Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

1. Übungsblatt

1. Betrachten Sie den MSE für Schätzer $\hat{\vartheta}$ eines Parameters $\vartheta \in \mathbb{R}^d$. Formulieren und beweisen Sie ein Analogon der Bias-Varianz-Zerlegung für beliebige Dimension $d \geqslant 1$.

2. Wir betrachten eine gleichmäßig auf dem Intervall $[a, b]$ verteilte mathematische Stichprobe $X_1, \dots, X_n$ (d.h. $X_1, \dots, X_n \sim U([a, b])$ i.i.d.) mit unbekannten Parametern $-\infty < a < b < \infty$.

   (a) Formalisieren Sie das statistische Modell.

   (b) Bestimmen Sie Maximum-Likelihood-Schätzer für a und b.

   (c) Welchen MSE haben die Maximum-Likelihood-Schätzer?

3. Die Beta-Verteilung $B(a, b)$ auf $[0, 1]$ ist gegeben durch die Dichte

$$f_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in (0,1),$$

   wobei $a, b > 0$ und Γ die Gamma-Funktion bezeichnet. $B(a, b)$ hat Erwartungswert $\mu_{a,b} = \frac{a}{a+b}$ und Varianz $\sigma^2_{a,b} = \frac{ab}{(a+b)^2(a+b+1)}$.

   (a) Skizzieren Sie $f_{a,b}$ für $(a, b) \in \{0.5; 1; 10\}^2$ (Computereinsatz gestattet).

   (b) Die Beobachtung X sei $\mathrm{Bin}(n, p)$-verteilt, wobei $n \geqslant 1$ bekannt ist sowie der unbekannte Parameter p gemäß $B(a, b)$ a priori verteilt ist. Zeigen Sie, dass der Bayes-optimale Schätzer (bzgl. MSE und $B(a, b)$) gegeben ist durch

$$\hat{p}_{a,b} = \frac{a + X}{a + b + n}.$$

   (c) Bestimmen Sie $MSE_p(\hat{p}_{a,b})$ für $p \in [0, 1]$. Finden Sie $a^*, b^* > 0$ so, dass $p \mapsto MSE_p(\hat{p}_{a^*,b^*})$ konstant ist.

   (d) Zeigen Sie mittels Bayes-Optimalität, dass $p \mapsto MSE_p(\hat{p}_{a^*,b^*})$ konstant impliziert, dass $\hat{p}_{a^*,b^*}$ minimax-Schätzer ist. Folgern Sie, dass der MLE $\hat{p} = \frac{X}{n}$ nicht minimax ist.

4. In einem Krankenhaus soll zur Hebammenplanung mit 95% Sicherheit eine Obergrenze für die Verteilung der Geburtenzahl pro Tag angegeben werden. Bekannt sind die Geburtenzahlen $N_1, \dots, N_n$ der vergangenen n Tage.

   (a) Begründen Sie, weshalb $N_1, \dots, N_n$ näherungsweise als unabhängig und Poiss(λ)-verteilt mit unbekanntem Parameter $\lambda > 0$ angesehen werden können. Geben Sie das entsprechende statistische Modell an.

   (b) Untersuchen Sie in dem Modell den Schätzer $\hat{\lambda} = (N_1 + \cdots + N_n)/n$ auf Erwartungstreue, Konsistenz und MSE. Weisen Sie nach, dass $\hat{\lambda}$ MLE von λ ist.

   (c) Zeigen Sie für große n unter Verwendung der Normalapproximation und des $N(0,1)$-Quantils q_α, dass

$$I = \left[0, \hat{\lambda} + \frac{q_{0,95}}{\sqrt{n}} \sqrt{\hat{\lambda} + \frac{q_{0,95}^2}{4n}} + \frac{q_{0,95}^2}{2n}\right]$$

   approximativ ein einseitiges 95%-Konfidenzintervall für λ ist. Für $n \to \infty$ und

$$\tilde{I}_n = \left[0, \hat{\lambda} + \frac{\sqrt{\hat{\lambda}}\, q_{0,95}}{\sqrt{n}}\right]$$

   gilt $\lim_{n\to\infty} \mathbb{P}_\lambda(\lambda \in \tilde{I}_n) = 0,95$, $\lambda > 0$ beliebig ($\tilde{I}_n$ ist asymptotisches 95%-Konfidenzintervall).
   Hinweis: Sie dürfen das Lemma von Slutsky aus Stochastik II bzw. der Vorlesung verwenden.

