alpha} \sup_{\theta\in\Theta_0} \mathbb{E}_\theta \varphi_\alpha,$$

d.h. der p-Wert ist das minimale Niveau zu dem H_0 bei den gegebenen Daten verworfen werden kann.

Bemerkung 1.31 (Wahrscheinlichkeit für extremere Ereignisse). Ist in der Situation von Definition 1.30 die Familie $(\varphi_\alpha)_{\alpha\in(0,1)}$ gegeben durch

$$\varphi_\alpha(x) := \mathbf{1}_{\{|T(x)|\geq c_\alpha\}}, \qquad x \in \mathfrak{X}$$

mit $\alpha \mapsto c_\alpha \geq 0$ monoton fallend. Dann ist

$$p(x) = \inf_{\alpha : T(x)\geq c_\alpha} \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\{|T(X)| \geq c_\alpha\} = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta\{|T(X)| \geq T(x)\}.$$

Schließlich notieren wir noch einen wichtigen Zusammenhang zwischen Tests und Konfidenzmengen.

Satz 1.32 (**Korrespondenzsatz**)**.** *Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $\alpha \in (0,1)$. Dann gilt*

(i) Ist C eine $1-\alpha$-Konfidenzmenge, dann ist

$$\varphi_{\theta_0}(x) := \mathbf{1}_{\{\theta_0 \notin C(x)\}} = 1 - \mathbf{1}_{\{\theta_0\in C(x)\}}$$

ein Test vom Niveau α für $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

(ii) Ist φ_{θ_0} für jedes $\theta_0 \in \Theta$ ein nicht randomisierter Test der Hypothese $H_0 : \theta = \theta_0$ zum Niveau α, so definiert

$$C(x) := \{\theta \in \Theta : \varphi_\theta(x) = 0\}, \qquad x \in \mathfrak{X}$$

eine $1-\alpha$-Konfidenzmenge. In Worten sind das gerade die Parameter, für die der Test φ_θ bei gegebenen Daten H_0 nicht verwerfen würde..

Beweis. (i) Folgt sofort.

(ii) Es gilt $\mathbb{E}_\theta\varphi_\theta \leq \alpha, \theta \in \Theta$ und damit für jedes θ

$$1 - \alpha \leq \mathbb{E}_\theta(1 - \varphi_\theta) = \mathbb{P}_\theta\{x : \varphi_\theta(x) = 0\} = \mathbb{P}_\theta\{\theta \in C(x)\}.$$

□

2 Das lineare Modell

2.1 Lineares Modell und kleinste Quadrate

Das motivierende Beispiel für das lineare Modell ist die einfache Regressionsgerade.

Beispiel 2.1 (Regressionsgerade). Wir betrachten das Modell

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots n$$

mit deterministischen $x_1, \dots, x_n$ deterministisch und bekannt, deterministischen Parametern $\beta_0, \beta_1 \in \mathbb{R}$ und Zufallsvariablen $(\varepsilon_i) \sim$ i.i.d. mit $\mathbb{E}\varepsilon_i = 0$ und $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Man bestimmt eine Regressionsgerade der Form $y = \widehat{\beta}_0 x + \widehat{\beta}_1$ mit Schätzern $\widehat{\beta}_0$, $\widehat{\beta}_1$ aus den Beobachtungen. Die Methode der kleinsten Quadrate liefert

$$(\widehat{\beta}_0, \widehat{\beta}_1) \in \operatorname{argmin} \sum_{i=1}^{n} \big(Y_i - (\beta_0 + \beta_1 x_i)\big)^2.$$

Definition 2.2 (Lineares Modell)**.** Sei $n \in \mathbb{N}$ und $(Y_1, \dots, Y_n)^\top$ reellwertige Zufallsvariablen.

(a) Ein *lineares Modell* besteht aus $\beta \in \mathbb{R}^p, p \le n$, $X \in \mathbb{R}^{n \times p}$ mit vollem Rang $\operatorname{rk} X = p$ und einem Zufallsvektor ε mit $\mathbb{E}\varepsilon = 0$ und einer symmetrischen positiv definiten Kovarianzmatrix Σ so, dass

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

(b) Für $y \in \mathbb{R}^n \setminus \{0\}$ existiert ein $x \in \mathbb{R}^n \setminus \{0\}$ mit $y = \Sigma x$, d.h. $\langle \Sigma^{-1} y, y\rangle = \langle x, \Sigma x\rangle > 0$ und

$$\langle \Sigma^{-1} y, y\rangle = \langle (\Sigma^\top)^{-1} y, y\rangle = \langle (\Sigma^{-1})^\top y, y\rangle = \langle y, \Sigma^{-1} y\rangle \qquad \text{für alle } y \in \mathbb{R}^n.$$

Deshalb ist auch Σ^{-1} symmetrisch positiv definit, d.h. $\Sigma^{-1/2} := \big(\Sigma^{-1}\big)^{1/2}$ is wohldefiniert. Ein kleinster quadrate Schätzer ist ein Schätzer

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|\Sigma^{-1/2}(Y - X\beta)\|^2.$$

Im *gewöhnlichen linearen Modell* ist $(\Sigma = \sigma^2 I_n)$ und der gewöhnliche Kleinste-Quadrate-Schätzer (OLS) $\widehat{\beta} \in \operatorname{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|^2$ ist unabhängig von Σ.

Lemma 2.3 (Orthogonale Projektion)**.** *In der Situation von Definition 2.2, setze* $X_\Sigma := \Sigma^{-1/2} X$.

(i) Die orthogonale Projektion von $\mathbb{R}^n$ *auf* $\operatorname{im}(X_\Sigma)$ *ist gegeben durch* $\Pi_{X_\Sigma} = X_\Sigma (X_\Sigma^\top X_\Sigma)^{-1} X_\Sigma^\top$.

(ii) Der KQS ist wohldefiniert und gegeben durch $\widehat{\beta} = (X_\Sigma^T X_\Sigma)^{-1} X_\Sigma^T \Sigma^{-1/2} Y$. *Der gewöhnliche KQS ist* $\widehat{\beta} = (X^T X)^{-1} X^T Y$.

Beweis.

(i) Zunächst ist $(X_\Sigma^\top X_\Sigma)$ invertierbar, da $\big(X_\Sigma^\top X_\Sigma\big)\, v = 0$ impliziert, dass

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

auf Grund der Rangbedingung an X. Es gilt dann $\Pi_{X_\Sigma}^2 = \Pi_{X_\Sigma}$, $\Pi_{X_\Sigma} X_\Sigma b = X_\Sigma b$ und $\Pi_{X_\Sigma}^\top = \Pi_{X_\Sigma}$.

(ii) Das folgt sofort aus der Normalengleichung $X_\Sigma^\top X_\Sigma \widehat{\beta} = X_\Sigma^\top \Pi_{X_\Sigma} \Sigma^{-1/2} Y$.

□

Theorem 2.4 (**Gauß-Markov**)**.** *Im gewöhnlichen linearen Modell mit* $\Sigma = \sigma^2 I_n$ *für ein* $\sigma^2 > 0$*, erhalten wir:*

(i) Der KQS $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$ *ist ein erwartungstreuer Schätzer für* β.

(ii) Für den Parameter $\gamma := \langle \beta, v\rangle$ *mit* $v \in \mathbb{R}^p$ *gilt, dass* $\widehat{\gamma} := \langle \widehat{b}, v\rangle$ *ein linearer erwartungstreuer Schätzer, der die Varianz in der Klasse der linearen, erwartungstreuen Schätzer minimiert, wobei* $\mathrm{Var}(\widehat{\gamma}) = \sigma^2 \|X(X^\top X)^{-1} v\|^2$.

(iii) Die normalisierte Stichprobenvarianz $\widehat{\sigma}^2 := \frac{\|Y - X\widehat{\beta}\|^2}{n-p}$ *ist ein erwartungstreuer Schätzer für* σ^2.

Beweis.

(i) Es gilt

$$\mathbb{E}\widehat{\beta} = \mathbb{E}(X^\top X)^{-1} X^\top Y = \mathbb{E}(X^\top X)^{-1} X^\top (X\beta + \varepsilon) = \beta.$$

(ii) Aus (i), folgt, dass $\widehat{\gamma}$ linear und erwartungstreu ist. Sei $\tilde{\gamma}$ ein weiterer solcher Schätzer von γ, d.h. es existiert ein $w \in \mathbb{R}^n$ so, dass $\tilde{\gamma} = \langle Y, w\rangle$. Dann gilt

$$\langle \beta, v\rangle = \mathbb{E}\langle Y, w\rangle = \langle X\beta, w\rangle = \langle \beta, X^\top w\rangle \qquad \text{für alle } \beta \in \mathbb{R}^p.$$

Der Riesz'sche Darstellungssatz liefert also $X^\top w = v$. Aus dem Satz des Pythagoras folgern wir jetzt

$$\mathrm{Var}\,(\tilde{\gamma}) = \mathbb{E}\langle \varepsilon, w\rangle^2 = \mathbb{E}\Big(\sum_{j=1}^n \varepsilon_j w_j\Big)^2 = \sigma^2\|w\|^2 = \sigma^2(\|w - \Pi_X w\|^2 + \|\Pi_X w\|^2).$$

Gleichzeitig gilt jedoch

$$\begin{aligned}\mathrm{Var}\,(\widehat{\gamma}) &= \mathbb{E}\langle \widehat{\beta} - \beta, v\rangle^2 = \mathbb{E}\langle (X^\top X)^{-1} X^\top \varepsilon, v\rangle^2 = \mathbb{E}\langle \varepsilon, \Pi_X w\rangle \\ &= \sigma^2\|\Pi_X w\|^2 = \sigma^2\|X(X^\top X)^{-1} v\|^2.\end{aligned}$$

(iii) Sei $e_1, \ldots, e_{n-p}$ eine ONB von $(\mathrm{im}X)^\perp$. Dann gilt wie vorher

$$\begin{aligned}\mathbb{E}\|Y - X\widehat{\beta}\|^2 &= \mathbb{E}\|(I_n - \Pi_X)Y\|^2 = \mathbb{E}\|(I - \Pi_X)\varepsilon\|^2 \\ &= \mathbb{E}\Big(\sum_{j=1}^{n-p} |\langle \varepsilon, e_j\rangle|^2\Big) = \sum_{j=1}^{n-p} \sigma^2\|e_j\|^2 = (n-p)\sigma^2.\end{aligned}$$

□

Bemerkung 2.5 (Vorhersagefehler).

(a) Der Schätzer $\widehat{\gamma}$ im Gauß-Markov-Theorem 2.4 wird BLUE (Best Linear Unbiased Estimator) genannt. Falls wir die Bedingung der Linearität oder der Erwartungstreue fallen lassen, gibt es bessere Schätzer bezüglich des MSE.

(b) Oft sind wir auch an dem Vorhersagefehler $\|X\widehat{\beta} - X\beta\|^2$ interessiert. Wie in Teil (iii) des Beweises von Theorem 2.4 erhalten wir

$$\mathbb{E}\|X\widehat{\beta} - X\beta\|^2 = \mathbb{E}\|\Pi_X(Y - X\beta)\|^2 = \mathbb{E}\|\Pi_X\varepsilon\|^2 = p\sigma^2.$$

Der Vorhersagefehler ist also linear in der Dimension des Modells.

(c) Die Aussage des Gauß-Markov-Theorems kann auf das allgemeine lineare Modell ausgeweitet werden.

Beispiel 2.6 (Beispiele von Regressionsmodellen).

(a) Das einfaceh *shift Modell* ist gegeben durch

$$Y_i = \mu + \varepsilon_i, \qquad 1 \le i \le n$$

mit $\varepsilon \sim N(0, \sigma^2 I_n), \mu = \beta$ und $X \in \mathbb{R}^{n\times 1} = [1, \ldots, 1]^\top$. Dann gilt

$$\widehat{\mu} = \widehat{\beta} = (X^\top X)^{-1} X^\top Y = \frac{1}{n}\sum_{i=1}^n Y_i \qquad \text{und} \qquad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \widehat{\mu})^2.$$

(b) Im einfachen Regressionsgeradenmodell erhalten wir mit $p = 2$, dass $\beta = (\beta_0, \beta_1)^\top$ und $X = (X_{ij})$ with $X_{i1} = 1$. $\mathrm{rk}\,X = 2$, falls $X_{i2} \neq X_{i'2}$ existieren.

(c) Mit dem Regressionsmodell kann man nicht nur lineare Zusammenhänge untersuchen. Betrachte zum Beispiel die *polynomielle Regression* mit $p+1$ Parmetern, d.h.

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{pmatrix}$$

Hier besitzt X vollen Rang sofern $p+1$ Designpukte unterschiedlich sind.

2.2 Inferenz unter Normalverteilungsannahmen

Satz 2.7 ($\mathbf{t_n}$- und $\mathbf{F_{m,n}}$-Verteilung). *Seien $X_1, \ldots, X_m, Y_1, \ldots, Y_n \sim N(0,1)$ i.i.d. Dann sind die Zufallsvariablen*

$$t_n := \frac{X_1}{\sqrt{\frac{1}{n}\sum_{j=1}^n Y_j^2}} \qquad \textit{und} \qquad F_{m,n} := \frac{\frac{1}{m}\sum_{i=1}^m X_i^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2}$$

verteilt mit Dichten

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \qquad x \in \mathbb{R}.$$

und

$$f_{m,n}(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}}, \qquad x \ge 0,$$

wobei $B(m/2, n/2) = \frac{\Gamma(m/2)\Gamma(n/2)}{\Gamma((m+n)/2)}$.

Beweis. Für die zweite Aussage beachte, dass $X := \sum_{i=1}^m X_i^2 \sim \chi_m^2, Y := \sum_{j=1}^n Y_j^2 \sim \chi_n^2$ unabhängig, wobei $\chi_n^2 \sim \Gamma(\frac{n}{2}, \frac{1}{2})$ und $\Gamma(p,b), p, b > 0$ die Dichte

$$f_{p,b}(y) = \frac{b^p}{\Gamma(p)} y^{p-1} e^{-by} \qquad \forall y > 0$$

und die charakteristische Funktion $\mathbb{R} \ni t \mapsto (1 - \frac{it}{b})^{-p}$ besitzt. Deshalb erhält man mit einer Dichtetransformation mit $\Phi(X,Y) = (X/Y, Y)$, dass

$$\begin{aligned}\mathbb{P}\{X/Y \in B\} &= \mathbb{P}\left\{\Phi(X,Y) \in B \times \mathbb{R}\right\} = \mathbb{P}\left\{(X,Y) \in \Phi^{-1}(B \times \mathbb{R})\right\} \\ &= \int_{\Phi^{-1}(B\times\mathbb{R})} f^X(x) f^Y(y)\, dxdy \overset{\text{Trafo}}{=} \int_B \int_0^\infty f^X(wy) f^Y(y) y\, dy\, dw,\end{aligned}$$

d.h. mit $w > 0$ und $u = (w+1)y$

$$\begin{aligned}f^{X/Y}(w) &= \int f_X(wy) f_Y(y) y\, dy = \int \frac{\left(\frac{1}{2}\right)^{m/2}}{\Gamma(\frac{m}{2})} (wy)^{\frac{m}{2}-1} e^{-\frac{1}{2}wy} \frac{\left(\frac{1}{2}\right)^{n/2}}{\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{1}{2}y} y\, dy \\ &= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int (wy)^{\frac{m}{2}-1} y^{\frac{n}{2}} e^{-\frac{1}{2}y(w+1)}\, dy \\ &\overset{\text{Trafo}}{=} \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int \left(\frac{w}{w+1}u\right)^{\frac{m}{2}-1} \left(\frac{u}{w+1}\right)^{\frac{n}{2}} e^{-\frac{1}{2}u} \frac{1}{w+1}\, du \\ &= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{w}{w+1}\right)^{\frac{m}{2}-1} \left(\frac{1}{w+1}\right)^{\frac{n}{2}+1} \int u^{\frac{m+n}{2}-1} e^{-\frac{1}{2}u}\, du = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} w^{\frac{m}{2}-1} (w+1)^{-\frac{m+n}{2}}.\end{aligned}$$

Eine weitere Transformation liefert, dass $F_{m,n} = \frac{nX}{mY}$ die Dichte

$$f_{m,n}(x) = \frac{1}{B(m/2, n/2)} \frac{(\frac{m}{n}x)^{m/2-1}}{(\frac{m}{n}x+1)^{(m+n)/2}} \frac{m}{n} = \frac{m^{m/2} n^{n/2}}{B(m/2,n/2)} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}, \qquad x \geq 0$$

besitzt.

Für die erste Aussage beachte, dass $t_n^2 = F_{1,n}$. Dichtetransformation liefert yields $f^{|t_n|}(x) = f_{1,n}(x^2) 2|x|, \forall x \geq 0$. Da t_n symmetrisch verteilt ist, gilt $f^{t_n}(x) = f_{1,n}(x^2)|x|, x \in \mathbb{R}$. Therefore,

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})} n^{\frac{n}{2}} \frac{(x^2)^{-\frac{1}{2}}}{(x^2+n)^{\frac{1+n}{2}}} |x| = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad \forall x \in \mathbb{R}.$$

□

Bemerkung 2.8 (Asymptotik von t_n und $F_{m,n}$).

(a) Für $n = 1$ ist t_n Cauchy-verteilt. Aus dem SLLN folgt $\frac{1}{n}\sum_{j=1}^n Y_j^2 \to \mathbb{E}Y_1^2 = 1$ fast sicher. Slutzkys Lemma liefert also $t_n \xrightarrow{d} N(0,1)$.

(b) Analog gilt $mF_{m,n} \xrightarrow{d} \chi_m^2$ für $n \to \infty$.

Satz 2.9 (Konfidenzmengen unter Normalverteilungsannahmen)**.** *Unter der Annahme, dass $\varepsilon \sim N(0, \sigma^2 I_n)$ gelten die folgenden Aussagen:*

(i) Für $\alpha \in (0,1)$ und das $(1-\alpha)$-Quantil $q_{1-\alpha}$ der $F_{p,(n-p)}$-Verteilung,

$$C := \left\{\beta \in \mathbb{R}^p : \|X(\widehat{\beta} - \beta)\|^2 \leq p q_{1-\alpha} \widehat{\sigma}^2\right\}$$

ist eine $(1-\alpha)$-Konfidenzmenge.

(ii) Für $\alpha \in (0,1)$ und das $(1-\frac{\alpha}{2})$-Quantil $q_{1-\frac{\alpha}{2}}$ der $t_{(n-p)}$-Verteilung ist

$$I := \left[\widehat{\gamma} - \widehat{\sigma}\sqrt{v^T(X^TX)^{-1}v}\, q_{1-\frac{\alpha}{2}}, \widehat{\gamma} + \widehat{\sigma}\sqrt{v^T(X^TX)^{-1}v}\, q_{1-\frac{\alpha}{2}}\right]$$

eine $(1-\alpha)$-Konfindenzmenge für $v \in \mathbb{R}^p$ und $\gamma := \langle \beta, v\rangle$.

Beweis.

(i) Sei $(e_i)_{i=1,\dots,n}$ eine ONB so, dass $(e_i)_{i\leq p}$ eine ONB von $\mathrm{im}(\Pi_X)$ and $(e_{i\geq p+1})$ eine ONB von $\mathrm{im}(I_n - \Pi_X)$ (Beachte: $\mathbb{R}^n = \mathrm{im}(\Pi_X) \oplus \mathrm{im}(I_n - \Pi_X)$ und $\mathrm{im}(\Pi_X) \perp \mathrm{im}(I_n - \Pi_X)$). For the orthogonal matrix $O := (e_1, \dots, e_n) \in \mathbb{R}^{n\times n}$, we obtain

$$[\langle \varepsilon, e_1\rangle, \dots, \langle \varepsilon, e_n\rangle]^\top = O^\top \varepsilon \sim N(0, \sigma^2 O^\top I_n O) = N(0, \sigma^2 I_n).$$

Deshalb gilt, dass

$$\|X(\widehat{\beta}-\beta)\|^2 = \|\Pi_X(Y-X\beta)\|^2 = \|\Pi_X\varepsilon\|^2 \overset{\text{Parcev.}}{=} \sum_{i=1}^{p}\langle\varepsilon, e_i\rangle^2$$

$$\text{und} \qquad \|Y-X\widehat{\beta}\|^2 = \|Y-\Pi_X Y\|^2 = \|\varepsilon-\Pi_X\varepsilon\|^2 \overset{\text{Parcev.}}{=} \sum_{i=p+1}^{n}\langle\varepsilon, e_i\rangle^2$$

unabhängig verteilt sind mit

$$\frac{\|X(\widehat{\beta}-\beta)\|^2}{p\,\widehat{\sigma}^2} = \frac{\sigma^{-2}(n-p)\|X(\widehat{\beta}-\beta)\|^2}{\sigma^{-2}p\|Y-X\widehat{\beta}\|^2} \overset{d}{=} \frac{\frac{1}{p}\chi^2_p}{\frac{1}{n-p}\chi^2_{(n-p)}} \overset{d}{=} F_{p,(n-p)},$$

nach Satz 2.7.

(ii) Weil

$$\langle\widehat{\beta}, v\rangle = \langle Y, \underbrace{X(X^\top X)^{-1}v}_{=:w}\rangle = \langle Y, \Pi_X w\rangle = \langle\Pi_X Y, w\rangle = \langle X\beta, w\rangle + \langle\Pi_X\varepsilon, w\rangle,$$

ist $\langle\widehat{\beta}, v\rangle$ unabhänig von $\widehat{\sigma}^2$ wie in (i). Dann folgt

$$\frac{\frac{\widehat{\gamma}-\gamma}{\sqrt{v^\top(X^TX)^{-1}v}}}{\widehat{\sigma}} = \frac{\frac{\widehat{\gamma}-\gamma}{\sqrt{\sigma^2 v^\top(X^TX)^{-1}v}}}{\sqrt{\widehat{\sigma}^2/\sigma^2}} \overset{d}{=} \frac{N(0,1)}{\sqrt{\frac{1}{n-p}\chi^2_{(n-p)}}} \overset{d}{=} t_{(n-p)}$$

aus Satz 2.7.

□

Beispiel 2.10 (Shift-Modell). Im einfachen Shift-Modell

$$Y_i = \beta_0 + \varepsilon_i, \qquad i \le n$$

aus 2.6 (a) liefert Satz 2.9 (ii), dass

$$\left[\overline{Y}_n \pm \frac{\widehat{\sigma}}{\sqrt{n}} q^{t_{n-p}}_{1-\frac{\alpha}{2}}\right]$$

ein $(1-\alpha)$-Konfidenzintervall für β_0 ist.

Beispiel 2.11 (**Interpretation einer Regressionstabelle**). Mit obigen Aussagen sind wir jetzt in der Lage die Angaben in einer stadard Regressionstabelle zu interpretieren. In Tabelle 2 sind die Ergebnisse einer Regression des Einkommens von Individuen auf die entsprechenden Bildungsjahre, das Prestige ihres Berufs und den Prozentsatz von Frauen in der Berufsgruppe zusammengefasst. Um die Tabelle zu interpretieren notieren wir folgendes:

Tabelle 2: Beispiel einer Regressionstabelle.

	Dependent variable:
	income
education.c	177.199
	(187.632)
prestige.c	141.435***
	(29.910)
women.c	−50.896***
	(8.556)
Constant	6,797.902***
	(254.934)
Observations	102
R^2	0.643
Adjusted R^2	0.632
Residual Std. Error	2,574.709 (df = 98)
F Statistic	58.890*** (df = 3; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

(a) Für jede Variable (education, prestige, women), erhalten wir einen Regressionskoeffizienten, d.h. den Wert von $\widehat{\beta}_k$.

(b) Für jeden Koeffizienten erhalten wir seinen Standardfehler in Klammern. Dies sind die Werte

$$\sqrt{\widehat{\sigma}^2 (X^\top X)^{-1}_{kk}} = \sqrt{\widehat{\sigma}^2 e_k^\top (X^\top X)^{-1} e_k}$$

aus Satz 2.9 (ii), d.h. das der t-Wert für den k-ten Koeffizienten unter der Hypothese $H_0 : \beta_k = 0$, gerade das Verhältnis des Koeffizienten und seines Standardfehlers ist. Sein Signifikanzlevel ist durch die Sterne angezeigt, wobei $p := \mathbb{P}\left\{|t_{(n-p)}| \geq |t\text{-stat}|\right\}$.

(c) Die Anzahl der Beobachtugnen ist gerade unser n.

(d) R^2 das Bestimmtheitsmaß. Ein Maß für die Güte des fits. In einem linearen Modell mit Intercept, d.h. $X_{i,1} = 1, i \leq n$, kann der erste Vektor der ONB in Satz 2.9 als $e_1 := (1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$ gewählt werden. Also

$$\begin{aligned} \underbrace{\|Y - \bar{Y}_n\|^2}_{=:\text{TSS}} &= \|Y - \langle Y, e_1\rangle e_1\|^2 \overset{\text{Pyth.}}{=} \|\sum_{i=2}^{p} \langle Y, e_i\rangle e_i\|^2 + \|\sum_{i=p+1}^{n} \langle Y, e_i\rangle e_i\|^2 \\ &= \|X\widehat{\beta} - \langle Y, e_1\rangle e_1\|^2 + \|Y - X\widehat{\beta}\|^2 = \underbrace{\|X\widehat{\beta} - \bar{Y}_n\|^2}_{=:\text{ESS}} + \underbrace{\|Y - X\widehat{\beta}\|^2}_{=:\text{RSS}}., \end{aligned} \tag{2.1}$$

wobei TSS für "Total Sum of Squares", ESS für "Explained Sum Squared" und RSS für "Residual Sum of Squares" steht. Es gilt dann

$$R^2 := \frac{\|X\widehat{\beta} - \bar{Y}_n\|^2}{\|Y - \bar{Y}_n^2\|} = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Da $R^2 = 1 - \text{SSR/TSS}$, ist R^2 monoton in der Anzahl der Variablen des Models. Deswegen wird die adjusted R^2 Statistik als weiteres Maß angegeben, wobei

$$\text{Adjusted } R^2 := 1 - \frac{\frac{1}{n-p}\|Y - X\widehat{\beta}\|^2}{\frac{1}{n-1}\|Y - \bar{Y}_n\|^2}.$$

(e) Residual Std. Error ist gerade der Schätzer für die Standardabweichung der Fehlervariablen

$$\sqrt{\widehat{\sigma}^2} = \sqrt{\|Y - X\widehat{\beta}\|^2/(n-p)}.$$

(f) Die F-Statistik ist

$$F := \frac{\frac{1}{p-1}\|X\widehat{\beta} - \bar{Y}_n\|^2}{\frac{1}{n-p}\|Y - X\widehat{\beta}\|^2} = \frac{\frac{1}{p-1}\|X\widehat{\beta} - \langle Y, e_1\rangle e_1\|^2}{\frac{1}{n-p}\|Y - X\widehat{\beta}\|^2}.$$

Unter der Hypothese $H_0 : \beta_2 = 0 = \cdots = 0 = \beta_p$, liefert dieselbe Argumentation wie in 2.9, dass $F \overset{d}{=} F_{(p-1),(n-p)}$.

Man kann sich die Frage Stellen, inwiefern die Konstruktion der Konfidenzintervalle in Satz 2.9 natürlich ist. Darauf gibt es im wesentlichen zwei Antworten. Erstens ist aus mathematischer Sicht im linearen Modell nur der transformierte Parameter $X\beta$ und nicht β selbst von Interesse. Y ist um $X\beta$ herum isotop (d.h. rotationsinvariant) verteilt wegen $Y \sim N(X\beta, \sigma^2 I_n)$, also sind Kugeln natürliche Konfidenzbereiche für $X\beta$, die sich zu Ellipsoiden für β transformieren.

Eine zweite Antwort liefert der Liklihoodquotiententest (LQ-Test), den wir allgemeiner für *lineare Hypothesen* betrachten.

Definition 2.12 (Lineares Testproblem)**.** Im gewöhnlichen linearen Modell ist ein (zweiseitges) *lineares Testproblem* gegeben durch

$$H_0 : K\beta = c, \sigma > 0 \qquad \text{vs.} \qquad H_1 : K\beta \neq c, \sigma > 0.$$

für eine deterministische *Kontrastmatrix* $K \in \mathbb{R}^{r \times p}$ von vollem Rang $\operatorname{rk} K = r \leq p$ und $c \in \mathbb{R}^r$ vorgegeben. Beachte, dass $\{\beta \in \mathbb{R}^p : K\beta = c\}$ ein $(p-r)$-dim affiner Unterraum des $\mathbb{R}^p$ ist. Es werden unter H_0 also $r \leq p$ linear unabhängige Bedingungen an den Parameter β gestellt.

Beispiel 2.13 (Lineare Hypothesen).

(a) Die Hypothese $H_0 : \beta = \beta_0$ für ein $\beta_0 \in \mathbb{R}^p$ entspricht $r = p, K = I_p, c = \beta_0$.

(b) Die Hypothese $H_0 : \beta \in \mathbb{R}^p$ mit $\beta_1 = \ldots \beta_2$ entspricht $r = p - 1, c = 0$ und

$$K = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & & \\ & & & 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

In der Situation von Definition 2.12 erhält man den Likelihood-Quotienten als Z/N mit

$$Z = \sup_{\beta\in\mathbb{R}^p,\sigma>0} (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\|Y-X\beta\|^2}{2\sigma^2}} \qquad \text{und} \qquad N = \sup_{K\beta=c,\sigma>0} (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{-\|Y-X\beta\|^2}{2\sigma^2}}$$

Das Bilden des Supremums über β führt gerade auf die Minimierung von $|Y - X\beta|^2$. Wir betrachten also die „residual sum of squares“

$$\mathrm{RSS} := \inf_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 = \|Y - X\widehat{\beta}\|^2 \qquad \text{und} \qquad \mathrm{RSS}_{H_0} := \inf_{\beta: K\beta=c} \|Y - X\beta\|^2 = \|Y - X\widehat{\beta}_{H_0}\|^2,$$

wobei $\widehat{\beta}$ der KQ-Schätzer ist und $\widehat{\beta}_{H_0}$ der KQ-Schätzer unter Einschränkung auf H_0 ist. Als ML-Schätzer erhalten wir

$$\widehat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\|Y - X\widehat{\beta}\|^2 \qquad \text{und} \qquad \widehat{\sigma}^2_{\mathrm{ML},H_0} = \frac{1}{n}\|Y - X\widehat{\beta}_{H_0}\|^2.$$

Als Likelihood-Quotient erhalten wir also bis auf Multiplikation mit Konstanten und Potenzierung

$$\frac{\|Y - X\widehat{\beta}_{H_0}\|^2}{\|Y - X\widehat{\beta}\|^2} = 1 + \frac{\|Y - X\widehat{\beta}_{H_0}\|^2 - \|Y - X\widehat{\beta}\|^2}{\|Y - X\widehat{\beta}\|^2} = 1 + \frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\mathrm{RSS}}$$

Das führt auf die Teststatistik

$$F := \frac{\frac{1}{r}(\mathrm{RSS}_{H_0} - \mathrm{RSS})}{\frac{1}{n-p}\mathrm{RSS}}.$$

Satz 2.14 (F-Statistik für lineare Hypothesen)**.** *Im gewöhnlichen linearen Modell betrachten wir unter der Normalverteilungsannahme $\varepsilon \sim N(0, \sigma^2 I_n)$ das lineare Testproblem*

$$H_0 : K\beta = c \qquad vs. \qquad H_1 : K\beta \neq c$$

mit einer Kontrastmatrix $K \in \mathbb{R}^{r\times p}$ mit vollem Rang $\operatorname{rk} K = r \leq p$ und $c \in \mathbb{R}^r$. Es gilt dann die folgenden Aussagen.

(i) Der KQ-Schätzer eingeschränkt auf H_0 ist $\widehat{\beta}_{H_0} = \widehat{\beta} - (X^\top X)^{-1}K^\top\big(K(X^\top X)^{-1}K^\top\big)^{-1}(K\widehat{\beta} - c)$, wobei $\widehat{\beta}$ der unrestringierte KQ-Schätzer ist.

(ii) Es gilt

$$RSS_{H_0} - RSS = (K\widehat{\beta} - c)^\top (K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - c)$$

mit $(RSS_{H_0} - RSS)/\sigma^2 \sim \chi^2_r$ unter H_0.

(iii) Die Fisher-Statistik

$$F := \frac{\frac{1}{r}(RSS_{H_0} - RSS)}{\frac{1}{n-p}RSS} = \frac{\frac{1}{r}(\|Y - X\widehat{\beta}_{H_0}\|^2 - \|Y - X\widehat{\beta}\|^2)}{\frac{1}{n-p}\|Y - X\widehat{\beta}\|^2}$$

ist unter H_0 gemäß $F_{r,n-p}$-verteilt.

Beweis.

(i) Aus dem Satz des Pythagoras folgt $\|Y - X\beta\|^2 = \|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta} - X\beta\|^2$. Deswegen ist es hinreichend $\|X\widehat{\beta} - X\beta\|^2$ im Bezug auf $K\beta = c$ zu minimieren. Auf Grund der Nebenbedingung können wir schreiben $\beta = \beta_0 + \beta'$ mit einem beliebigen β_0, das $K\beta_0 = c$ erfüllt und $\beta' \in \ker K$. Das führt auf das Problem

$$\min_{\beta'\in\ker K} \|X(\widehat{\beta} - \beta_0 - \beta')\|^2 = \min_{\beta'\in\ker K} \|(\widehat{\beta} - \beta_0) - \beta'\|^2_X,$$

wobei wir das zusatzliche Skalarprodukt $\langle ., .\rangle_X := \langle X., X.\rangle$ mit der entsprechenden Norm $\|.\|_X$ einführen. Da $\ker K = (\operatorname{im} K^*)^\perp$, wobei K^* die Adjungierte von K bezüglich $\langle ., .\rangle_X$ ist und auch das Komplement bezüglich diesem Skalarprodukt gebildet wird, erhalten wir

$$\beta' = (I_p - K^*(KK^*)^{-1}K)(\widehat{\beta} - \beta_0) = \widehat{\beta} - \beta_0 - K^*(KK^*)^{-1}(K\widehat{\beta} - c)$$

und damit

$$\widehat{\beta}_{H_0} = \beta_0 + \beta' = \widehat{\beta} - K^*(KK^*)^{-1}(K\widehat{\beta} - c).$$

Für $u, v \in \mathbb{R}^p$ zeigt die Rechung

$$\begin{aligned}\langle XKu, Xv\rangle &= \langle u, K^\top(X^\top X)v\rangle = \langle (X^\top X)u, (X^\top X)^{-1}K^\top(X^\top X)v\rangle \\ &= \langle Xu, X(X^\top X)^{-1}K^\top(X^\top X)v\rangle,\end{aligned}$$

dass $K^* = (X^\top X)^{-1}K^\top(X^\top X)$. Das zeigt schließlich, dass

$$\begin{aligned}\widehat{\beta}_{H_0} &= \widehat{\beta} - K^*\big(K(X^\top X)^{-1}K^\top(X^\top X)\big)^{-1}(K\widehat{\beta} - c) \\ &= \widehat{\beta} - (X^\top X)^{-1}K^\top(K(X^\top X)^{-1}K^\top)^{-1}(K\widehat{\beta} - c).\end{aligned}$$

(ii) Es gilt nach dem Satz von Pythagoras

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

Unter Beachtung von der Tatsache, dass $K\widehat{\beta}$ r-dimensional gauß'sch-verteilt ist mit

$$\mathbb{E}(K\widehat{\beta}) = \mathbb{E}(K(X^\top X)^{-1}XY) = K(X^\top X)^{-1}XX^\top\beta = c$$

und

$$\mathrm{Cov}(K\widehat{\beta}) = \mathrm{Cov}(K(X^\top X)^{-1}X^\top\varepsilon) = \sigma^2 K(X^\top X)^{-1}K^\top$$

unter H_0, folgt $(\mathrm{RSS}_{H_0} - \mathrm{RSS})/\sigma^2 \sim \chi^2_r$.

(iii) Nach (ii) ist $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ eine Funktion von $\widehat{\beta}$ und damit auf dieselbe Art und Weise unabhängig von $\|Y - X\widehat{\beta}\|^2$ wie in der Situation von 2.9(ii). Die Verteilungsbehauptung folgt jetzt sofort.

□

Bemerkung 2.15 (χ^2_r-Test bei bekanntem σ^2). Ist in der Situation von Satz 2.14 $\sigma^2 > 0$ bekannt, so ist unter $H_0 : K\beta = c$ die Teststatistik

$$T := \frac{\|(K(X^\top X)^{-1}K^\top)^{-1/2}(K\widehat{\beta} - c)\|^2}{\sigma^2} \sim \chi^2_r$$

und $\varphi_\alpha = \mathbf{1}(T > q^{\chi^2_r}_{1-\alpha})$ ist der entsprechende Likelihood-Quotienten-Test.

Beispiel 2.16 (Jungengeburten). Wir betrachten folgende Geburtenstatistik

Monat	1	2	3	4	5	6	7	8	9	10	11	12
Jungengeburten	50.4	51.9	52.6	51.3	51.6	52.7	51.1	51.7	50.9	50.5	52.9	51.4

Notiz: Jeden Monat beträgt die Gesamtzahl der Geburten ca. 3000

Tabelle 3: Jungengeburten über Monate in %.

(a) Wir beobachten $Y_i \sim \mathrm{Bin}(n_i, p_i), i = 1, \ldots, 12$ mit $p_1, \ldots, p_{12}$ unbekannt und $n_1 \approx \cdots \approx n_2 \approx 3000$. Durch die Normalapproximation

$$\frac{Y_i - np_i}{\sqrt{np_i(1-p_i)}} \xrightarrow[n\to\infty]{d} N(0,1)$$

und $p_i(1-p_i) \approx 1/4$ erhalten wir folgendes approximatives Modell:

$$\frac{Y_i}{3000} = \overline{Y}_i = p_i + \varepsilon_i \qquad \text{mit} \qquad \varepsilon_i \sim N(0, 1/(4n)) \text{ i.i.d.}$$

Wir wollen testen

$$H_0 : p_1 = \cdots = p_{12} \qquad \text{vs.} \qquad H_1 : \exists j, k \in \{1, \ldots, 12\} : p_j \neq p_k.$$

Da in unserem Modell gilt $p = n$, können wir nicht die standard F-Statistik verwenden, aber $\sigma^2 > 0$ ist bekannt. Also gilt mit K wie in Beispiel 2.13 (b), dass

$$\begin{aligned}T = \frac{\|(KK^\top)^{-1/2}K\widehat{\beta}\|^2}{\sigma^2} = \frac{\langle (KK^\top)^{-1}K\widehat{\beta}, K\widehat{\beta}\rangle}{\sigma^2} = \frac{\langle K^\top(KK^\top)^{-1}KY, Y\rangle}{\sigma^2} = \frac{\langle \Pi_{K^\top}Y, Y\rangle}{\sigma^2} \\ = \frac{\langle (I_n - \Pi_{\ker K})Y, Y\rangle}{\sigma^2} = \frac{\|(I_n - \Pi_{\ker K})Y\|^2}{\sigma^2} = \frac{\|Y - \overline{Y}\mathbf{1}\|^2}{\sigma^2} \sim \chi^2_{p-1}.\end{aligned}$$

Für die Beispieldaten erhalten wir

$$T = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \overline{Y})^2 \approx 6.19$$

Unter H_0 gilt aber bereits $\mathbb{E}T = 11$, d.h. H_0 wird zu keinem vernünftigen Niveau verworfen.

(b) Man kann zeigen, dass dieser Test asymptotisch gut ist vgl. Übung.

2.3 Varianzanalyse

Die Varianzanalyse (engl. analysis of variance, ANOVA) ist eine Gruppe statistischer Verfahren mit verschiedenen Anwendungen, wobei sie alle Varianzen und Prüfgrößen nutzen, um Erkenntnisse über Strukturen in gegebenen Daten zu erlangen.

Beispiel 2.17 (Düngemittel 1). Wir vergleichen den Ernteertrag unter verschiedenen Anbaumethoden/Düngemitteln und wollen wissen, ob es einen Einfluss des Faktors Dünger auf den Ertrag gibt. Z.B. 3 verschiedene Düngemittel werden auf n_1, n_2, n_3 Felder ausgebracht. Wir erhalten $Y_{11}, \dots Y_{1n_1}$ Erträge für Dünger 1 etc.

Wir betrachten ein einfaches statistisches Modell

$$\begin{aligned} Y_{1j} &= \mu_1 + \varepsilon_{1j}, \quad j = 1, \dots, n_1 \qquad \text{mit} \qquad \varepsilon_{1,j} \sim N(0, \sigma^2) \text{ i.i.d.}, \\ Y_{2j} &= \mu_2 + \varepsilon_{2j}, \quad j = 1, \dots, n_2 \qquad \text{mit} \qquad \varepsilon_{2,j} \sim N(0, \sigma^2) \text{ i.i.d.}, \\ Y_{3j} &= \mu_3 + \varepsilon_{3j}, \quad j = 1, \dots, n_3 \qquad \text{mit} \qquad \varepsilon_{3,j} \sim N(0, \sigma^2) \text{ i.i.d.}, \end{aligned}$$

wobei die Fehler auch gemeinsam unabhängig sind.

Definition 2.18 (ANOVA1)**.** Das Modell der *einfaktoriellen Varianzanalyse* ANOVA1 ist gegeben durch die Beobachtungen

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

mit Störgrößen $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. Den ersten Index bezeichnen wir als *Faktor*. Mit $n := \sum_i n_i$ bezeichnen wir den *Gesamtstichprobenumfang*. Für $n_1 = \dots = n_k$ spricht man von *balanciertem Design*.

Bemerkung 2.19 (Spezialfall des linearen Modells).

(a) Das ANOVA1 Modell in Definition 2.18 ist ein Spezialfall des linearen Modells. Es gilt

$$Y = X\mu + \varepsilon = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_1} \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}}_{n \times k} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \vdots \\ \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn_1} \end{pmatrix}$$

Es gilt sofort $\operatorname{rk} X = k$.

(b) Die klassische Frage der Varianzanalyse lautet: „Existieren Unterschiede zwischen den Faktor-spezifischen Mittelwerten μ_i?“ Dies führt auf das Testproblem

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

Allgemein gesprochen können wir die Nullhypothese ablehnen, wenn wir große Abstände zwischen den Durchschnitten mit kleinen Varianzen innerhalb der Gruppen haben.

Umgekehrt werden wir die Nullhypothese nicht ablehnen können, wenn wir kleine Abstände zwischen den Durchschnitten mit großen Varianzen innerhalb der Gruppen haben.

Satz 2.20 (ANOVA1 Streuungszerlegung)**.** *Im ANOVA1-Modell definiert man für $i = 1, \dots, k$ das Gruppenmittel beziehungsweise das Gesamtmittel*

$$\overline{Y}_{i\cdot} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \qquad \textit{bzw.} \qquad \overline{Y}_{\cdot\cdot} := \frac{1}{n} \sum_{i=1} \sum_{j=1}^{n_i} Y_{ij},$$

sowie die between group sum of squares *und die* within sum of squares

$$BSS := \sum_{i=1}^{k} n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \qquad \textit{und} \qquad WSS := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2.$$

Für die total sum of squares *gilt dann*

$$TSS := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = BSS + WSS.$$

Beweis. Es gilt

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i})^2 + 2(Y_{ij}-\overline{Y}_{i})(\overline{Y}_{i}-\overline{Y}_{..}) + (\overline{Y}_{i}-\overline{Y}_{..})^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i})^2 + 0 + \sum_{i=1}^{k} n_i(\overline{Y}_{i}-\overline{Y}_{..})^2.$$

□

Satz 2.21 (Inferenz im ANOVA1-Modell)**.** *Im ANOVA1-Modell gilt:*

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

(ii) Es gilt $WSS/\sigma^2 \sim \chi^2_{n-k}$ und unter $H_0 : \mu_1 = \dots = \mu_k$ gilt $BSS/\sigma^2 \sim \chi^2_{k-1}$ mit $\frac{BSS/(k-1)}{WSS/(n-k)} \sim F_{k-1,n-k}$.

Beweis. Für (i) gilt nach Bemerkung 2.19 (a), dass

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

Für (ii) wenden wir Satz 2.14 an. Nach (i) gilt

$$\mathrm{RSS} = \|Y - X\widehat{\mu}\|^2 = \sum_{i,j}(Y_{ij}-\overline{Y}_{i.})^2 = \mathrm{WSS}.$$

Weiter gilt

$$\mathrm{RSS}_{H_0} = \min_{\mu_1=\dots=\mu_k} \|Y - X(\mu_1, \dots, \mu_k)^\top\|^2 = \|Y - \overline{Y}\mathbf{1}\|^2 = \mathrm{TSS}.$$

Die Behauptung folgt jetzt, da wegen Satz 2.20 gilt, dass $\mathrm{BSS} = \mathrm{TSS} - \mathrm{WSS} = \mathrm{RSS}_{H_0} - \mathrm{RSS}$. □

Oft ist es sinnvoll die Ergebnisse der ANOVA1 in eine Tabelle zu schreiben. FG steht für Freiheitsgerade.

	FG	Quadratsummen	Quadratmittel	F-Stat
Between	$k-1$	BSS	$\mathrm{BSS}/(k-1)$	
Within	$n-k$	WSS	$\mathrm{WSS}/(n-k)$	$\frac{\mathrm{BSS}/(k-1)}{\mathrm{WSS}/(n-k)}$
Total	$n-1$	TSS	$\mathrm{TSS}/(n-1)$	

Tabelle 4: ANOVA1-Tafel

Im zwei Faktormodell können wir noch einen weiteren einfachen Test einführen.

Corollar 2.22 (Zweistichproben-t-Test)**.** *Im ANOVA1-Modell mit $k = 2$ und $H_0 : \mu_1 = \mu_2$ ist*

$$\frac{\overline{Y}_{1.} - \overline{Y}_{2.}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\,WSS/(n-2)}} \sim t_{n-2}.$$

Der Zweistichproben-t-Test zum Niveau $\alpha \in (0,1)$ ist damit gegeben durch $\varphi(|t_{n-2}| \geq q_{1-\alpha/2})$.

Beweis. Es gilt $F_{1,n-2} \sim \frac{BSS}{WSS/(n-2)}$ mit

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

Da t_n symmetrisch ist und $t_n^2 = F_{1,n}$ folgt die Behauptung. □

Bemerkung 2.23 (Effektdarstellung). Das ANOVA1-Modell lässt sich umformen zu

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

und $\sum_{i=1}^k n_i\alpha_i = 0$. Diese Form heißt Effektdarstellung. Wir schätzen also den Parameter $(\mu_0, \alpha_1, \dots, a_{k-1})$. An dem Ergebnis lässt sich der Effekt des Faktors direkt ablesen. Wegen der Äquivalenz der Probleme

$$\min_{\mu\in\mathbb{R}^k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2 \qquad \text{und} \qquad \min_{(\mu_0,\alpha_1,\dots,\alpha_{k-1})\in\mathbb{R}^k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu_0-\alpha_i)^2$$

muss gelten $\widehat{\mu}_0 + \widehat{\alpha}_i = \widehat{\mu}_i, i = 1, \dots, k$ und damit

$$n\widehat{\mu}_0 = \sum_{i=1}^{k} n_i(\widehat{\mu}_0 + \widehat{\alpha}_i) = \sum_{i=1}^{k} n_i \widehat{\mu}_i$$

Auf Grund dieser Äquivalenz können die Modelle ineinander überführt werden und alle Teststatistiken die wir hergeleitet haben können hier verwendet werden.

3 Exponentialfamilien und verallgemeinerte lineare Modelle

3.1 Die Informationsungleichung

Definition 3.1 (Reguläre statistische Modelle)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ ein vom σ-endlichen Maß μ dominiertes Modell.

(a) Solch ein Modell heißt *regulär* falls

(i) $\Theta \subset \mathbb{R}^d$ ist offen.

(ii) Die Likelihood-Funktion $L(\theta, x)$ ist auf $\Theta \times \mathfrak{X}$ strikt positiv und nach θ stetig differenzierbar. Insbesondere existiert die *Scorefunktion*

$$U_\theta(x) := \nabla_\theta \ln L(\theta, x) = \frac{\nabla_\theta L(\theta, x)}{L(\theta, x)}, \qquad x \in \mathfrak{X}.$$

(iii) Für alle $\theta \in \Theta$ existiert die *Fisher-Information* $I(\theta) := \mathbb{E}_\theta\big(U_\theta U_\theta^\top\big)$ und ist positiv definit.

(iv) Es gilt die Vertauschungsrelation

$$\nabla_\theta \int h(x)L(\theta, x)\,\mu(dx) = \int h(x)\nabla_\theta L(\theta, x)\,\mu(dx)$$

mit $h(x) := 1, x \in \mathfrak{X}$.

(b) In einem regulären Modell heißt ein Schätzer $T : \mathfrak{X} \to \mathbb{R}$ *regulär* falls $\mathbb{E}_\theta T^2 < \infty$ für alle $\theta \in \Theta$ und die Relation in (iv) auch für $h(x) = T(x)$ gilt.

Bemerkung 3.2 (Hinreichende Bedingung über DCT und Fisher-Information)*.*

(a) Der Satz über die dominierte Konvergenz liefert eine hinreichende Bedingung für Vertauschbarkeit von Integration und Differentiation in Definition 3.1 (iv). Sie gilt falls für jedes $\theta_0 \in \Theta$ eine Umgebung V_{θ_0} existiert so, dass

$$\int \sup_{\theta\in V_{\theta_0}} |\nabla_\theta L(\theta, x)|\,\mu(dx) < \infty.$$

Oft kann man die Bedingung aber auch für jedes Modell einzeln prüfen.

(b) Aus Definition 3.1 (iv) erhalten wir außerdem sofort

$$\int \nabla_\theta L(\theta, x)\,\mu(dx) = \nabla_\theta \int L(\theta, x)\,\mu(dx) = \nabla_\theta 1 = 0.$$

Das impliziert $\mathbb{E}U_\theta = 0$ und damit $\mathrm{Cov}_\theta(U_\theta) = I(\theta)$.

(c) Die Fisher-Information ist tatsächlich informativ: $I(\theta) = 0$ gilt auf einer Umgebung $\Theta_0 \subset \Theta$ genau dann wenn U_θ konstant, d.h. identisch 0 fast überall ist. In diesem Fall ist $L(\theta, x)$ konstant in θ fast überall, d.h. keine Beobachtung ist in der Lage zwischen den Paramtern zu unterscheiden. Weiter gilt für die Fisher-Information $I_n(\theta)$ eines n-fachen Produktmodells, dass

$$I_n(\theta) = \mathrm{Cov}_\theta\Big(\nabla_\theta \sum_{i=1}^{n} \ln L(\theta, X_i)\Big) = \sum_{i=1}^{n} \mathrm{Cov}_\theta(\nabla_\theta \ln L(\theta, X_i)) = nI(\theta).$$

Satz 3.3 (**Cramér-Rao-Informationsschranke**)**.** *In einem regulären statistischen Modell $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ sei $g : \Theta \to \mathbb{R}$ eine stetig differenzierbare Funktion und T ein erwartungstreuer regulärer Schätzer von $g(\theta)$. Dann gilt*

$$Var_\theta(T) \geq \nabla g(\theta)^\top I(\theta)^{-1} \nabla g(\theta) \qquad \text{für alle } \theta \in \Theta.$$

Beweis. Da $\mathbb{E}_\theta U_\theta = 0$ folgt aus der Regularität von T für beliebiges $v \in \mathbb{R}^k$, dass

$$\begin{aligned}\mathrm{Cov}_\theta(\langle U_\theta, v\rangle, T) &= \mathbb{E}_\theta(T\langle U_\theta, v\rangle) = \langle \int T(x)\nabla_\theta L(\theta, x)\,\mu(dx), v\rangle \\ &= \langle \nabla_\theta \int T(x)L(\theta, x)\,\mu(dx), v\rangle = \langle \nabla_\theta \mathbb{E}_\theta T, v\rangle = \langle \nabla g, v\rangle.\end{aligned}$$

Somit erhalten wir aus Cauchy-Schwarz

$$\langle \nabla g, v\rangle^2 = \mathrm{Cov}_\theta(\langle U_\theta, v\rangle, T)^2 \leq \mathrm{Var}_\theta(\langle U_\theta, v\rangle)\mathrm{Var}(T) = \langle v, I(\theta)v\rangle \mathrm{Var}(T).$$

Die Behauptung folgt jetzt mit $v = I(\theta)^{-1}\nabla g(\theta)$. □

Man kann zeigen, dass Gleichheit in der Cramér-Rao-Informationsschranke auf eine bestimmte Klasse von statistischen Modellen führt, siehe Buchprojekt.

Definition 3.4 (Exponentialfamilien)**.** Sei $(\mathfrak{X}, \mathscr{F}, (\mathbb{P})_{\theta\in\Theta})$ ein statistisches Modell, das von einem σ-endlichen Maß μ dominiert wird.

(a) $(\mathbb{P}_\theta)_{\theta\in\Theta}$ heißt *Exponentialfamilie* falls $k \in \mathbb{N}$ und $\eta : \Theta \to \mathbb{R}^k$ existieren, so dass

$$\frac{d\mathbb{P}_\theta}{d\mu}(x) = h(x) \exp\big(\langle \eta(\theta), T(x)\rangle - b(\eta(\theta))\big), \qquad x \in \mathfrak{X},$$

für messbare Abbildungen $T : \mathfrak{X} \to \mathbb{R}^k$, $h : \mathfrak{X} \to (0, \infty)$.

(b) Für eine Exponentialfamilie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ wird die Menge

$$H := \Big\{\eta \in \mathbb{R}^k : \int h(x) \exp\big(\langle \eta, T(x)\rangle\big)\, \mu(dx) \in (0, \infty)\Big\}$$

natürliche Parametermenge genannt. Die Familie $(\mathbb{P}_\eta)_{\eta\in H}$ wird *natürliche Exponentialfamilie* genannt.

Bemerkung 3.5 (Mehrdeutigkeit und Normalisierungskonstante).

(a) Beachte, dass

$$b(\eta) = \ln \int h(x) \exp\big(\langle \eta, T(x)\rangle\big)\, \mu(dx)$$

nur eine Normalisierungskonstante ist.

(b) Exponentialfamilien sind nicht eindeutig bestimmt. Z.B. kann h in das dominierende Maß μ absorbiert werden.

(c) (Man will lineare Struktur zwischen Parameter und x.)

Beispiel 3.6 (Wichtige Fälle von Exponentialfamilien).

(a) Betrachte $\mathbb{P}_p = \mathrm{Bin}(n, p)$ für $p \in (0, 1)$. Für das Zählmaß μ auf $\mathfrak{X} = \mathbb{N}_0$ ausgestattet mit der Potenzmenge als σ-Algebra erhalten wir

$$\frac{d\mathbb{P}_p}{d\mu}(k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \exp\Big(k \ln \frac{p}{1-p} + n \ln(1-p)\Big), \qquad k \in \mathfrak{X}.$$

Bezüglich $\eta = \ln \frac{p}{1-p}$ bilden die Maße $\mathbb{P}_\eta = \mathrm{Bin}(n, \frac{e^\eta}{1+e^\eta})$ eine natürliche Exponentialfamilie mit

$$\frac{d\mathbb{P}_\eta}{d\mu}(k) = \binom{n}{k} \exp\Big(k \cdot \eta + n \ln\Big(1 - \frac{e^\eta}{1+e^\eta}\Big)\Big), \qquad k \in \mathbb{N}_0.$$

Der natürlicher Parameterbereich ist gegeben durch $H = \big\{\eta \in \mathbb{R} : \sum_{k=0}^n \binom{n}{k} e^{\eta k} \in (0, \infty)\big\} = \mathbb{R}$.

(b) Die Normalverteilungen $\mathbb{P}_\theta = N(\mu, \sigma^2)$ für $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ bilden eine Exponentialfamilie bezüglich des Lebesguemaßes λ mit Parameter $\eta = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$, da

$$\begin{aligned}\frac{d\mathbb{P}_\theta}{d\lambda}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\ln\sigma^2 - \frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2}\Big) \\ &= \frac{1}{\sqrt{2\pi}} \exp\Big(\frac{\mu}{\sigma^2}x + \frac{1}{2\sigma^2}(-x^2) - \frac{1}{2}\ln\sigma^2 - \frac{\mu^2}{2\sigma^2}\Big).\end{aligned}$$

Die natürliche Parametermenge ist $\mathbb{R} \times (0, \infty)$. Ist σ^2 bekannt, so ist $\eta = \mu$ der natürlicher Parameter.

(c) Für eine Exponentialfamilie $(\mathbb{P}_\eta)_{\eta\in H}$ ist auch die Produktfamilie $(\mathbb{P}_\eta^{\otimes n})_{\eta\in H}$ eine Exponentialfamilie, da

$$\frac{d\mathbb{P}_\eta^{\otimes n}}{d\mu^{\otimes n}}(x) = \prod_{i=1}^n \frac{d\mathbb{P}_\eta}{d\mu}(x_i) = \prod_{i=1}^n h(x_i) \exp\Big(\langle \eta, \sum_{i=1}^n T(x_i)\rangle - nb(\eta)\Big), \qquad x \in \mathfrak{X}.$$

Satz 3.7 (**M.g.f. für Exponentialfamilien**)**.** *Betrachte ein statistisches Modell $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\eta)_{\eta\in H})$ mit einer natürlichen Exponentialfamilie*

$$\frac{d\mathbb{P}_\eta}{d\mu}(x) = h(x) \exp\big(\langle \eta, T(x)\rangle - b(\eta)\big), \qquad x \in \mathfrak{X}.$$

(i) Sei $\eta_0 \in \mathring{H}$ und $f \in L^1(\mathbb{P}_\eta)$ für alle $\eta \in B_\varepsilon(\eta_0)$. Dann ist die Funktion

$$B_\varepsilon(\eta_0) \ni \eta \mapsto \int f(x)h(x)e^{\langle \eta, T(x)\rangle}\, \mu(dx)$$

unendlich oft differenzierbar und ihre Ableitungen können durch Differenzieren unter dem Integral berechnet werden.

(ii) In der Situation von (i) ist die Funktion

$$\mathbb{C}^k \supset B_\varepsilon(\eta_0) \times \mathbb{R}^k \ni \eta + i\tilde{\eta} \mapsto \int f(x)h(x)e^{\langle \eta + i\tilde{\eta}, T(x)\rangle}\, \mu(dx)$$

wohldefiniert und analytisch in jedem der Argumente $(\eta_j + i\tilde{\eta}_j)$. Ihre Ableitungen können ebenfalls durch Differenzieren unter dem Integral berechnet werden.

(iii) Für $\eta_0 \in \mathring{H}$ ist die momentengenerierende Funktion ψ_{η_0} von $\mathbb{P}^T_{\eta_0}$ in einer Nullumgebung endlich und unendlich oft differenzierbar mit

$$\psi_{\eta_0}(u) = \mathbb{E}_{\eta_0} e^{\langle u, T\rangle} = \exp\big(b(\eta_0 + u) - b(\eta_0)\big).$$

Weiter gilt

$$\mathbb{E}_{\eta_0} T_j = \partial_{\eta_j} b(\eta_0) \qquad \textit{und} \qquad Cov(T_j, T_l) = \partial^2_{\eta_j \eta_l} b(\eta_0).$$

Beweis.

(i) O.E. $f \geq 0$. Als Funktion von η_1 können wir das Integral also schreiben als

$$\int e^{\eta_1 T_1(x)}\, \tilde{\mu}(dx) \qquad \text{mit} \qquad \tilde{\mu} = fhe^{\langle \eta_{-1}, T_{-1}\rangle}$$

Für $|h| < \delta$ hinreichend klein erhalten wir

$$\Big|\frac{e^{(\eta_1+h)T_1} - e^{\eta_1 T_1}}{h}\Big| = e^{\eta_1 T_1}\Big|\frac{e^{hT_1} - 1}{h}\Big| \leq \frac{1}{\delta} e^{\eta_1 T_1} e^{\delta |T_1|} \in L^1(\tilde{\mu}).$$

Damit folgt (i) für die ersten partiellen Ableitungen aus dominierter Konvergenz. Für höhere Ableitungen erhalten wir das Ergebnis, wenn wir $\tilde{f} = fT$ setzen.

(ii) Folgt wie in (i).

(iii) Folgt ebenfalls aus (i).

□

Corollar 3.8 (Exponentialfamilien sind regulär)**.** *Sei ein statistisches Modell, wobei $(\mathbb{P}_\eta)_{\eta \in H}$ eine natürliche Exponentialfamilie mit offenem $H \subset \mathbb{R}^k$ ist.*

(i) $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\eta)_{\eta \in H})$ ist regulär und T ist ein regulärer Schätzer.

(ii) T ist komponentenweise Cramér-Rao effizient für $\nabla b(\theta)$.

Beweis.

(i) Nach Voraussetzung ist H offen. Satz 3.7 liefert die Vertauschungsrelation für $h := 1$ und $h := T$. Die Likelihood-Funktion ist per Definition strikt positiv und nach η stetig differenzierbar. Für die Score-Funktion erhalten wir o.E.

$$U_\eta = \nabla_\eta \ln L(\eta, x) = \nabla_\eta(\langle T, \eta\rangle - b(\eta)) = T - \nabla b(\eta),$$

d.h. $I(\eta) = \mathrm{Cov}_\eta(T) = \nabla^2 b(\eta)$. O.E. ist diese Matrix positiv definit. Andernfall kann eine Komponente von T als Linearkombination der anderen plus einer konstanten geschrieben werden, d.h. die Exponentialfamilie könnte um einen Parameter reduziert werden.

(ii) Wir erhalten sofort

$$\mathrm{Var}_\eta(T_j) = (\nabla^2 b(\eta))_{jj} = \nabla(\nabla b(\eta)_j)^\top \nabla^2 b(\eta)^{-1} \nabla(\nabla b(\eta)_j) == \nabla(\nabla b(\eta)_j)^\top I(\eta)^{-1} \nabla(\nabla b(\eta)_j),$$

was die Behauptung beweist.

□

Corollar 3.9 (MLE für $b(\eta)$)**.** *In der Situation von Satz 3.7 (iii) ist T ein erwartungstreuer Schätzer von $\nabla b(\eta_0)$ mit Kovarianz $\nabla^2 b(\eta_0)$. T ist auch ein ML-Schätzer, sofern dieser in $\mathring{H}$ existiert.*

Beweis. Es bleibt die ML-Eigenschaft zu zeigen.

$$\ell(\eta, x) = \ln L(\eta, x) = \ln \frac{d\mathbb{P}_\eta}{d\mu}(x) = \ln h(x) + \langle \eta, T(x)\rangle - b(\eta)$$

Existiert ein ML-Schätzer in $\mathring{H}$, so muss $\nabla \ell(\widehat{\eta}) = 0$ gelten, d.h.

$$T(x) - \nabla b(\widehat{\eta}) = 0 \Rightarrow T(x) = \nabla b(\widehat{\eta}) = \mathbb{E}_{\widehat{\eta}} T.$$

NOCHMAL NACHFRAGEN. □

3.2 Verallgemeinerte lineare Modelle

Wir wollen das lineare Modell verallgemeinern, um insbesondere auch diskrete Beobachtungen und nicht approximativ normalverteilte Beobachtungen adäquat modellieren zu können.

Definition 3.10 (Verallgemeinertes lineares Modell)**.** Die Beobachtungen $Y_1, \dots Y_n$ folgen einem *verallgemeinerten linearen Modell* (GLM), falls sie unabhängig sind und die Randverteilung einer natürlichen Exponentialfamilie

$$\frac{d\mathbb{P}_{\eta_i}^{Y_i}}{d\mu}(y_i) = h(y_i, \varphi) \exp\Big(\frac{\eta_i y_i - b(\eta_i)}{\varphi}\Big), \qquad i = 1, \dots, n$$

folgen für einen unbekannten *Dispersionsparameter* $\varphi > 0$,

$$\eta_i \in H = \big\{\eta \in \mathbb{R} : \int h(y, \varphi) \exp\Big(\frac{y\eta}{\varphi}\Big)\, \mu(dy) \in (0, \infty)\big\}$$

und $b''(\eta) > 0$ für alle $\eta \in \mathring{H}$. Weiter muss gelten, dass für $\varrho(\eta_i) := \mathbb{E}_{\eta_i} Y_i$ und einen Parameter $\beta \in \mathbb{R}^k$, eine Designmatrix $X \in \mathbb{R}^{n\times k}$ mit $\operatorname{rk} X = k \le n$ und eine bijektive stetig differenzierbare Funktion $g : \mathbb{R} \to \mathbb{R}$

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

g heißt *Linkfunktion*. Falls $\varrho = g^{-1}$ sprechen wir von der *kanonischen Linkfunktion* und es gilt $(\eta_1, \dots \eta_n)^\top = X\beta$.

Beispiel 3.11 (Das lineare Modell als verallgemeinertes lineares Modell). Im Fall des linearen Modells $Y = X\beta + \varepsilon$ mit $\operatorname{rk} X = p \le n, \varepsilon \sim N(0, \sigma^2 I_n), \beta \in \mathbb{R}^p$ und $\sigma^2 > 0$ bekannt gilt

$$\frac{d\mathbb{P}^{Y_i}}{d\lambda}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(\frac{-(y_i - (X\beta)_i)^2}{2\sigma^2}\Big) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-y_i^2/(2\sigma^2)} \exp\Big(\frac{(X\beta)_i y_i - (X\beta)_i^2/2}{\sigma^2}\Big), \qquad i = 1, \dots, n$$

mit $\eta_i := (X\beta)_i \in \mathbb{R}$. Weiter gilt

$$b''(\eta_i) := \partial_{\eta_i} \eta_i = 1 > 0.$$

Damit ist das lineare Modell ein verallgemeinertes lineares Modell mit kanonischer Linkfunktion und Dispersionsparameter $\sigma^2 > 0$.

Satz 3.12 (MLE im verallgemeinerten linearen Modell)**.** *Im verallgemeinerten linearen Modell mit kanonischer Linkfunktion und den Zeilen x_i der Designmatrix X gilt für die Loglikelihoodfunktion*

$$\nabla_\beta \ell(\beta, \varphi) = \frac{1}{\varphi} \sum_{i=1}^n \big(Y_i - b'(\langle x_i, \beta\rangle)\big) x_i \in \mathbb{R}^p.$$

Ist die Fisher-Informationsmatrix

$$I(\beta) = \frac{1}{\varphi} \sum_{i=1}^n b''(\langle x_i, \beta\rangle) x_i x_i^\top \in \mathbb{R}^{p\times p}$$

positiv definit für alle $\beta \in \mathbb{R}^p$ und erfüllt ein $\widehat{\beta} \in \mathbb{R}^p$, dass $\nabla_\beta \ell(\widehat{\beta}, \varphi) = 0$,, so ist $\widehat{\beta}$ der eindeutig bestimmte ML-Schätzer von β.

Beweis. Es gilt

$$\nabla_\beta \ell(\beta, \varphi, Y) = \frac{1}{\varphi} \nabla_\beta \sum_{i=1}^n \big(\eta_i Y_i - b(\eta_i)\big) = \frac{1}{\varphi} \sum_{i=1}^n \big(Y_i - b'(\langle x_i, \beta\rangle)\big) x_i.$$

Für die Fisher-Information erhalten wir wegen $\mathbb{E}_\beta Y_i = b'(\langle x_i, \beta\rangle)$ und $\operatorname{Var}_\beta(Y_i) = \varphi b''(\langle x_i, \beta\rangle)$, dass

$$I(\beta) = \mathbb{E}_\beta\big(\nabla_\beta \ell(\beta, \varphi, Y) \nabla_\beta \ell(\beta, \varphi, Y)^\top\big) = \frac{1}{\varphi^2} \sum_{i=1}^n \operatorname{Var}_\beta(Y_i) x_i x_i^\top = \frac{1}{\varphi} \sum_{i=1}^n b''(\langle x_i, \beta\rangle) x_i x_i^\top.$$

Man rechnet außerdem nach, dass $-I(\beta)$ die Hessematrix der Loglikelihoodfunktion ist. Sofern diese positiv definit ist für alle β ist die Loglikelihoodfunktion strikt konkav und ein stationärer Punkt ist bereits ein eindeutiges globales Maximum. □

In der Situation von Satz 3.12 kann der ML-Schätzer für β oft nicht symbolisch berechnet werden. Durch die positive Definitheit der Fisher-Information können wir diesen jedoch durch konsekutive Newton-Schritte approximieren. Im allgemeinen sucht das Newton-Verfahren die Nullstelle einer Funktione $F \in C^1(\mathbb{R}^n, \mathbb{R}_m)$, indem man ihre Linearisierung null setzt, d.h. man setzt

$$\begin{aligned} &F(x_0) + DF(x_0)(x - x_0) = 0 \\ \Leftrightarrow &x = x_0 - DF(x_0)^{-1} F(x_0) \end{aligned}$$

sofern alle Terme wohldefiniert sind. Im Fall des verallgemeinerten linearen Modells führt das auf *Fishers Scoring Methode*.

Algorithmus 3.13 (Fishers Scoring Methode)**.** *In der obigen Situation setzen wir* $F(\beta) := \nabla\ell(\beta)$ *mit* $DF(\beta) = \nabla^2\ell(\beta) = -I(\beta)$*. Wir setzen dann* $\widehat{\beta}_0 := 0$ *und erhalten iterativ für* $k = 1, \dots$

$$\begin{aligned}\widehat{\beta}^{k+1} &= \widehat{\beta}^k + I(\widehat{\beta}^k)^{-1}\nabla_\beta\ell(\widehat{\beta}^k, \varphi, Y)\\ &= \widehat{\beta}^k + \Big(\frac{1}{\varphi}\sum_{i=1}^n b''(\langle x_i, \widehat{\beta}^k\rangle)x_i x_i^\top\Big)^{-1}\frac{1}{\varphi}\sum_{i=1}^n \big(Y_i - b'(\langle x_i, \widehat{\beta}^k\rangle)\big)x_i\\ &= \widehat{\beta}^k + \Big(\sum_{i=1}^n b''(\langle x_i, \widehat{\beta}^k\rangle)x_i x_i^\top\Big)^{-1}\sum_{i=1}^n \big(Y_i - b'(\langle x_i, \widehat{\beta}^k\rangle)\big)x_i.\end{aligned}$$

Als Anwendung eines verallgemeinerten linearen Modells betrachten wir die *Logistische Regression* genauer.

Definition 3.14 (Logistische Regression)**.** Ein verallgemeinertes lineares Modell auf $(\{0,1\}^n, \mathscr{P}(\{0,1\})^{\otimes n})$ heißt *logistische Regression* falls $Y_i \sim \mathrm{Ber}(p_i), i = 1, \dots, n$ i.i.d. und die Linkfunktion g kanonisch ist. Wegen $\eta_i = \ln\frac{p_i}{1-p_i}$ und $\varrho(\eta_i) = \mathbb{E}Y_i = p_i$ ist die kanonische Linkfuntion gerade

$$g : (0,1) \to \mathbb{R}, \qquad g(p) := \ln\frac{p}{1-p}.$$

und wir betrachten das Modell

$$\eta = \Big(\ln\frac{p_1}{1-p_1}, \dots, \ln\frac{p_n}{1-p_n}\Big)^\top = X\beta.$$

Die Linkfuntion g heißt *Logit-Funktion*. Ihre Umkehrfunktion $g^{-1} : \mathbb{R} \to (0,1), g^{-1}(\eta) = e^\eta/(1+e^\eta)$ heißt *logistische Funktion*.

Beispiel 3.15 (Prostatakrebs). Wenn bei Patienten Prostatakrebs diagnostiziert wird ist es von höchstem Interesse festzustellen, ob der Krebs auf die umliegenden Lymphknoten gestreut hat. Man versucht dies durch verschiedenen Kovariablen vorherzusagen:

1. `aged` - Ist der Patient jünger als 60 Jahre.
2. `stage` - Das Ergebnis einer groben Messung deutet auf ernsteres Stadium hin oder nicht.
3. `grade` - Eine Biopsi mittels Spritze deutet auf ein ernsteres Stadium hin.
4. `xray` - Röntgenaufnahme deutet auf ein ernsteres Stadium hin.
5. `acid` - Anteil von Phosphaten im Blut hat Schwellenwert überschritten.

Es stellt sich heraus, dass die Variablen `acid` und `xray` ausschlaggebend sind. (Vergleiche R script).

4 Klassifikation

4.1 Logistische Regression, KNN und LDA

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 es, 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).

Beispiel 4.1 (`Default`-Datensatz). Wir betrachten den Datensatz `Default` aus dem `R`-Paket `ISLR`. Wir wollen auf Grundlage vom durchschnittlichen monatlichen Kontostand der Kreditkarte `balance`, dem Jahreseinkommen `income` und dem Studentenstatus `student` vorhersagen, ob eine Person zahlungsunfähig wird `default` oder nicht.

Abbildung 1: Illustration der ersten 1000 Einträge des `default`-Datensatzes. Rot bedeutet `default`

Definition 4.2 (Klassifizierer)**.**

(a) Seien $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, \dots, K\}$ Jede messbare Funktion $C : \mathbb{R}^d \to \{1, \dots, K\}$ heißt *Klassifizierer*. Der *Klassifizierungsfehler* ist definiert als

$$R(C) := \mathbb{P}\{C(X) \neq Y\} = \mathbb{E}\mathbf{1}_{\{C(X)\neq Y\}}.$$

Beachte, dass der Erwartungwert bezüglich einer neuen Beobachtung X, Y gebildet wird.

(b) Im Falle, dass die Menge der Label durch $\{0,1\}$ gegeben ist erhalten wir

$$R(C) := \mathbb{P}\{C(X) \neq Y\} = \mathbb{E}(Y - C(X))^2 = \int \mathbf{1}_{\{y\neq C(x)\}}\,\mathbb{P}^{X,Y}(d(x,y)).$$

Sind alle theoretischen Größen bekannt, lässt sich ein Klassifikationsproblem optimal lösen.

Satz 4.3 (**Bayes-Klassifizierer**)**.**

(i) In der Situation von Definition 4.2 wird der Klassifizierungsfehler minimiert durch den Bayes-klassifizierer

$$C^{Bayes}(x) := \operatorname*{argmax}_{k=1,\dots,K} \mathbb{P}\{Y = k|X = x\}.$$

(ii) Sind die Label gegeben durch $\{0, 1\}$*, so gilt*

$$C^{Bayes}(x) := \mathbf{1}\Big(\eta(x) \geq \frac{1}{2}\Big) \qquad \textit{mit} \qquad \eta(x) := \mathbb{P}\{Y = 1|X = x\}.$$

Beweis. Für (i) erhalten wir durch Bedingen auf X für einen beliebigen Klassifizierer C, dass

$$R(C) = 1 - \mathbb{E}\mathbb{E}(\mathbf{1}_{C(X)=Y}|X) = 1 - \mathbb{E}\sum_{k=1}^{K}\mathbb{E}(\mathbf{1}_{C(X)=k}\mathbf{1}_{Y=k}|X) = 1 - \mathbb{E}\sum_{k=1}^{K}\mathbf{1}_{C(X)=k}\mathbb{P}\{Y = k|X\},$$

was die Optimalität des Bayes-Klassifizierers beweist. (ii) folgt sofort. □

Beispiel 4.4 (Entscheidungsgrenze durch Bayesklassifikation). Seien $\mathbb{P}(Y = 1) = 1/2 = \mathbb{P}(Y = 0)$ und

$$\mathbb{P}^{X|Y=0} = N(\mu_0, 1), \qquad \mathbb{P}^{X|Y=0} = N(\mu_1, 1).$$

Abbildung 2:

In der Notation von Satz 4.3 (ii) folgt mit Bayes Formel, dass

$$\eta(x) = \mathbb{P}\{Y = 1|X = x\} = \frac{f_1(x)\mathbb{P}\{Y = 1\}}{\frac{1}{2}(f_0(x) + f_1(x))} = \frac{f_\mu(x)}{f_\mu(x) + f_{\tilde{\mu}}(x)}$$

Also ist der Bayesklassifizierer gegeben durch $C^{\text{Bayes}}(x) = \mathbf{1}_{\eta(x)\geq 1/2} = \mathbf{1}_{f_1(x)\geq f_0(x)}$.

Auf Grund der Optimalität wird der Bayes-Klassifizierer wird auch Orakel-Klassifizierer genannt. Für jeden anderen Klassifizierer definieren wir entsprechend das *Exzess-Risiko.*

Definition 4.5 (Exzess-Risiko)**.** Für jeden Klassifizierer C heißt

$$\mathcal{E}(C) := \mathbb{P}\{C(X) \neq Y\} - \mathbb{P}\{C^{\text{Bayes}}(X) \neq Y\} \geq 0$$

Excess-Verlust/Risiko.

In der Realität können wir den Bayes-Klassifizierer natürlich nicht nutzen. Wir betrachten Methoden die versuchen das Verhalten des Bayes-Klassifizierers auf Basis der empirischen Daten zu immitieren.

Logistische Regression: In der Situation von Definition 3.14 können wir den Schätzer $\widehat{\beta}$ zur Klassifikation benutzen.

Definition 4.6 (Klassifikation mittels logistischer Regression)**.** In der Situtation von Definition 3.14 können wir nach Schätzung des Parametervektors $\widehat{\beta}$ auf der Trainingsmenge $(x_1, Y_1), \dots, (x_n, Y_n)$ für eine jede neue Kovariablenrealisierung x_{n+1} einen zugehörigen Wert

$$\widehat{p}_{n+1} = p(x_{n+1}, \widehat{\beta}) = \frac{\exp(\langle x_{n+1}, \widehat{\beta}\rangle)}{1 + \exp(\langle x_{n+1}, \widehat{\beta}\rangle)}$$

vorhersagen und klassifizieren dann entsprechend $\widehat{C}^{\text{Log}}(x) := \mathbf{1}\{p(x, \widehat{\beta}) \geq \tau\}$ für einen Schwellenwert $\tau \in [0, 1]$. Dieser wird entsprechend der Risikoaversion gewählt.

Bemerkung 4.7 (Probabilitische Eigenschaften logistische Regression).

(a) Der ML-Schätzer $\widehat{\beta}$ in der logistischen Regression erfüllt

$$\nabla_\beta \ell(\widehat{\beta}, y) = \sum_{i=1}^{n}(y_i - p(x_{i\cdot}, \widehat{\beta}))x_{i\cdot} = 0.$$

Wenn der erste Koeffizient von $x_{i\cdot}$ gleich 1 ist (Intercept) folgt $\sum_{i=1}^n y_i = \sum_{i=1}^n p(x_{i\cdot}, \widehat{\beta})$, d.h. die beobachtete Anzahl von Werte $y_i = 1$ entspricht der Anzahl, die wir gegeben des geschätzten Modells erwarten würden.

(b) Inferenz im logistischen Modell beruht auf asymptotischen Überlegungen, auf die wir hier nicht weiter eingehen.

K-Nächste-Nachbarn: Approximieren wir in der Situation von Definition 4.2 $\mathbb{P}\{Y = j|X = x\} \approx \frac{1}{K}\sum_{X_i\in N_K(x)}\mathbf{1}_{Y_i=j}$, wobei $N_K(x)$ die Menge der K nächsten Nachbarn ist, so erhalten wir den KNN-Klassifizierer.

Definition 4.8 (KNN-Klassifizierer)**.**

(a) Gegeben sei $K \in \mathbb{N}$, ein Trainingsdatensatz $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, \ldots, J\}$ von i.i.d. Daten. Für $x \in \mathbb{R}^d$ sei die Menge der K-nächste Nachbarn bezüglich der euklidischen Norm definiert als $N_j(x)$. Der KNN-Klassifizierer ist jetzt gegeben durch

$$\widehat{C}^{\mathrm{KNN}}(x) := \operatorname*{argmax}_{j=1,\ldots,J} \frac{1}{K} \sum_{X_i \in N_K(x)} \mathbf{1}_{Y_i = j}$$

(b) Im Spezialfall, dass die Y_i Werte in $\{0, 1\}$ annehmen erhalten wir

$$\widehat{C}^{\mathrm{KNN}}(x) := \mathbf{1}\{\widehat{\eta}(x) \geq 1/2\} \qquad \text{mit} \qquad \widehat{\eta}(x) := \frac{1}{K} \sum_{X_i \in N_K(x)} \mathbf{1}_{Y_i = 1} =: \sum_{i=1}^{n} w_i(x) Y_i,$$

wobei $w_i := \mathbf{1}\{X_i \in N_K(x)\}/K$ mit $\sum_{i=1}^n w_i = 1$.

Bemerkung 4.9 (Wahl von K)*.*

(a) Obwohl das KNN-Verfahren einen recht einfachen Ansatz verfolgt, können die erreichten KLassifizierungen bei Tests oft sehr gute Ergebnisse erzielen.

(b) Die Wahl von $K \in \mathbb{N}$ hat einen drastischen Effekt auf die erzielte Güte der Klassifizierung. Es bietet sich an für mehrere K, die in-sample und out-of-sample performance zu betrachten und ein dementsprechend optimales K zu wählen.

Zur Untersuchung der Eigenschaften des KNN-Klassifizierers beschränken wir uns im Folgenden auf den Fall (b) aus obiger Definition.

Lemma 4.10 (Rückführung auf die Regressionsfunktion)**.** *In der Situation von Definition 4.8 (b) gilt für den Bayes-Klassifizierer C^{Bayes}, dass*

$$|\mathbb{E}_{\leq n} R(\widehat{C}^{KNN}) - R(C^{Bayes})| \leq \sqrt{\mathbb{E}_{\leq n+1} |\widehat{\eta}(X) - \eta(X)|^2}.$$

Beweis. Es gilt für jeden Klassifizierer C

$$\mathbb{P}\{C(X) = Y | X\} = \mathbf{1}_{C=1}\eta + \mathbf{1}_{C=0}(1-\eta) = \eta + \mathbf{1}_{C=0}(1 - 2\eta).$$

Damit erhalten wir

$$\begin{aligned}
|\mathbb{P}\{\widehat{C}^{\mathrm{KNN}}(X) \neq Y | X\} - \mathbb{P}\{\widehat{C}^{\mathrm{Bayes}}(X) \neq Y | X\}| &= |\mathbb{P}\{\widehat{C}^{\mathrm{KNN}}(X) = Y | X\} - \mathbb{P}\{\widehat{C}^{\mathrm{Bayes}}(X) = Y | X\}| \\
&= |\widehat{\eta} + \mathbf{1}_{\widehat{C}^{\mathrm{KNN}}=0}(1 - 2\widehat{\eta}) - \eta - \mathbf{1}_{C^{\mathrm{Bayes}}=0}(1 - 2\eta)| \\
&= \begin{cases} |\widehat{\eta} - \eta| & , \widehat{C}^{\mathrm{KNN}} = C^{\mathrm{Bayes}} \\ |1 - \widehat{\eta} - \eta| & , \widehat{C}^{\mathrm{KNN}} \neq C^{\mathrm{Bayes}} \end{cases} \\
&\leq \begin{cases} |\widehat{\eta} - \eta| & , \widehat{C}^{\mathrm{KNN}} = C^{\mathrm{Bayes}} \\ |1/2 - \widehat{\eta} \wedge \eta| & , \widehat{C}^{\mathrm{KNN}} \neq C^{\mathrm{Bayes}} \end{cases} \leq |\widehat{\eta} - \eta|.
\end{aligned}$$

und damit durch Bedingen auf X und Jensens Ungleichung

$$|\mathbb{E}_{\leq n} R(\widehat{C}^{\mathrm{KNN}}) - R(C^{\mathrm{Bayes}})|^2 = |\mathbb{E}_{\leq n+1}\big(\mathbf{1}_{\widehat{C}^{\mathrm{KNN}} \neq Y} - \mathbf{1}_{C^{\mathrm{Bayes}} \neq Y}\big)|^2 \leq \mathbb{E}_{\leq n+1} |\widehat{\eta} - \eta|^2.$$

□

Satz 4.11 (**Konsistenz des KNN-Klassifizierers**)**.** *In der Situation von Definition 4.8 (b) gelte $k \to \infty, k/n \to 0$ und $x \mapsto \eta(x)$ sei gleichmäßig stetig. Dann ist der KNN-Klassifizierer $\widehat{C}^{KNN}$ konsistent, d.h.*

$$|\mathbb{E}_{\leq n} R(\widehat{C}) - R(C_B)| \xrightarrow{n \to \infty} 0.$$

Beweis. Aus Lemma 4.10 erhalten wir mit Hilfe der Dreiecksungleichung

$$\begin{aligned}
|\mathbb{E}_{\leq n} R(\widehat{C}) - R(C_B)| \leq \sqrt{\mathbb{E}_{\leq n+1} |\widehat{\eta}(X) - \eta(X)|^2} &= \sqrt{\mathbb{E}_{\leq n+1} |\sum_{i=1}^n w_i(X)\big(Y_i - \eta(X)\big)|^2} \\
&\leq \sqrt{\mathbb{E}_{\leq n+1} |\sum_{i=1}^n w_i(X)\big(Y_i - \eta(X_i)\big)|^2} + \sqrt{\mathbb{E}_{\leq n+1} |\sum_{i=1}^n w_i(X)\big(\eta(X_i) - \eta(X)\big)|^2}.
\end{aligned}$$

Im Folgdenden betrachten wir die beiden Terme einzeln.

Für den ersten Term gilt wegen der Unabhängigkeit

$$\begin{aligned}
\mathbb{E}_{\leq n+1} |\sum_{i=1}^n w_i(X)\big(Y_i - \eta(X_i)\big)|^2 &= \sum_{i=1}^n \mathbb{E}_{\leq n+1} w_i(X)^2 \big(Y_i - \eta(X_i)\big)^2 \\
&\leq \mathbb{E}_{\leq n+1}\Big(\max_{i \leq n} w_i(X) \sum_{i=1}^n w_i(X)\Big) \leq \frac{1}{K} \xrightarrow{n \to \infty} 0.
\end{aligned}$$

Für den zweiten Term erhalten wir für beliebiges $\varepsilon > 0$ und dazu gehöriges $\delta > 0$, dass

$$\mathbb{E}_{\leq n+1}|\sum_{i=1}^{n} w_i(X)\big(\eta(X_i) - \eta(X)\big)|^2 \leq \mathbb{E}_{\leq n+1}|\sum_{i=1}^{n} w_i(X)\mathbf{1}_{|X_i - X|\geq\delta}| + \varepsilon$$
$$\leq \mathbb{E}_{n+1}\frac{1}{K}\sum_{i=1}^{K}\mathbf{1}\Big\{\sum_{l=1}^{n}\mathbf{1}_{|X_l - X|<\delta} \leq i\Big\} + \varepsilon \leq \mathbb{P}\Big\{\sum_{l=1}^{n}\mathbf{1}_{|X_l - X|<\delta} \leq K/n\Big\} + \varepsilon,$$

da im Fall, dass der i-nächste Nachbar einen Abstand von mehr als δ zu X besitzt höchstens i Beobachtungen so nahe an X liegen (O.E. X_i ordnen und von 1 bis K durchgehen). Daraus folgt die Behauptung, da $k/n \to 0$. □

Lineare Diskriminanzanalyse: Als dritte Klassifikationsmethode betrachten wir die linearen Diskriminanzanalyse (LDA). Wir nehmen an, dass in der Situation von Definition 4.2 die Verteilung von $X|Y = k$ eine Dichte f_k bezüglich einem Maß μ besitzt und wir setzen $\pi_k := \mathbb{P}\{Y = k\}, k = 1, \ldots, K$. Dann besitzt X die μ-Dichte

$$f(x) = \sum_{k=1}^{K}\pi_k f_k(x), \qquad x \in \mathbb{R}^d.$$

Man sagt, f sei Mischung der Verteilungen $f_1, \ldots, f_K$. In der Tat gilt

$$\int_A f\,d\mu = \sum_{k=1}^{K}\pi_k\int_A f_k\,d\mu = \sum_k \mathbb{P}\{Y = k\}\mathbb{P}\{X \in A|Y = k\} = \mathbb{P}\{X \in A\}, \qquad A \in \mathscr{B}_{\mathbb{R}^d}.$$

Genauer gilt, dass $\mathbb{P}$ die Dichte $p(x, k) = \pi_k f_k(x)$ besitzt. Unter diesen Voraussetzungen ist der Bayes-Klassifizierer gegeben durch

$$C^{\text{Bayes}}(x) \in \operatorname*{argmax}_{k=1,\ldots K}\mathbb{P}\{Y = k|X = x\} = \operatorname*{argmax}_{1\leq k\leq K}\frac{f_k(x)\pi_k}{f(x)} = \operatorname*{argmax}_{1\leq k\leq K} f_k(x)\pi_k.$$

Für die LDA ist unsere Modellannahme, dass die Dichten (f_k) durch Normalverteilungsdichten mit Mittelwertvektoren $(\mu_k) \subset \mathbb{R}^d$ und gemeinsame positiv definite Kovarianzmatrix $\Sigma \in \mathbb{R}^{d\times d}$ gegeben sind.

Abbildung 3: Mischung f für $d = 1$.

Angenommen man kennt das Modell, erhält man

$$C^{\text{Bayes}} \in \operatorname*{argmax}_{1\leq k\leq K}\ln(f_k\pi_k) = \operatorname*{argmax}_{1\leq k\leq K} -\ln\sqrt{(2\pi)^d|\det\Sigma|} - \frac{1}{2}\langle x - \mu_k, \Sigma^{-1}(x - \mu_k)\rangle + \ln\pi_k$$
$$= \operatorname*{argmax}_{k=1,\ldots,K}\langle\Sigma^{-1}\mu_k, x\rangle - \frac{1}{2}\langle\Sigma^{-1}\mu_k, \mu_k\rangle + \ln\pi_k =: \delta_k(x).$$

Die Diskriminante $\delta_k(x)$ ist dabei linear in x. Für die Decision boundary zwischen den Klassen k und l erhalten wir

$$\{x : \delta_k(x) = \delta_l(x)\} = \{x : \langle\Sigma^{-1}(\mu_l - \mu_k), x\rangle - \langle\Sigma^{-1}(\mu_l - \mu_k), \frac{\mu_l + \mu_k}{2}\rangle + \ln\pi_l/\pi_k = 0\},$$

d.h. die decision boundary ist eine Hyperebene im $\mathbb{R}^d$, orthogonal zu $\mu_k - \mu_l$ bzgl. $\langle\Sigma^{-1}.,.\rangle$.

Abbildung 4: Decision boundary $\{x : \delta_1(x) = \delta_2(k)\}$

Das statistische Problem besteht darin, dass die $\pi_1, \ldots, \pi_k$, $f_1, \ldots f_K$ geschätzt werden müsssen.

Definition 4.12 (LDA-Klassifizierer)**.** Es seien $(x_1, y_1), \ldots, (x_n, y_n)$ Daten in $\mathbb{R}^d \times \{1, \ldots, K\}$. Setze $n_k := |\{i : y_i = k\}|$ sowie

$$\widehat{\mu}_k := \frac{1}{n_k}\sum_{i=1}^{n} x_i\mathbf{1}_{\{y_i=k\}} = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \qquad \widehat{\pi}_k := \frac{n_k}{n} \qquad \text{und} \qquad \widehat{\Sigma} := \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^\top.$$

Dann heißt

$$\widehat{C}^{\text{LDA}}(x) \in \operatorname*{argmax}_{k=1,\ldots,K}\{\langle\widehat{\Sigma}^{-1}\widehat{\mu}_k, x\rangle - \frac{1}{2}\langle\widehat{\Sigma}^{-1}\widehat{\mu}_k, \widehat{\mu}_k\rangle + \ln\widehat{\pi}_k\}$$

LDA-Klassifizierer.

Bemerkung 4.13 (QDA und Vergleich mit Logistischer Regression).

(a) Eine weitere Verallgemeinerung stellt die quadratische Diskriminanzanalyse (QDA) dar, bei der jede Klasse k eine eigene im allgemeinen unterschiedliche Kovarianzmatrix Σ_k besitzt. Dies führt zu einer quadratischen Klassifizierungsregel.

(b) Im Fall, dass die Label durch $\{0,1\}$ gegeben sind Klassifizieren LDA und Logistische Regression qualitativ identisch, nur die Schätzmethoden sind unterschiedlich. In dieser Situation hängt die Entscheidung nur von den den Log-odds ab. Für die Logistische Regression haben wir per Konstruktionem

$$\ln \frac{p^{\mathrm{Log}}(x)}{1-p^{\mathrm{Log}}(x)} = \langle x, \beta \rangle$$

Bei der LDA haben wir bereits gesehen, dass die Decision boundary linear ist. Es gilt aber nochmal

$$\ln \frac{p^{\mathrm{LDA}}(x)}{1-p^{\mathrm{LDA}}(x)} = \ln \frac{\widehat{f}_1(x)\pi_1}{\widehat{f}_0(x)(1-\pi_1)} = -\|\widehat{\Sigma}^{-1}(x-\widehat{\mu}_1)\|^2 + \|\widehat{\Sigma}^{-1}(x-\widehat{\mu}_0)\|^2 + c = \langle x, v\rangle + c'.$$

Wir stellen uns die Frage, wie groß der Exzess-Verlust $\mathcal{E}(\widehat{C}^{LDA})$ des LDA-Klassifizierers ist. Genauer: Angenommen, $(X_1,Y_1),\dots(X_n,Y_n)$ ist eine mathematische Stichprobe gemäß dem LDA Modell. Was kann man dann über $\mathbb{E}\mathcal{E}(\widehat{C}_n^{LDA})$ aussagen, wo jetzt jetzt $\widehat{\pi}_k, \widehat{\mu}_k, \widehat{\Sigma}$ Schätzer sind? An dieser Stelle können wir diese Frage nicht abschließend beantworten. Wir betrachten aber ein „toy example“, das uns die richtige Intuition gibt.

Beispiel 4.14 (Toy example). Wir betrachten den Fall zweier eindimensionaler Gaußdichten mit Erwartungswerten $\mu_0 < \mu_1$ und Varianz $\sigma^2 > 0$. Der LDA-Klassifizierer betrachtet

$$\widehat{f}_0 = g_{\widehat{\mu}_0,\widehat{\sigma}^2}, \qquad \widehat{f}_1 = \varphi_{\widehat{\mu}_1,\widehat{\sigma}^2}, \qquad \widehat{\pi}_1 = n_1/n$$

und klassifiziert entsprechend

$$\widehat{C}^{\mathrm{LDA}}(x) = \mathbf{1}\{x \geq \widehat{x}^*\} \qquad \text{mit} \qquad \widehat{x}^* = \frac{\widehat{\mu}_1+\widehat{\mu}_2}{2} - \frac{\widehat{\sigma}^2}{\widehat{\mu}_2-\widehat{\mu}_1}\ln\frac{\widehat{\pi}_1}{\widehat{\pi}_0}$$

Für den Bayes-Klassifizierer gilt

$$x^* = \frac{\mu_1+\mu_2}{2} - \frac{\sigma^2}{\mu_2-\mu_1}\ln\frac{\pi_2}{\pi_1}$$

unter der Annahme, dass $\widehat{\pi}_1 \in (0,1)$. Für $n\to\infty$ erwartet man, dass $|\widehat{x}^* - x^*| = O_p(n^{-1/2})$, d.h. $|\widehat{x}^* - x^*|$ ist von der Größenordnung $n^{-1/2}$ im stochastischen Mittel.

Nun gilt für $\widehat{\mu}_0 < \widehat{\mu}_1$

$$\mathcal{E}(\widehat{C}^{LDA}) = \int_{-\infty}^{\widehat{x}^*} \pi_1 f_1(x)\,dx + \int_{\widehat{x}^*}^{\infty} \pi_0 f_0(x)\,dx - \left(\int_{-\infty}^{x^*} \pi_1 f_1(x)\,dx + \int_{x^*}^{\infty} \pi_0 f_0(x)\,dx\right) = \dots$$

Im Fall $\widehat{x}^* > x^*$ folgt

$$\cdots = \int_{x^*}^{\widehat{x}^*} \pi_1 f_1(x)\,dx - \int_{x^*}^{\widehat{x}^*} \pi_0 f_0(x)\,dx = \int_{x^*}^{\widehat{x}^*} (\pi_1 f_1(x) - \pi_0 f_0(x))\,dx$$

Worauf man hinaus will: $\pi_1 f_1(x^*) = \pi_0 f_0(x^*)$, so dass der Integrand klein sein sollte für $\widehat{x}^*$ nahe bei x^*.

Betrachte nun den Exzess-Verlust als Funktion von $\widehat{x}^*$ für $\widehat{x}^*$ nahe x^* (n groß):

$$e(\widehat{x}^*) = \int_{x^*}^{\widehat{x}^*} (\pi_1 f_1(x) - \pi_0 f_0(x))\,dx$$

also $e(x^*) = 0$. Entwickeln

$$e'(x^*) = \pi_1 f_1(x^*) - \pi_0 f_0(x^*) = 0$$

und

$$e''(x^*) = \pi_2 f_2'(x^*) - \pi_1 f_1'(x^*) \overset{f_k=gauss}{=} \pi_2 f_2(x^*)\frac{\mu_2-x^*}{\sigma^2} f_2(x^*) - \frac{\mu_1-x^*}{\sigma^2} f_1(x^*) = \frac{\mu_2-\mu_1}{\sigma^2}\pi_k f_k(x^*)$$

Daraus folgt mit Taylor

$$e(\widehat{x}^*) \approx \frac{1}{2}\frac{\mu_2-\mu_1}{\sigma^2}\pi_k f_k(x^*)(\widehat{x}^*-x^*)^2$$

Benutzt man noch $\mathcal{E}(\widehat{C}) \leq 1$ (Wkeit) und Asymptotic von $\widehat{\mu}_1-\mu, \widehat{\mu}_2-\mu_2, \widehat{\sigma}^2 - s^2, \frac{n_k}{n} - \pi_k$ (nicht trivial), so erhält man

$$\mathbb{E}\mathcal{E}(\widehat{C}) = O(n^{-1})$$

ohne Beweis. Die Rate n^{-1} heißt fast rate, weil sie schneller ist als die Standardkonvergenzrate $n^{-1/2}$.

Der Klassifikationsfehler von $\widehat{C}$ is also recht klein, wenn das LDA Modell erfüllt ist.

5 Modellwahl

Zitat: „All models are wrong but some are useful“ (G. Box). Wir betrachten eine Polynomregression

Abbildung 5: Polynomregression mit Polynomen unterschiedlichen Grades.

Die Grundlegende Frage ist, welchen Polynomgrad man auswählt. Etwas allgemeiner betrachten wir im folgenden ein Regressionsmodell der Form

$$Y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \dots, n$$

mit einer unbekannten Funktion $f : \mathfrak{X} \to \mathbb{R}$ und $x_1, \dots, x_n \in \mathfrak{X}$. Den Raum dieser Funktionen statten wir aus mit der empirischen Seminorm

$$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^n f(x_i)^2.$$

Ist $e_1, \dots, e_n$ eine ONB diese Funktionenraums bezüglich $\|.\|_n$, so gilt

$$f(x_i) = \sum_{j=1}^n \langle f, e_j \rangle_n e_j(x_i) =: \sum_{j=1}^n \beta_j e_j(x_i), \qquad i = 1, \dots, n.$$

Wir können also schreiben

$$Y_i = \sum_{j=1}^n \beta_j e_j(x_i) + \varepsilon_i, \qquad i = 1, \dots, n.$$

Oft kann man annehmen, dass $\sum_{j=1}^k \beta_j e_j$ bereits eine gute Approximation von f ist für kleine k (z.B. f glatt, (e_j) orthogonale Polynome). Wir erhalten somit verschiedene Modelle mit Parameterraumdimension $k = 1, \dots, n$:

$$Y = X^{(k)} \beta^{(k)} + \varepsilon = \begin{pmatrix} e_1(x_1) & e_2(x_2) & \dots & e_k(x_1) \\ \vdots & \ddots & & \vdots \\ & & \ddots & \vdots \\ e_1(x_n) & e_2(x_n) & \dots & e_k(x_n) \end{pmatrix} \begin{pmatrix} \beta_1^{(k)} \\ \vdots \\ \beta_k^{(k)} \end{pmatrix} + \varepsilon,$$

wobei wir im Folgenden $\varepsilon \sim N(0, \sigma^2 I_k)$ annehmen. Hierbei wird also eine gewisse Ordnung der Modelle vorausgesetzt, d.h. die $\beta_j = \langle f, e_j \rangle_n$ sind für große j eher klein.

Der allgemeine Ansatz ist also, dass die Beobachtungen Y von einem wahren Modell mit Verteilung $\mathbb{P}$ (z.B. $\mathbb{P} = N(f(x_1), \dots, f(x_n))^\top, \sigma^2 I_n)$) erzeugt werden, wir aber zur Schätzung Modelle mit Verteilungen $(\mathbb{P}_\theta)_{\theta \in \Theta_k}$ für Parametermengen $\Theta_1, \dots, \Theta_k$ annehmen. Dort sei jeweils $\widehat{\theta}_k$ ein vernünftiger Schätzer und wir wollen ein Modell $k \in \{1, \dots, K\}$ auswählen und dann $\widehat{\theta}_{\widehat{k}}$ als Schätzer benutzen.

5.1 Akaike-Informationskriterium (AIC)

Im obigen Rahmenn sei jeweils $\widehat{\theta}_k$ ein ML-Schätzer in der Parametermenge $\Theta_k \subset \mathbb{R}^k$, d.h es liegt ein k-dimensionaler Parameter vor. Es gelte Außerdem $(\mathbb{P}_\theta)_{\theta \in \Theta_k} \subset (\mathbb{P}_\theta)_{\theta \in \Theta_{k+1}}, k = 1, \dots K-1$. Wir betrachten also geschachtelte Familien von Wahrscheinlichkeitsmaßen auf dem Raum $(\mathcal{Y}, \mathscr{F})$, in dem Zielraum von Y.

Abbildung 6: Wahrscheinlichkeitsmaße auf $(\mathcal{Y}, \mathscr{F})$ mit Modellen.

Aufgabe ist es ein sinnvolles Abstandsmaß zu definieren und dann den Abstand von $\mathbb{P}_{\widehat{\theta}_k}$ zum unbekannten aber wahren $\mathbb{P}$ geeignet zu schätzen.

Definition 5.1 (Kullback-Leibler-Divergenz)**.** Für Wahrscheinlichkeitsmaße $\mathbb{P}, \mathbb{Q}$ auf $(\mathfrak{X}, \mathscr{F})$ heißt

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) := \begin{cases} \int \ln \frac{d\mathbb{P}}{d\mathbb{Q}} \, d\mathbb{P} & , \mathbb{P} \ll \mathbb{Q}, \\ \infty & , \text{sonst.} \end{cases}$$

Kullback-Leibler-Divergenz von $\mathbb{P}$ bezüglich $\mathbb{Q}$.

Lemma 5.2 (Eigenschaften KLD)**.** *In der Situation von Definition 5.1 gilt*

(i) $KL(\mathbb{P}|\mathbb{Q}) \geq 0$ *und* $KL(\mathbb{P}|\mathbb{Q}) = 0$ *genau dann, wenn* $\mathbb{P} = \mathbb{Q}$.

(ii) $KL(\mathbb{P}^{\otimes n}|\mathbb{Q}^{\otimes n}) = nKL(\mathbb{P}|\mathbb{Q})$ *für* $n \in \mathbb{N}$.

(iii) Bildet $(\mathbb{P}_\theta)_{\eta\in H}$ eine natürliche Exponentialfamilie der Form

$$\frac{d\mathbb{P}_\eta}{d\mu} = \exp\big(\langle\eta, T(x)\rangle - b(\eta)\big), \qquad x \in \mathfrak{X}, \eta \in H$$

und ist $\eta_0 \in \mathring{H}$, so gilt

$$KL(\mathbb{P}_{\eta_0}|\mathbb{P}_\eta) = b(\eta) - b(\eta_0) - \langle\nabla b(\eta_0), \eta - \eta_0\rangle.$$

(iv) Im Produktmodell $(\mathfrak{X}^n, \mathscr{F}^{\otimes n}, (\mathbb{P}_\theta^{\otimes n})_{\theta\in\Theta})$ gelte $\mathbb{P}_\theta \ll \mathbb{P}_{\theta'}$ für alle $\theta, \theta' \in \Theta$. Sei $\ell(\theta, x) := \ln d\mathbb{P}_\theta/d\mathbb{P}_{\theta^*}$ für ein $\theta^* \in \Theta$. Dann gilt für alle $\theta_0 \in \Theta$, dass

$$-\sum_{i=1}^n \ell(\theta, X_i) \xrightarrow[n\to\infty]{f.s.} = KL(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta) - KL(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta^*),$$

sofern die Kullback-Leibler-Divergenzen endlich sind.

Beweis.

(i) O.E. $\mathbb{P} \ll \mathbb{Q}$. Es gilt dann

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \int \ln\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big)\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{Q} = \int f\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\mathbf{1}_{d\mathbb{P}/d\mathbb{Q}>0}\Big)\,d\mathbb{Q} \geq f(1) = 0,$$

wobei die Ungleichheit aus der Jensen'schen Ungleichung folgt, da $f(y) := y\ln y, y > 0$ strikt konvex ist und $f(0) = 0$. Gleichheit gilt genau dann, wenn $d\mathbb{P}/d\mathbb{Q} =$ konstant ist fast sicher.

(ii) Es gilt

$$\mathrm{KL}(\mathbb{P}^{\times n}|\mathbb{Q}^{\otimes n}) = \int \ln\frac{d\mathbb{P}^{\otimes n}}{d\mathbb{Q}^{\otimes n}}\,d\mathbb{P}^{\otimes n} = \sum_{i=1}^n \int \ln\frac{d\mathbb{P}}{d\mathbb{Q}}\,d\mathbb{P}^{\otimes n} = n\mathrm{KL}(\mathbb{P}|\mathbb{Q}).$$

(iii) Es gilt

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\eta_0}|\mathbb{P}_\eta) &= \mathbb{E}_{\eta_0}\ln\frac{d\mathbb{P}_{\eta_0}/d\mu}{d\mathbb{P}_\eta/d\mu} = \mathbb{E}_{\eta_0}\Big(\ln\frac{d\mathbb{P}_{\eta_0}}{d\mu} - \ln\frac{d\mathbb{P}_\eta}{d\mu}\Big)\\ &= \mathbb{E}_{\eta_0}[\langle\eta_0, T\rangle - b(\eta_0) - (\langle\eta, T\rangle - b(\eta))]\\ &= b(\eta) - b(\eta_0) - \langle\eta - \eta_0, \nabla b(\eta_0)\rangle\end{aligned}$$

nach Satz 3.7 (iii).

(iv) Es gilt

$$\mathrm{KL}(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta^*) = \int \ln\frac{d\mathbb{P}_{\theta_0}}{d\mathbb{P}_\theta} - \ln\frac{d\mathbb{P}_{\theta_0}}{d\mathbb{P}_\theta^*}\,d\mathbb{P}_{\theta_0} = \int \ln\frac{d\mathbb{P}_{\theta^*}}{d\mathbb{P}_\theta}\,d\mathbb{P}_{\theta_0} \in \mathbb{R}.$$

Damit ist $\ell(\theta, X_1) \in L^1(\mathbb{P}_{\theta_0})$ und es gilt nach dem SLLN unter $\mathbb{P}_{\theta_0}$, dass

$$-\frac{1}{n}\sum_{i=1}^n \ell(\theta, X_i) \xrightarrow[n\to\infty]{\text{f.s.}} -\mathbb{E}_{\theta_0}\ell(\theta, X_1) = \mathrm{KL}(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{P}_{\theta_0}|\mathbb{P}_\theta^*).$$

□

Corollar 5.3 (KLD im linearen Modell)**.** *Für $\Sigma \in \mathbb{R}^{d\times d}$ symmetrisch und positiv definit und $\theta_0, \theta \in \mathbb{R}^d$ gilt*

$$KL(N(\mu_0, \Sigma)|N(\mu, \Sigma)) = \frac{1}{2}\|\Sigma^{-1/2}(\mu - \mu_0)\|^2 = \frac{1}{2}\langle\Sigma^{-1}(\mu - \mu_0), \mu - \mu_0\rangle.$$

Beweis. Für die entsprechenden Gaußdichten g_0, g gilt nach Ausmultiplizieren der Skalarprodukte und berechnen des Erwartungswertes

$$\int \ln\frac{g_0}{g}g_0\,\lambda = \int -\frac{1}{2}\big(\|\Sigma^{-1/2}(x - \mu_0)\|^2 - \|\Sigma^{-1/2}(x - \mu)\|^2\big)g_0\,d\lambda = \frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu - \mu_0)\|^2.$$

□

Bemerkung 5.4 (Natürlicher Limes des MLE). In der Situation von Lemma 5.2 sei das Modell missspezifiziert und die wahre Verteilung sei $\mathbb{P}$ mit $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\theta) < \infty\theta \in \Theta$. Wir erhalten auf dieselbe Weise unter $\mathbb{P}^{\otimes\mathbb{N}}$, dass

$$-\sum_{i=1}^n \ell(\theta, X_i) \xrightarrow[n\to\infty]{\text{f.s.}} = \mathrm{KL}(\mathbb{P}|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{P}|\mathbb{P}_\theta^*).$$

Die Rechte Seite wird also minimiert durch den Parameter der $\mathrm{KL}(\mathbb{P}|\mathbb{P}_\theta)$ minimiert. Dieser Wert stellt einen natürlichen Kandidaten für einen Limes des ML-Schätzers dar.

Idee: Wir minimieren $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\theta}_k})$. Beachte, dass $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\theta}_k})$ unbekannt ist, da $\mathbb{P}$ unbekannt ist. Wir haben aber Beobachtungen, die von $\mathbb{P}$ generiert werden. Wir wollen daher $\mathrm{KL}(\mathbb{P}|\mathbb{P}_{\widehat{\theta}_k})$ schätzen. Dazu seien alle $(\mathbb{P}_\theta)_{\theta\in\Theta_k}, k = 1,\dots,K$ und auch $\mathbb{P}$ selbst durch ein Maß μ dominiert, so dass insbesondere die Likelihoodfunktion

$$L(\theta) = \frac{d\mathbb{P}_\theta}{d\mu}, \qquad \theta \in \bigcup_{k=1}^K \Theta_k$$

existiert. Es gilt also

$$\mathrm{KL}(\mathbb{P}|\mathbb{P}_\theta) = \int \ln \frac{d\mathbb{P}/d\mu}{d\mathbb{P}_\theta/d\mu}\, d\mathbb{P} = \int \ln \frac{d\mathbb{P}}{d\mu}\, d\mathbb{P} - \int \ln L_k(\theta)\, d\mathbb{P}$$

sofern das linke Integral existiert. In diesem Fall reicht es also für die Kullback-Leibler-Deviance

$$\theta \mapsto d(\theta) := \mathbb{E}_{\mathbb{P}}(-2\ln L(\theta))$$

den Ausdruck $k \mapsto d(\widehat{\theta}_k)$ zu minieren. Eine empirische Version von $d(\theta)$ ist gegeben durch $-2\ln L(\theta, X)$, da ja X gemäß $\mathbb{P}$ verteilt ist. Da $\widehat{\theta}_k$ als ML-Schätzer gerade $\theta \mapsto -2\ln L(\theta, X)$ minimiert kann man andererseits vermuten, dass reines Einsetzen den Wert $d(\widehat{\theta}_k)$ unterschätzt. In der Tat kann man zeigen, dass oft (z.B. asymptotisch für i.i.d. Beobachtungen oder im linearen Modell s.u.) $\mathbb{E}_{\mathbb{P}}(-2\ln L(\widehat{\theta}_k(X), X)) = \mathbb{E}d(\widehat{\theta}_k) - 2k$.

Definition 5.5 (AIC)**.** In der obigen Situation ist das Akaike-Informations-Kriterium (AIC) definiert als

$$\mathrm{AIC}(k) := -2\ln L(\widehat{\theta}_k, X) + 2k, \qquad k = 1,\dots,K.$$

Das Modell $\widehat{k}$ wird nun gewählt als $\widehat{k} \in \mathrm{argmin}_k\, \mathrm{AIC}(k)$ und $\widehat{\theta}_{\widehat{k}}$ mit dem ML-Schätzer $\widehat{\theta}_k$ ist ein gemäß AIC gewählter Schätzer.

Satz 5.6 (AIC im linearen Modell, $\sigma^2 > 0$ bekannt)**.** *Betrachte das wahre lineare Modell unter* $\mathbb{P}$

$$Y = \mu + \varepsilon \qquad \textit{mit} \qquad \mu \in \mathbb{R}^n, \varepsilon \sim N(0, \sigma^2 I_n),$$

sowie für $k = 1,\dots,K$ *die Modelle*

$$Y = X^{(k)}\beta^{(k)} + \varepsilon \qquad \textit{mit} \qquad \beta^{(k)} \in \mathbb{R}^k, X^{(k)} \in \mathbb{R}^{n\times k}, \mathrm{rk}\, X^{(k)} = k.$$

Dann gilt für $\sigma^2 > 0$ *bekannt, dass*

(i) der KQ-Schätzer $\widehat{\beta}_k = (X^{(k)\top}X^{(k)})^{-1}X^{(k)\top}$ *ist ein ML-Schätzer;*

(ii) $AIC(k) = n\ln(2\pi\sigma^2) + \frac{RSS_k}{\sigma^2} + 2k, k = 1,\dots,K$ *mit* $RSS_k = \|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2$*;*

(iii) es gilt $\mathbb{E}AIC(k) = \mathbb{E}d(\widehat{\beta}^{(k)})$.

Beachte: Das AIC besteht aus einem empirischen Verlustanteil RSS_k/σ^2*, der mit wachsender Dimension* k *des Parameters fällt und einem Strafterm* $2k$ *der mit* k *wächst.* $\widehat{k}$ *balanciert also in gewisser Weise die Güte der Datenanpassung mit der Komplexität des statistischen Modells.*

Beweis.

(i) Das ist klar.

(ii) Es gilt

$$\begin{aligned}\mathrm{AIC}(k) &= -2\ln L(\widehat{\beta}^{(k)}) + 2k = -2\Big(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2\Big) + 2k \\ &= n\ln(2\pi\sigma^2) + \frac{\mathrm{RSS}_k}{\sigma^2} + 2k.\end{aligned}$$

(iii) Es gilt

$$\begin{aligned}d(\beta^{(k)}) &= \mathbb{E}_{\mathbb{P}}(-2\ln L(\beta^{(k)})) = n\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}_{\mathbb{P}}\|Y - X^{(k)}\beta^{(k)}\|^2 \\ &= n\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}\big(\|\mu - X^{(k)}\beta^{(k)}\|^2 - \mathbb{E}\|\varepsilon\|^2 + 2\cdot 0\big) \\ &= n\ln(2\pi\sigma^2) + \frac{\|\mu - X^{(k)}\beta^{(k)}\|^2}{\sigma^2} + n.\end{aligned}$$

Also ist

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}d(\widehat{\beta}^{(k)}) &= n\ln(2\pi\sigma^2) + n + \mathbb{E}_{\mathbb{P}}\frac{\|\mu - X^{(k)}\widehat{\beta}^{(k)}\|^2}{\sigma^2} \\ &= n\ln(2\pi\sigma^2) + n + \frac{\mathbb{E}_{\mathbb{P}}\|\mu - \Pi_k(\mu+\varepsilon)\|^2}{\sigma^2} \\ &= n\ln(2\pi\sigma^2) + n + \frac{\|\mu - \Pi_k\mu\|^2}{\sigma^2} + k.\end{aligned}$$

Andererseits gilt

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\mathrm{AIC}(k) &= n\ln(2\pi\sigma^2) + 2k + \frac{\mathbb{E}_{\mathbb{P}}\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{\sigma^2} \\ &= n\ln(2\pi\sigma^2) + 2k + \frac{\|\mu - \Pi_k\mu\|^2}{\sigma^2} + n - k,\end{aligned}$$

was die Behauptung liefert.

□

Corollar 5.7 (Unbiased risk estimation)**.** *In der Situation von Satz 5.6 gilt*

$$\mathbb{E}_{\mathbb{P}}(RSS_k + 2k\sigma^2 - n\sigma^2) = \mathbb{E}_{\mathbb{P}}\|\mu - X^{(k)}\widehat{\beta}^{(k)}\|^2,$$

d.h. die linke Seite ist eine unverzerrte Schätzung des quadratischen Vorhersagefehlers $\|\mu - X^{(k)}\widehat{\beta}^{(k)}\|^2$ und das AIC kann als unbiased risk estiamtion Kriterium interpretiert werden.

Beweis. Es gilt

$$\mathbb{E}_{\mathbb{P}}\mathrm{RSS}_k = \mathbb{E}_{\mathbb{P}}\|(I_n - \Pi_k)Y\|^2 = \|\mu - \Pi_k\mu\|^2 + (n-k)\sigma^2$$

und nach dem Satz des Pythagoras

$$\mathbb{E}_{\mathbb{P}}\|\mu - X^{(k)}\widehat{\beta}^{(k)}\|^2 = \|\mu - \Pi_k\mu\|^2 + k\sigma^2.$$

□

Bemerkung 5.8 (Mallow's C_p-Kriterium). Der Beweis von Corollar 5.7 zeigt insbesondere, dass für die unverzerrte Fehlerschätzung die Normalverteilung unerheblich ist, es reicht $\mathbb{E}\varepsilon = 0, \mathrm{Cov}\varepsilon = \sigma^2 I_n$ vorauszusetzen. Dies ist genau der Ansatz in Mallow's C_p-Kriterium, das im linearen Modell mit AIC übereinstimmt (Modulo in k konstanten Summanden).

Satz 5.9 (AIC im linearen Modell, $\sigma^2 > 0$ unbekannt)**.** *Betrachte das wahre lineare Modell unter $\mathbb{P}$*

$$Y = \mu + \varepsilon \qquad mit \qquad \mu \in \mathbb{R}^n, \varepsilon_0 \sim N(0, \sigma_0^2 I_n),$$

sowie für $p = 1, 2, \dots P$ die Modelle

$$Y = X^{(p)}\beta^{(p)} + \varepsilon \qquad mit \qquad \beta^{(p)} \in \mathbb{R}^p, X^{(p)} \in \mathbb{R}^{n\times p}, \mathrm{rk}\, X^{(p)} = p, \varepsilon \sim N(0, \sigma^2 I_n).$$

Für $k = p + 1$ ist $\theta_k = (\beta^{(p)}, \sigma^2)$ der unbekannte Parameter in $\mathbb{R}^k$. Dann gilt

(i) $\widehat{\theta}_k = (\widehat{\beta}^{(k)}, \widehat{\sigma})_k^2$ mit KQS $\widehat{\beta}^{(k)} \in \mathbb{R}^p$ und $\widehat{\sigma}_k^2 = RSS_k/n$ ist ML-Schätzer. Es gilt

$$AIC(k) = n(\ln(2\pi\widehat{\sigma}_k^2) + 1) + 2k.$$

(ii) Im Fall, dass $\mu = X^{(p)}\beta^{(p)}$ für ein p, gilt für $k = p - 1$, dass $\mathbb{E}AIC(k) = \mathbb{E}d(\widehat{\theta}_k) - 2\frac{k(k+1)}{n-k-1}$, d.h. dass $AIC(k)$ die Devianz $d(\widehat{\theta}_k)$ nicht unverzerrt schätzt, aber der Fehler ist nur von der Ordnung $O(1/n)$ in der Stichprobengröße n.

Beweis.

(i) Die ML-Eigenschaft ist klar und es gilt

$$\begin{aligned}\mathrm{AIC}(k) &= -2\ln L(\widehat{\theta}_k) + 2k = -2\Big(\ln(2\pi\widehat{\sigma}_k^2)^{-\frac{n}{2}} - \frac{\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{2\widehat{\sigma}_k^2}\Big) + 2k \\ &= n(\ln(2\pi\widehat{\sigma}_k^2) + 1) + 2k.\end{aligned}$$

(ii) Es gilt

$$d(\theta) = \mathbb{E}_{\mathbb{P}} - 2\ln L(\beta^{(p)}, \sigma^2) = n\ln(2\pi\sigma^2) + \frac{1}{\sigma^2}(\|\mu - X^{(p)}\beta^{(p)}\|^2 + n\sigma_0^2).$$

Also erhalten wir

$$\mathbb{E}_{\mathbb{P}}d(\widehat{\theta}_k) = n\mathbb{E}_{\mathbb{P}}\ln(2\pi\widehat{\sigma}_k^2) + \mathbb{E}\frac{\|\mu - \Pi_p Y\|^2 + n\sigma_0^2}{\widehat{\sigma}_k^2}$$

und

$$\mathbb{E}_{\mathbb{P}}\mathrm{AIC}(k) = n\mathbb{E}_{\mathbb{P}}\ln(2\pi\widehat{\sigma}_k^2) + n + 2k.$$

Damit ist

$$\mathbb{E}_{\mathbb{P}}(\mathrm{AIC}(k) - d(\widehat{\theta}_k)) = n + 2k - \mathbb{E}\frac{\|\mu - \Pi_p Y\|^2 + n\sigma_0^2}{\|(I_n - \Pi_p)Y\|^2}.$$

Auf Grund der Orthogonalität der Projektion sind in dem letzten Term Zähler und Nenner voneinander unabhängig. Es gilt also

$$\begin{aligned}\mathbb{E}\frac{\|\mu-\Pi_pY\|^2+n\sigma_0^2}{\|(I_n-\Pi_p)Y\|^2} &= \mathbb{E}_{\mathbb{P}}(\|\mu-\Pi_pY\|^2+n\sigma_0^2)\cdot\mathbb{E}_{\mathbb{P}}\|(I_n-\Pi_p)Y\|^{-2}\\ &= \big(\|\mu-\Pi_p\mu\|^2+\sigma_0^2(p+n)\big)\cdot\sigma_0^{-2}\mathbb{E}\chi_{n-p}^{-1}.\end{aligned}$$

Nun gilt

$$\mathbb{E}\chi_{n-p}^{-1}=\mathbb{E}\Gamma_{(n-p)/2,1/2}^{-1}=\int_0^\infty\frac{(1/2)^{(n-p)/2}}{\Gamma(\frac{n-p}{2})}t^{\frac{n-p}{2}-2}e^{-t/2}\,dt=\frac{1}{n-p-2}\qquad\text{für } n-p\geq 3$$

durch auswerten des Gamma-Integrals. Einsetzen liefert jetzt die Behauptung.

□

5.2 Das Bayes'sche Informationskriterium (BIC)

Anstelle des Ansatzes, die Kullback-Leibler-Devianz bzw. das Risiko unverzerrt zu schätzen, ist der Bayes'sche Ansatz (nach Schwarz), eine gleichmäßige a-priori-Verteilung für die Modelle $k=1,\dots,K$ anzunehmen. Wir leiten das BIC im Fall des wahren linearen Modells $Y=\mu+\varepsilon,\varepsilon\sim N(0,\sigma^2I_n)$ mit $\sigma^2>0$ bekannt her. Betrachte die Teilmodelle

$$Y=X^{(k)}\beta^{(k)}+\varepsilon,\qquad \beta^{(k)}\in\mathbb{R}^k,\operatorname{rk}X^{(k)}=k\leq n,k=1,\dots,K.$$

Gegeben dem Modell k sei der Parameter $\beta^{(k)}$ der Einfachheit halber gemäß $\pi_k\ll\lambda^{\otimes k}$ verteilt und das k-te Modell sei gemäß einem uninformativen Prior $\delta:=\sum_{k=1}^K\frac{1}{k}\delta_k$ gewählt. Diese Auswahl sei unabhänig von ε.

Abbildung 7: Illustration des Bayes-Ansatzes.

Wir erhalten also die gemeinsame Verteilung von Modell κ, Parameter $\tilde\beta$ und Beobachtung Y als

$$\mathbb{P}(A)=\int\mathbf{1}_A(k,\beta^{(k)},y)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(\frac{-\|y-X^{(k)}\beta^{(k)}\|}{2\sigma^2}\Big)\lambda^{(n)}(dy)\pi_k(d\beta^{(k)})\delta(dk)$$

Die gemeinsame Verteilung von (κ,Y) ist also gegeben durch die Dichte

$$f^{\kappa,Y}(k,y)=\int_{\mathbb{R}^k}\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\Big(\frac{-\|y-X^{(k)}\beta^{(k)}\|}{2\sigma^2}\Big)\pi_k(d\beta^{(k)})$$

bezüglich $\delta\otimes\lambda^{\otimes n}$. Der BIC-Ansatz besteht darin, das Modell $\widehat k\in\operatorname{argmax}_{k=1,\dots,K}\mathbb{P}\{\kappa=k|Y\}$ auszuwählen, wobei κ die zufällige Ziehung des Modells bezeichnet. Das BIC maximiert also die a-posteriori Wahrscheinlichkeit des Modells gegeben der Daten (MAP). Dabei gilt (beachte, dass sich $(2\pi\sigma^2)^{-n/2}$ herauskürzt durch die Bedingung)

$$\begin{aligned}\mathbb{P}\{\kappa=k|Y\} &= \Big(\frac{-\|Y-X^{(k)}\widehat\beta^{(k)}\|^2}{2\sigma^2}\Big)\int_{\mathbb{R}^k}\exp\Big(\frac{-\|X^{(k)}\widehat\beta^{(k)}-X^{(k)}\beta^{(k)}\|^2}{2\sigma^2}\Big)\pi_k(d\beta^{(k)}) \qquad (5.1)\\ &= \exp\Big(\frac{-\|Y-X^{(k)}\widehat\beta^{(k)}\|^2}{2\sigma^2}\Big)\int_{\mathbb{R}^k}\exp\Big(\frac{-\|X^{(k)}h/\sqrt{n}\|^2}{2\sigma^2}\Big)\frac{d\pi_k}{d\lambda^{\otimes k}}\Big(\widehat\beta^{(k)}+\frac{h}{\sqrt n}\Big)n^{-\frac{k}{2}}\,dh\\ &= n^{-\frac{k}{2}}\exp\Big(\frac{-\|Y-X^{(k)}\widehat\beta^{(k)}\|^2}{2\sigma^2}\Big)\int_{\mathbb{R}^k}\exp\Big(\frac{-\langle X^{(k)\top}X^{(k)}/nh,h\rangle}{2\sigma^2}\Big)\frac{d\pi_k}{d\lambda^{\otimes k}}\Big(\widehat\beta^{(k)}+\frac{h}{\sqrt n}\Big)\,dh.\end{aligned}$$

Betrachte nun die Asymptotik $n\to\infty$ und

$$\Sigma_n^{(k)}:=\frac{1}{n}X^{(k)\top}X^{(k)}\xrightarrow{n\to\infty}\Sigma^{(k)}\in\mathbb{R}^{k\times k}. \qquad (5.2)$$

Beispiel 5.10 (Asymptotik der Designmatrix im linearen Modell). Betrachte $\varphi_1,\dots,\varphi_k:[0,1]\to\mathbb{R}$ stetig (z.B. $\varphi_k(x)=x^{k-1}$) und

$$X^{(k)}=\begin{pmatrix}\varphi_1(1/n) & \dots & \varphi_k(1/n)\\ \varphi_1(2/n) & \dots & \varphi_k(2/n)\\ \vdots & \vdots & \vdots\\ \varphi_1(n/n) & \dots & \varphi_k(n/n)\end{pmatrix}.$$

Dann gilt

$$\Sigma_n^{(k)}=\begin{pmatrix}\sum_{i=1}^n\varphi_1(i/n)^2 & \dots & \sum_{i=1}^n\varphi_1(i/n)\varphi_k(i/n)\\ \vdots & \ddots & \vdots\\ \sum_{i=1}^n\varphi_n(i/n)\varphi_1(i/n) & \vdots & \sum_{i=1}^n\varphi_k(i/n)\varphi_k(i/n)\end{pmatrix}/n\xrightarrow{n\to\infty}\Big(\int_0^1\varphi_j\varphi_l\,d\lambda\Big)_{jl}$$

Vergleiche auch die Übung für i.i.d. random-Design.

Nehmen wir nun weiter an, dass $d\pi_k/d\lambda^{\otimes k} > 0$ auf $\mathbb{R}^k$ und stetig und beschränkt ist, so erhalten wir in (5.1), mit dominierter Konvergenz, dass für Konstanten $C, C' > 0$

$$\ln \mathbb{P}\{\kappa = k|Y\} = \ln C - \frac{k}{2}\ln n - \frac{\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{2\sigma^2} + \underbrace{\ln \int \dots \, dh}_{\xrightarrow{n\to\infty} c\in\mathbb{R}}.$$

Also ist

$$-2\ln \mathbb{P}\{\kappa = k|Y\} = k\ln n\big(1 + o(1)\big) + \frac{\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{2\sigma^2}$$

mit einem Term $o(1) \xrightarrow{n\to\infty} 0$. Vernachlässigt man diesen Term wählt das BIC entsprechend $\widehat{k} \in \operatorname{argmin}(\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2 + (\sigma^2 k \ln n)(1 + o(1)))$. Allgemein definieren wir analog zum AIC.

Definition 5.11 (BIC)**.** Betrachte die aufsteigenden Modelle $(\mathfrak{X}, \mathscr{F}, (\mathbb{P}_\theta)_{\theta\in\Theta_k})$ mit $\Theta_k \subset \mathbb{R}^k$ mit Raumdimension $\dim \mathfrak{X} = n$, Likelihoodfunktion L und ML-Schätzer $\widehat{\theta}_k$ in Modell k.

(a) Das allgemeine BIC ist definiert als

$$\mathrm{BIC}(k) := -2\ln L(\widehat{\theta}_k) + k\ln n, \qquad k = 1, \dots, K.$$

Man wählt entsprechend $\widehat{k} \in \operatorname{argmin} \mathrm{BIC}$.

(b) In der Situation der linearen Modelle mit bekanntem $\sigma^2 > 0$ wie oben ergibt sich

$$\mathrm{BIC}(k) = n\ln(2\pi\sigma^2) + \frac{\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{\sigma^2} + k\ln n.$$

Bemerkung 5.12 (Vergleich AIC vs. BIC).

(a) Im Vergleich zum AIC wird im BIC die Modelldimension mit $k\ln n$ statt mit $2k$ penalisiert. BIC wählt also für $n \geq 8$ tendenziell ein kleineres Modell aus.

(b) Im linearen Modell wird mit dem BIC im Allgemeinen eher das richtige Modell gefunden. Andererseits ist das AIC besser darin den Vorhersagefehler $\|\mu - X^{(k)}\widehat{\beta}^{(k)}\|^2$ zu minimieren.

(c) Beachte noch, dass $\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2$ tendenziell von der Größenarndnung n ist. Reformuliert man das BIC als $\|Y - X^{(k)}\widehat{\beta}^{(k)}\|_n^2 + k\frac{\ln n}{n}$, so sieht man, dass die Erhöhung der Modelldimension mit steigendem n geringer bestraft wird.

5.3 Hauptsatz der penalisierten Modellwahl

Bei AIC und BIC haben wir bei Beobachtung von $Y = \mu + \varepsilon$ die KQ-Schätzer $\widehat{\beta}^{(k)}$ für die Modelle $Y = X^{(k)}\beta^{(k)} + \varepsilon$ gewählt und $\widehat{k} \in \operatorname{argmin} \mathrm{AIC}(k)$ bzw. $\widehat{k} \in \operatorname{argmin} \mathrm{BIC}(k)$ bestimmt um den finalen Schätzer $\widehat{\beta}^{(\widehat{k})}$ auszuwählen. Beide Modellwahlkriterien waren von der Form

$$\widehat{k} \in \operatorname*{argmin}_{k=1,\dots,K} \big(\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2 + \mathrm{Pen}(k)\big)$$

mit Penalisierungstermen $\mathrm{Pen}(k)$. Betrachtet man $S_k := \operatorname{im} X^{(k)}$ und $\widehat{\mu}^{(k)} := X^{(k)}\widehat{\beta}^{(k)}$, bettet sich das in die folgende allgemeine Modellwahl ein:

Definition 5.13 (Allgemeine penalisierte Modellwahl)**.** Für eine Beobachtung $Y \in \mathbb{R}^n$ seien lineare Unterräume $S_m \subset \mathbb{R}^n, m = 1, \dots, M$, von dimension d_m gegeben. Für eine Funktion $\mathrm{Pen} : \mathbb{N} \to [0, \infty)$ betrachten wir die Modellwahl

$$\widehat{m} \in \operatorname*{argmin}_{m=1,\dots,M} \big(\|Y - \widehat{\mu}^{(m)}\|^2 + \mathrm{Pen}(d_m)\big)$$

mit KQ-Schätzern $\widehat{\mu}^{(m)} := \Pi_{S_m} Y \in S_m$.

Im Fall, dass das datenerzeugende Modell gegeben ist durch

$$Y = \mu + \varepsilon, \qquad \mu \in \mathbb{R}^n, \varepsilon \sim N(0, \sigma^2 I_n),$$

können wir die folgende Orakelungleichung beweisen.

Theorem 5.14 (**Hauptsatz zur Modellwahl**)**.** *In der obigen Situation gelte $K > 1, Pen(d_m) \geq K\sigma^2(d_m + 1)$ und $\kappa \in (0, \sqrt{K} - 1)$. Wir setzen weiter $\mu^{(m)} := \Pi_m \mu$, wobei Π_m die Projektion auf S_m ist.*

(i) Für jedes $\tau > 0$ gilt mit Wahrscheinlichkeit $1 - \sum_{m=1}^M e^{-d_m\kappa^2/2}e^{-\tau/2}$ die Orakelungleichung

$$\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 \leq C(K,\kappa)\Big(\min_{m=1,\dots,M} \big(\|\mu - \Pi_{S_m}\mu\|^2 + Pen(d_m)\big) + \sigma^2\tau\Big)$$

mit einer Konstanten $C(K, \kappa)$, die nur von K und κ abhängt, wobei $C(K,\kappa) \xrightarrow{\kappa\to\sqrt{K}-1} \infty$.

(ii) Es gilt

$$\mathbb{E}\|\widehat{\mu}^{(\widehat{m})}-\mu\|^2 \le C(K,\kappa)\Big(\min_{m=1,\dots,M}\big(\|\mu-\Pi_{S_m}\mu\|^2+Pen(d_m)\big)+\sigma^2\sum_{m=1}^{M}e^{-d_m\kappa^2/2}\Big)$$

mit einer Konstanten $C(K,\kappa)$, die nur von K und κ abhängt, wobei $C(K,\kappa)\xrightarrow{\kappa\to\sqrt{K}-1}\infty$.

Beweis.

(i) **Schritt 1: Anwenden der Fundamentalungleichung.** Aus der Fundamentalungleichung 5.15 folgt sofort für beliebiges $m^*\in\{1,\dots,M\}$, dass

$$\|\widehat{\mu}^{(\widehat{m})}-\mu\|^2 \le \|\mu^{(m^*)}-\mu\|^2+\mathrm{Pen}(d_{m^*})-\mathrm{Pen}(d_{\widehat{m}})+2\langle\varepsilon,\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\rangle.$$

Es gilt $\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\in\mathrm{span}(S_{\widehat{m}},\mu^{(m^*)})=:S^*_{\widehat{m}}$ mit $\dim S^*_{\widehat{m}}\le d_{\widehat{m}}+1$ und damit

$$\langle\varepsilon,\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\rangle \le \|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\sup_{s\in S^*_{\widehat{m}}}\frac{\langle\varepsilon,s\rangle}{\|s\|}=\|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\|\Pi^*_{\widehat{m}}\varepsilon\|$$

mit der entsprechenden Projektion $\Pi^*_{\widehat{m}}:\mathbb{R}^n\to S^*_{\widehat{m}}$.

Schritt 2: Konzentrationsungleichung für χ^2_p. Für $X_p\sim\chi^2_p=\Gamma_{\frac{p}{2},\frac{1}{2}}$ gilt für $\varrho>1$ und $u<1/2$

$$\mathbb{P}\{X_p\ge\varrho p\}\le\mathbb{E}e^{uX_p}e^{-u\varrho p}=\frac{1}{(1-2u)^{p/2}}e^{-u\varrho p}=\varrho^{p/2}e^{-(\varrho-1)/2\cdot p}=e^{\frac{-p}{2}(\varrho-1-\ln\varrho)}.$$

für $u=(\varrho-1)/(2\varrho)$. Daraus folgt, wegen $\ln(1+t)\le t, t\ge 0$, dass

$$\begin{aligned}\mathbb{P}\Big\{X_p\ge\Big(1+\kappa+\sqrt{\frac{\tau}{p}}\Big)^2p\Big\}&\le\exp\Big(-\frac{p}{2}\Big[\Big(\kappa+\sqrt{\frac{\tau}{p}}\Big)^2+2\Big(\kappa+\sqrt{\frac{\tau}{p}}\Big)-2\ln\Big(1+\kappa+\sqrt{\frac{\tau}{p}}\Big)\Big]\Big)\\&\le\exp\Big(-\frac{p}{2}\Big(\kappa+\sqrt{\frac{\tau}{p}}\Big)^2\Big)\le\exp\Big(-\frac{p}{2}\kappa^2-\frac{\tau}{2}\Big).\end{aligned}$$

Schritt 3: Kontrolle des Mischterms in Wahrscheinlichkeit. Wir normieren und korrigieren um einen Term der Größenordnung von $\mathbb{E}\|\Pi^*_{\widehat{m}}\varepsilon\|/\sigma$ und erhalten damit aus Schritt 1, dass

$$\begin{aligned}\|\Pi^*_{\widehat{m}}\varepsilon\|&\le\sigma\Big((\kappa+1)\sqrt{d_{\widehat{m}}+1}+\|\Pi^*_{\widehat{m}}\varepsilon\|/\sigma-(\kappa+1)\sqrt{d_{\widehat{m}}+1}\Big)\\&\le\sigma\Big((\kappa+1)\sqrt{d_{\widehat{m}}+1}+\max_{m=1,\dots,M}\big(\|\Pi^*_m\varepsilon\|/\sigma-(\kappa+1)\sqrt{d_m+1}\big)\Big).\end{aligned}$$

Eine Unionbound Abschätzung zusammen mit Schritt 2 liefert

$$\begin{aligned}&\mathbb{P}\Big\{\max_{m=1,\dots,M}\big(\|\Pi^*_m\varepsilon\|-(\kappa+1)\sqrt{d_m+1}\big)\ge\sqrt{\tau}\Big\}\\&\le\sum_{m=1}^{M}\mathbb{P}\big\{\|\Pi^*_m\varepsilon\|-(\kappa+1)\sqrt{d_m+1}\ge\sqrt{\tau}\big\}\le\sum_{m=1}^{M}\mathbb{P}\Big\{\sqrt{\chi^2_{d_m+1}}\ge\Big(1+\kappa+\sqrt{\frac{\tau}{d_m+1}}\Big)\sqrt{d_m+1}\Big\}\\&\le\sum_{m=1}^{M}\exp\Big(-\frac{d_m+1}{2}\kappa^2-\frac{\tau}{2}\Big)\le\sum_{m=1}^{M}e^{-d_m\kappa^2/2}e^{-\tau/2}.\end{aligned}$$

Mit der gesuchten Wahrscheinlichkeit erhalten wir also

$$\begin{aligned}\langle\varepsilon,\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\rangle&<\|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\sigma\big((\kappa+1)\sqrt{d_{\widehat{m}}+1}+\sqrt{\tau}\big)\\&\le\|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\sigma\Big(\frac{(\kappa+1)}{\sqrt{K}\sigma}\sqrt{\mathrm{Pen}(d_{\widehat{m}})}+\sqrt{\tau}\Big).\end{aligned}$$

Schritt 4: Umordnung der Terme. Mit der gesuchten Wahrscheinlichkeit erhalten wir jetzt

$$\|\widehat{\mu}^{(\widehat{m})}-\mu\|^2\le\|\mu^{(m^*)}-\mu\|^2+\mathrm{Pen}(d_{m^*})-\mathrm{Pen}(d_{\widehat{m}})+2\|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\Big(\frac{1+\kappa}{\sqrt{K}}\sqrt{\mathrm{Pen}(d_{\widehat{m}})}+\sigma\sqrt{\tau}\Big).$$

Mit $\|\widehat{\mu}^{(\widehat{m})}-\mu^{(m^*)}\|\le\|\widehat{\mu}^{(\widehat{m})}-\mu\|+\|\mu^{(m^*)}-\mu\|$ und durch zweimalige Anwendung von $2ab\le\eta a^2+\eta^{-1}b^2,\eta>0$ gilt, dass

$$\begin{aligned}&2\|\widehat{\mu}^{(\widehat{m})}-\mu\|\Big(\frac{1+\kappa}{\sqrt{K}}\sqrt{\mathrm{Pen}(d_{\widehat{m}})}+\sigma\sqrt{\tau}\Big)\\&\le\big(\eta_1^{-1}+\eta_2^{-1}\big)\|\widehat{\mu}^{(\widehat{m})}-\mu\|^2+\eta_1\frac{(1+\kappa)^2}{K}\mathrm{Pen}(d_{\widehat{m}})+\eta_2\sigma^2\tau\end{aligned}$$

Behandelt man den zweiten Term genauso folgt

$$\begin{aligned}\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 \le \big(1 + \eta_3^{-1} + \eta_4^{-1}\big)\|\mu^{(m^*)} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) + (\eta_2 + \eta_4)\sigma^2\tau \\ + \Big((\eta_1 + \eta_3)\frac{(1+\kappa)^2}{K} - 1\Big)\mathrm{Pen}(d_{\widehat{m}}) + \big(\eta_1^{-1} + \eta_2^{-1}\big)\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2\end{aligned}$$

Wählt man $(\eta_1^{-1} + \eta_2^{-1}) < 1$, $\eta_1 + \eta_3 = K/(1+\kappa)^2 > 1$ und $\eta_4 = 1$, so folgt durch Umstellen

$$\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 \le C(\eta_1, \dots, \eta_3)\big(\|\mu^{(m^*)} - \mu\|^2 + \mathrm{Pen}(d_{m^*}) + \sigma^2\tau\big).$$

(ii) Setze $t^* := C(K,\kappa)\min_m(\|\mu - \Pi_{S_m}\mu\|^2 + \mathrm{Pen}(d_m))$. Es gilt dann

$$\begin{aligned}\mathbb{E}\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 &= \int_0^\infty \mathbb{P}\{\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 > t\}\,dt \\ &\le \int_0^{t^*} \mathbb{P}\{\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 > t\}\,dt + \int_0^\infty \mathbb{P}\big\{\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 > t^* + C(K,\kappa)\sigma^2\tau\big\}C(K,\kappa)\sigma^2\,d\tau \\ &\le t^* + C(K,\kappa)\sigma^2\sum_{m=1}^M e^{-d_m\kappa^2/2}.\end{aligned}$$

□

Lemma 5.15 (**Fundamentalungleichung**). *In der Situation von Theorem 5.14 gilt für beliebiges $m^* \in \{1, \dots, M\}$*

$$\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2 + \mathit{Pen}(d_{\widehat{m}}) \le \|\mu^{(m)} - \mu\|^2 + \mathit{Pen}(m) + 2\langle\varepsilon, \widehat{\mu}^{(\widehat{m})} - \mu^{(m)}\rangle.$$

Beweis. Per Konstruktion gilt

$$\|Y - \widehat{\mu}^{(\widehat{m})}\|^2 + \mathrm{Pen}(\widehat{m}) \le \|Y - \widehat{\mu}^{(m)}\|^2 + \mathrm{Pen}(m) \le \|Y - \mu^{(m)}\|^2 + \mathrm{Pen}(m).$$

Einsetzen von $Y = \mu + \varepsilon$ und Ausmultiplizieren liefert jetzt die Behauptung. □

Bemerkung 5.16 (Interpretation des Hauptsatzes).

(a) Das m^*, mit dem das Minimum auf der rechten Seite angenommen wird, nennt man auch *Orakelmodell*. aus der Bias-Varianz-Zerlegung folgt

$$\mathbb{E}\|\widehat{\mu}^{(m)} - \mu\|^2 = \|\mu^{(m)} - \mu\|^2 + \sigma^2 d_m \approx \|\mu^{(m)} - \mu\|^2 + \mathrm{Pen}(d_m)$$

für $\mathrm{Pen}(d_m) \approx \sigma^2(d_m + 1)$. Damit liegt $\|\mu^{(m^*)} - \mu\|^2 + \mathrm{Pen}(d_{m^*})$ nahe dem *Orakelfehler* $\min_m \mathbb{E}\|\widehat{\mu}^{(m)} - \mu\|^2$. Mit $\tau \approx d_{m^*}$ ist der Restterm von der Ordnung $\mathrm{Pen}(d_{m^*})$. Dann ist Wahrscheinlichkeit gegeben durch

$$\sum_{m=1}^M \exp\big(-d_m\kappa^2/2\big)\exp(-d_{m^*}/2).$$

Falls asymptotisch $n \to \infty$, $d_{m^*} \xrightarrow{n\to\infty} \infty$ und die Summe gleichmäßig in M beschränkt ist, so ist der Abfall der Gegenwahrscheinlichkeit exponentiell.

(b) In Teil (b) des Hauptsatzes kann man für festes M und $n \to \infty$ z.B. im Fall des BIC mit $\mathrm{Pen}(d_m) = \ln(n)d_m\sigma^2$, κ zunehmend größer wählen. Der hintere Term kann asymptotisch sehr klein gewählt werden.

(c) Im Satz wird nicht gefordert, dass die Modelle geordnet sind. In der multiplen linearen Regression können alle 2^k Untermodelle betrachtet werden. Im Satz muss dann κ hinreichend groß gewählt werden, um eine große Wahrscheinlichkeit sicherzustellen. In der Tat zeigt sich bei dieser '"fullsubset„-Variablenselektion, dass in der Praxis AIC und BIC nicht so gut funktionieren wie größere $\mathrm{Pen}(d_m)$-Penalisierungen.

5.4 Kreuzvalidierung (CV)

Die erste Idee zur Modellwahl ist „sample splitting“, d.h. „hold out“ oder Validierung: Gegeben die Beobachtungen $Y_1, \dots, Y_n$ spalte sie in eine Trainingsmenge $Y_1, \dots, Y_m, m < n$ und eine Validierungsmenge $Y_{m+1}, \dots, Y_n$. Schätze ein Modell anhand der Trainingsmenge und überprüfe es anhand der Modellvorhersage für $Y_{m+1}, \dots, Y_n$. Bei der Auswahl aus K verschiedenen Modellen führt das führe das für jedes Modell $k \le K$ durch und wähle das Modell, das die Vorhersage minimiert.

Das wird in der Praxis sehr häufig gemacht, allerdings „verschenkt“ man dür die Schätzung $n - m$ Daten. Dies führt auf die Idee, die Auswahl der Trainingsdaten und Validierungsmengen zu permutieren und die entsprechenden Schätzer und Vorhersagefehler zu mitteln.

Leave p out-Kreuzvalidierung (Lpo-CV). In der obigen Situation betrachtet man jede Validierungsmenge $V \subset \{Y_1, \dots, Y_n\}$ der Kardinalität $p < n$ und benutzt $V^c = \{Y_1, \dots, Y_n\} \setminus V$ als Trainingsmenge. Bestimme dann den Vorhersagefehler $E_V^{(k)}$ und verwende das Modellwahlkriterium

$$\widehat{k} \in \operatorname*{argmin}_{k=1,\dots,K} \mathrm{CV}_{\mathrm{Lpo}}(k) := \operatorname*{argmin}_{k=1,\dots,K} \sum_{V\subset\{Y_1,\dots,Y_n\}} E_V^{(k)}.$$

Im Folgenden betrachten wir das lineare Modell mit „leave one out“-Kreuzvalidierung (Loo-CV). Wir betrachten wieder die linearen Modelle

$$Y = X^{(k)}\beta^{(k)} + \varepsilon \qquad \text{mit} \qquad X^{(k)} \in \mathbb{R}^{n\times k}, \operatorname{rk} X^{(k)} = k, \beta^{(k)} \in \mathbb{R}^k, \varepsilon \sim N(0, \sigma^2 I_n)$$

für $k = 1, \dots, K$. Wir bestimmen die KQ-Schätzer $\widehat{\beta}_{-i}^{(k)}$ in Modell k ohne Beobachtung Y_i, d.h.

$$\widehat{\beta}_{-i}^{(k)} = \big(X_{-i}^{(k)\top} X_{-i}^{(k)}\big)^{-1} X_{-i}^{(k)\top} Y \qquad \text{mit} \qquad X_{-i}^{(k)} := I_n^{(-i)} X^{(k)},$$

wobei $I_n^{(-i)} := I_n - I_n^{(ii)}$, die Indentitäsmatrix der dimension n ist, in der die i-te Zeile genullt wird und wir annehmen, dass $\operatorname{rk} X_{-i}^{(k)} = k \le n-1$ für alle i. Der quadratische Vorhersagefehler auf der Trainingsmenge ist dann $(Y_i - (X^{(k)}\widehat{\beta}_{-i}^{(k)})_i)^2$. Wir erhalten also

$$\mathrm{CV}(k) := \mathrm{CV}_{Loo}(k) = \sum_{i=1}^{n} \big(Y_i - (X^{(k)}\widehat{\beta}_{-i}^{(k)})_i\big)^2$$

Lemma 5.17 (Berechenbarkeit von CV aus den Daten)**.** *In der obigen Situation gilt*

$$CV(k) = \|(I_n - \tilde{\Pi}_k)^{-1}(I_n - \Pi_k)Y\|^2, \qquad k = 1, \dots, K$$

mit der orthogonalen Projektion Π_k *auf* $\operatorname{im} X^{(k)}$ *und* $\tilde{\Pi}_k := diag((\Pi_k)_{11}, \dots, (\Pi_k)_{nn})$.

Beweis. Aus Notationsgründen lassen wir die Abhängigkeit von k fallen. Es gilt

$$\mathrm{CV}(k) = \sum_{i=1}^{n} \big(Y_i - (X\widehat{\beta}_{-i})_i\big)^2 = \|Y - \sum_{i=1}^{n} I_n^{(ii)} X\widehat{\beta}_{-i}\|^2 =: \|Y - MY\|^2.$$

Es gilt aber für beliebige $\beta \in \mathbb{R}^k$, dass

$$MX\beta = \sum_{i=1}^{n} I_n^{(ii)} X\big(X_{-i}^\top X_{-i}\big)^{-1} X_{-i}^\top X\beta = \sum_{i=1}^{n} I_n^{(ii)} X I_n \beta = X\beta,$$

d.h., dass M auf $\operatorname{im} X$ die Identität ist. Es folgt, dass $\mathrm{CV}(k) = \|(I_n - M)(Y - \Pi Y)\|^2$. Indem wir nun $X^\top(Y - \Pi X) = 0$ einsetzen erhalten wir

$$\begin{aligned}
\mathrm{CV}(k) &= \|\big(I_n - \sum_{i=1}^{n} I_n^{(ii)} X (X_{-i}^\top X_{-i})^{-1}((I_n - I_n^{(ii)})X)^\top\big)(Y - \Pi Y)\|^2 \\
&= \|\big(I_n + \sum_{i=1}^{n} I_n^{(ii)} X (X_{-i}^\top X_{-i})^{-1}(I_n^{(ii)}X)^\top\big)(Y - \Pi Y)\|^2 \\
&= \|\big(I_n + \sum_{i=1}^{n} I_n^{(ii)} X (X_{-i}^\top X_{-i})^{-1} X^\top I_n^{(ii)}\big)(Y - \Pi Y)\|^2 =: \|(I_n + A)(Y - \Pi Y)\|^2.
\end{aligned}$$

Es gilt aber nun $(I_n + A)(I_n - \tilde{\Pi}) = I_n$, da der Eintrag in der i-ten Zeile der Diagonalmatrix $A\tilde{\Pi}$ gegeben ist durch

$$\begin{aligned}
& I_n^{(ii)} X (X_{-i}^\top X)^{-1} X^\top I_n^{(ii)} X (X^\top X)^{-1} X^\top \\
&= I_n^{(ii)} X (X_{-i}^\top X)^{-1} X^\top (I_n - I_n^{(-i)}) X (X^\top X)^{-1} X^\top \\
&= I_n^{(ii)} A - I_n^{(ii)} \tilde{\Pi}.
\end{aligned}$$

□

In der Situation von Lemma 5.17 gilt

$$\operatorname{tr}(I_n - \tilde{\Pi}_k) = \operatorname{tr}(I_n - \Pi_k) = n - k,$$

da $I_n - \Pi_k$ eine orthogonale Projektion auf einen $(n-k)$-dimensionalen Unterraum ist (Diagonalisieren, Eigenwerte alle 1). Im Mittel hat $(I_n - \tilde{\Pi}_k)_{ii}$ also den Wert $(n-k)/n = 1 - k/n$. Daher definiert man:

Definition 5.18 (Allgemeine Kreuzvalidierung im Linearen Modell)**.** Im linearen Modell definieren wir

$$\mathrm{GCV}(k) := \frac{\mathrm{RSS}_k}{(1 - k/n)^2} = \frac{\|Y - X^{(k)}\widehat{\beta}^{(k)}\|^2}{(1 - k/n)^2}, \qquad k = 1, \dots, K$$

und wählt $\widehat{k}_{\mathrm{GCV}} \in \operatorname{argmin}_k \mathrm{GCV}(k)$.

Ein Vorteil von CV und GCV ist, dass zur Definition σ^2 nicht benötigt wird. Wir vergleichen GCV mit AIC bei unbekanntem $\sigma^2 > 0$:

$$\mathrm{GCV}(k) = \frac{\mathrm{RSS}_k}{(1-k/n)^2} \qquad \text{vs.} \qquad \mathrm{AIC}(k) = n\big(\ln(2\pi\widehat{\sigma}_k^2) + 1\big) + 2k.$$

Nun gilt aber mit einer monotonen Transformation

$$\begin{aligned}\widehat{k}_{\mathrm{AIC}} &= \operatorname*{argmin}_{k=1,\dots,K} \big(n\ln(\mathrm{RSS}_k) + 2k\big) = \operatorname*{argmin}_{k=1,\dots,K} \Big(\ln(\mathrm{RSS}_k) + 2\frac{k}{n}\Big) \\ &= \operatorname*{argmin}_{k=1,\dots,K} \Big(\mathrm{RSS}_k + e^{2k/n}\Big) = \operatorname*{argmin}_{k=1,\dots,K} \Big(\frac{\mathrm{RSS}_k}{e^{-k/n\cdot 2}}\Big)\end{aligned}$$

Falls n groß ist im Vergleich zu $k = 1, \dots, K$, so ist $e^{-k/n} \approx 1 - k/n$ und $\widehat{k}_{\mathrm{GCV}}$ stimmt für $n \to \infty$ mit $\widehat{k}_{\mathrm{AIC}}$ bei unbekanntem $\sigma^2 > 0$ überein. Man kann unter geeigneter Asymptotic zeigen, dass beide dann auch mit $\widehat{k}_{\mathrm{AIC}}$ bei bekanntem $\sigma^2 > 0$ übereinstimmen.

5.5 Der LASSO-Schätzer

LASSO: Least Absolute Shrinkate and Selection Operator. Wir betrachten wieder das lineare Modell

$$Y = X\beta + \varepsilon \qquad \text{mit} \qquad \beta \in \mathbb{R}^p, X \in \mathbb{R}^{n\times p}, \operatorname{rk} X = p, \varepsilon \sim N(0, \sigma^2 I_n).$$

Eine wichtige Anwendung der Modellwahl ist Variablenselektion in der multiplen linearen Regression

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon i, \qquad i = 1, \dots, n.$$

Die Modellannahme ist in vielen Fällen, dass nur wenige β_j signifikant von null verschieden sind. Die dimensionspenalisierte Modellwahl betrachtet für einen Tuningparameter $\lambda > 0$

$$\widehat{S} := \operatorname*{argmin}_{S\subset\{1,\dots,p\}} \|Y - \Pi_S Y\|^2 + \lambda|S| = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + \lambda\|\beta\|_0,$$

wobei Π_S die orthogonale Projektion auf $\{\sum_{j\in S}\beta_j x_{\cdot j} : \beta_1, \dots, \beta_p \in \mathbb{R}\}$ und $\|\beta\|_0$ die Anzahl der von null verschiedenen Komponenten von β bezeichnet. Numerisch ist das Bestimmen von $\widehat{S}$ ein NP-vollständiges Probelm, d.h. wir brauchen nicht viel weniger als 2^p Berechnungen von $\|Y - \Pi_S Y\|^2$. Die Funktion $\beta \mapsto \|\beta\|_0$ ist auch nicht konvex, so dass auch Näherungsverfahren schwierig sind. Daher wird bei der LASSO-Schätzung $\|\beta\|_0$ durch die kleinste konvexe Majorante (in gewissem Sinne) ersetzt, nämlich durch $\|\beta\|_1 = \sum_{j=1}^p |\beta|_j$.

Abbildung 8: Verallgemeinerte L^p-Normen.

Definition 5.19 (LASSO-Schätzer)**.** Im multiplen Regressionsmodell

$$Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \dots, n$$

mit $x_{\cdot j} \in \mathbb{R}^n, \varepsilon \sim N(0, \sigma^2 I_n)$ (Beachte voller Rang von X nicht benötigt) ist der *LASSO-Schätzer* definiert als

$$\widehat{\beta} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|Y - X\beta\|^2 + \lambda\|\beta\|_1$$

für einen Penalisierungsparameter $\lambda > 0$.

Man kann zeigen, dass der LASSO-Schätzer nur wenige Koeffizienten besitzt die ungleich null sind und gute Variablenselektionseigenschaften besitzt. Um die Eigenschaften besser zu verstehen und insbesondere die Wahl von $\lambda > 0$ zu klären wollen wir wieder eine Orakelungleichung herleiten.

Lemma 5.20 (Fundamentalungleichung für LASSO)**.** *Für jedes $\beta^* \in \mathbb{R}^p$ gilt im LASSO-Modell mit wahrem $\beta \in \mathbb{R}^p$, dass*

$$\|X\widehat{\beta} - X\beta\|^2 + \lambda\|\widehat{\beta}\|_1 \le \|X\beta^* - X\beta\|^2 + \lambda\|\beta^*\|_1 + 2\langle\varepsilon, X\widehat{\beta} - X\beta^*\rangle.$$

Beweis. Es gilt nach Konstruktion

$$\|Y - X\widehat{\beta}\|^2 + \lambda\|\widehat{\beta}\|_1 \le \|Y - X\beta^*\|^2 + \lambda\|\beta^*\|_1.$$

Einsetzen von $Y = X\beta + \varepsilon$ und Ausmultiplizieren liefert jetzt die Behauptung. □

Wir werden jetzt die folgende ℓ^1-ℓ^∞-Abschätzung aus der Hölderungleichung verwenden

$$|\langle\varepsilon, X\widehat{\beta} - X\beta^*\rangle| \le \|X^\top\varepsilon\|_\infty \|\widehat{\beta} - \beta^*\|_1.$$

Satz 5.21 (**Orakelungleichung für den LASSO-Schätzer**). *Für jedes β^* setze $S^* := \{j : \beta_j^* \neq 0\}$ („active set"). Allgemein sei $\|b\|_S := \sum_{j\in S} |b_j|, b \in \mathbb{R}^p$. Auf dem Ereignis $G := \{\|X^\top \varepsilon\|_\infty \leq \lambda/8\}$ gilt dann*

$$\|X\widehat{\beta} - X\beta\|^2 + \lambda\|\widehat{\beta}\|_{S^{*c}} \leq \frac{5}{3}\|X\beta^* - X\beta\|^2 + \frac{25}{6}\frac{\lambda^2|S^*|}{\lambda_{\min}(X^\top X)} \qquad \text{für alle } \beta^* \in \mathbb{R}^p,$$

wobei der kleinste Eigenwert $\lambda_{\min}(X^\top X)$ von $X^\top X$ echt gößer null sei.

Beweis. Auf G gilt mit der Fundamentalungleichung für den LASSO 5.20, dass

$$\begin{aligned}
4\|X\widehat{\beta} - X\beta\|^2 + 4\lambda\|\widehat{\beta}\|_1 &\leq 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\beta^*\|_1 + 8|\langle \varepsilon, X\widehat{\beta} - X\beta^*\rangle| \\
&\leq 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\beta^*\|_1 + 8\|X^\top\varepsilon\|_\infty\|\widehat{\beta} - \beta^*\|_1 \\
&\leq 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\beta^*\|_1 + \lambda\|\widehat{\beta} - \beta^*\|_1
\end{aligned}$$

Wir erhalten also wegen $\|\beta\|_1 = \|\beta\|_{S^*} + \|\beta\|_{S^{*c}}$ und $\|\beta^*\|_{S^{*c}} = 0$, dass

$$\begin{aligned}
4\|X\widehat{\beta} - X\beta\|^2 + 4\lambda\|\widehat{\beta}\|_{S^{*c}} &\leq 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\beta^*\|_1 + \lambda\|\widehat{\beta} - \beta^*\|_1 - 4\lambda\|\widehat{\beta}\|_{S^*} \\
&\leq 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\widehat{\beta} - \beta^*\|_{S^*} + \lambda\|\widehat{\beta} - \beta^*\|_1 \pm \lambda\|\widehat{\beta}\|_{S^{*c}} \\
&= 4\|X\beta^* - X\beta\|^2 + 4\lambda\|\widehat{\beta} - \beta^*\|_{S^*} + \lambda\|\widehat{\beta} - \beta^*\|_{S^*} + \lambda\|\widehat{\beta}\|_{S^{*c}}
\end{aligned}$$

und damit

$$4\|X\widehat{\beta} - X\beta\|^2 + 3\lambda\|\widehat{\beta}\|_{S^{*c}} \leq 4\|X\beta^* - X\beta\|^2 + 5\lambda\|\widehat{\beta} - \beta^*\|_{S^*}.$$

Um den hinteren Term in relation zu dem vorderen zu setzen beachte noch, dass für beliebige $b \in \mathbb{R}^p$ gilt

$$\|Xb\|^2 = \langle X^\top Xb, b\rangle \geq \lambda_{\min}(X^\top X)\|b\|^2$$

mit Jensen erhalten wir dann

$$\|\widehat{\beta} - \beta^*\|_{S^*}^2 \leq |S^*| \sum_{j\in S^*} |\widehat{\beta}_j - \beta_j^*|^2 \leq \frac{|S^*|\|X\widehat{\beta} - X\beta^*\|^2}{\lambda_{\min}(X^\top X)}.$$

Verwende nun $\|X\widehat{\beta} - \beta^*\| \leq \|X\widehat{\beta} - X\beta\| + \|X\beta^* - X\beta\|$ und $ab \leq (a/2)^2 + b^2$, so dass gilt

$$\begin{aligned}
5\lambda\|\widehat{\beta} - \beta^*\| &\leq 5\lambda\sqrt{\frac{|S^*|}{\lambda_{\min}(X^\top X)}}\|X\widehat{\beta} - X\beta^*\| \\
&\leq 5\lambda\sqrt{\frac{|S^*|}{\lambda_{\min}(X^\top X)}}\big(\|X\widehat{\beta} - X\beta\| + \|X\beta^* - X\beta\|\big) \\
&\leq 2\frac{25\lambda^2}{4}\frac{|S^*|}{\lambda_{\min}(X^\top X)} + \|X\widehat{\beta} - X\beta\|^2 + \|X\beta^* - X\beta\|^2,
\end{aligned}$$

woraus die Behauptung folgt. □

Wir müssen nun $\mathbb{P}(G^c)$ kontrollieren, um eine Gültigkeit der Orakelungleichung mit hoher Wahrscheinlichkeit zu garantieren. Sei dazu $\sigma_{\max}^2 := \max_{j\leq p} \operatorname{Var}(X^\top\varepsilon)_j$, d.h. $(X^\top\varepsilon)_j \sim N(0, \sigma_j^2)$ mit $\sigma_j^2 \leq \sigma_{\max}^2$. Dann gilt mit einer Union-bound, dass

$$\begin{aligned}
\mathbb{P}(G^c) = \mathbb{P}\Big(\bigcup_{j=1}^p \{|(X^\top\varepsilon)_j| > \lambda/8\}\Big) &\leq \sum_{j=1}^p \mathbb{P}\{|(X^\top\varepsilon)_j| > \lambda/8\} \\
\leq p\mathbb{P}\{|N(0,\sigma_{\max}^2)| > \lambda/8\} &\leq 2p\exp\Big(\frac{-\lambda^2}{2\cdot 64\sigma_{\max}^2}\Big) = 2\exp\Big(\frac{-\lambda^2}{128\sigma_{\max}^2} + \ln p\Big)
\end{aligned}$$

Für $\lambda^2 := 128\tau^2\sigma_{\max}^2 \ln p, \tau > 1$ gilt also

$$\mathbb{P}(G^c) \leq 2e^{-(\tau^2-1)\ln p} = 2p^{-(\tau^2-1)}$$

Dies führt auf folgendes Korollar:

Corollar 5.22 (Ungleichung in hoher Wahrscheinlichkeit). *Falls $\lambda_{\min}(X^\top X) > 0$ und $\lambda^2 = 128\tau^2\sigma_{\max}^2 \ln p$ für $\tau > 1$, gilt mit Wahrscheinlichkeit größer als $1 - 2p^{-(\tau^2-1)}$*

$$\|X\widehat{\beta} - X\beta\|^2 \leq \inf_{\beta^*\in\mathbb{R}^p}\Big(\frac{5}{3}\|X\beta^* - X\beta\|^2 + \frac{1600}{3}\frac{\tau^2\sigma_{\max}^2\ln p}{\lambda_{\min}(X^\top X)}|S^*|\Big).$$

Dabei gilt stets, dass $\sigma_{\max}^2 \leq \sigma^2\lambda_{\max}(X^\top X)$.

Beweis. Die Ungleichheit folgt sofort aus der Orakelungleichung 5.21 und der obigen Überlegung wenn wir den nicht negativen Term $\lambda\|\widehat{\beta}\|_{S^{*c}}$ wegfallen lassen. Der Zusatz folgt aus

$$\sigma_{\max}^2 = \max_{j\le p} \operatorname{Var}(X^\top \varepsilon)_j = \max_{j\le p} \sigma^2 (X^\top X)_{jj} = \max_{j\le p} \sigma^2 \langle e_j, X^\top X e_j\rangle \le \sigma^2 \lambda_{\max}(X^\top X).$$

□

Bemerkung 5.23 (Interpretation der Orakelungleichung).

(a) Insbesondere liefert $\beta^* = \beta$ in der Orakelungleichung, dass

$$\|X\widehat{\beta} - X\beta\|^2 \le \frac{1600}{3} \frac{\tau^2 \sigma_{\max}^2 \ln p}{\lambda_{\min}(X^\top X)} |\{j : \beta_j \neq 0\}|$$

Wenn die aktive Menge $\{j : \beta_j \neq 0\}$ des wahren Parameters bekannt wäre, so würde man das lineare Modell darauf einschränken und als Vorhersagefehler $\sigma^2|j : \beta_j \neq 0|$ in Erwartung erhalten. Der LASSO-Schätzer erreicht diesen Wert also bis auf einen Faktor $O(\frac{\lambda_{\max}}{\lambda_{\min}} \ln p)$.

(b) Oft hat ein wahres β viele Einträge, die nur nahe bei null sind. Dann wird man in der Orakelungleichung β^* so wählen, dass $\beta_j^* = \beta_j$ für große $|\beta_j|$ und $\beta_j^* = 0$ für kleine $|\beta_j|$ gilt. Damit verringert man den zweiten Term mit $|S^*|$ auf Kosten eines Bias $\|X\widehat{\beta} - X\beta^*\|^2$. Dies ist ein wichtiges Robustheitsmerkmal des LASSO für β die nur approximativ sparse sind.

(c) Die Konditionszahl $\lambda_{\max}/\lambda_{\min}$ auf der rechten Seite kommt aus einer pessimistischen Abschätzung und ist in vielen Fällen sehr groß. Dies kann sehr verfeinert werden, indem man die Abbildungseigenschaften von X nur auf $b \in \mathbb{R}^p$ betrachtet wobei b sparse ist. Dies führt z.B. auf die Restricted Isometry property R.I.P.

(d) Es gibt vielerlei Modifikation von LASSO und auch weitergehende Resultate, z.B. zu $\mathbb{P}(\{j : \widehat{\beta}_j \neq 0\} =, \le, \ge \{j : \beta_j \neq 0\})$ oder $\|\widehat{\beta} - \beta\|^2$.