5.* Praktische Aufgabe.

   (a) Lesen Sie Kapitel 2.3 “Introduction to R” im Buch “Introduction to Statistical Learning” aus der Literaturliste, um sich mit den grundlegenden Funktionen von R vertraut zu machen. Die dort verwendete Datei “Auto.data” finden Sie unter `www-bcf.usc.edu/~gareth/ISL/Auto.data`.

   (b) Bearbeiten Sie Aufgabe 2.4.10 aus demselben Buch.

Die Abgabe (in Zweier-Gruppen) erfolgt **vor** der Vorlesung am Donnerstag, 1.11.18.
Klausurtermin: Do 14.2.19, 9-11 Uhr; Ersatz: Mo 8.4.19, 13-15 Uhr

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

2. Übungsblatt

1. Beweisen Sie den *Korrespondenzsatz* für Tests und Konfidenzbereiche:

   (a) Ist $C_{1-\alpha}$ eine $(1-\alpha)$-Konfidenzmenge für $\vartheta \in \Theta$, so ist für jedes $\vartheta_0 \in \Theta$

   $$\varphi^{(\vartheta_0)}(x) = \mathbf{1}(\vartheta_0 \notin C_{1-\alpha}(x))$$

   ein Test vom Niveau α für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$.

   (b) Ist andererseits $(\varphi^{(\vartheta_0)})_{\vartheta_0 \in \Theta}$ eine Familie von nicht-randomisierten Tests für $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$ zum Niveau α, so ist

   $$C_{1-\alpha}(x) = \{\vartheta_0 \in \Theta \,|\, \varphi^{(\vartheta_0)}(x) = 0\}$$

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

2. Es sei X $\mathrm{Bin}(n,p)$-verteilt mit $p \in [0,1]$ unbekannt. Zeigen Sie:

   (a) Für $p_0 \in (0,1)$ und das Testproblem $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$ hat der Likelihood-Quotienten-Test die Form

   $$\varphi_\alpha = \mathbf{1}\Big(\frac{\hat{p}^X(1-\hat{p})^{n-X}}{p_0^X(1-p_0)^{n-X}} > c_\alpha\Big) + \gamma_\alpha \mathbf{1}\Big(\frac{\hat{p}^X(1-\hat{p})^{n-X}}{p_0^X(1-p_0)^{n-X}} = c_\alpha\Big)$$

   mit $\hat{p} = X/n$ und $c_\alpha \geqslant 0$, $\gamma_\alpha \in [0,1]$ geeignet.

   (b) Eine Münze wird sechsmal unabhängig geworfen und die Anzahl von 'Kopf' notiert. Bestimmen Sie einen Likelihood-Quotienten-Test für die Hypothese einer fairen Münze ($p_0 = 0,5$), der das Niveau $\alpha = 0,05$ exakt ausschöpft.

   (c) 2015 wurden in Berlin 38 030 Kinder geboren, darunter 19 614 Jungen. Die Wahrscheinlichkeit einer Jungengeburt sei p. Bestimmen Sie für den Test auf $H_0 : p = 0,5$ die kritischen Werte mittels Normalapproximation und klären Sie, ob bei den Zahlen für Berlin die Nullhypothese zum Niveau $\alpha = 5\%$ abgelehnt wird.

3. p-Werte.

   (a) Recherchieren Sie die Definition des p-Wertes eines Tests und erläutern Sie dies kurz in eigenen Worten.

   (b) Bestimmen Sie den p-Wert in Aufgabe 2(c).

4. Betrachten Sie das lineare Modell $Y = X\beta + \varepsilon$, wobei die Fehler ε $N(0, \sigma^2 E_n)$-verteilt seien. Bestimmen Sie den Likelihood-Quotiententest für die Nullhypothese $H_0 : \beta = \beta_0$ gegen $H_1 : \beta \neq \beta_0$ zum Niveau $\alpha \in (0, 1)$. Hierbei seien zunächst $X \in \mathbb{R}^{n \times p}, \beta_0 \in \mathbb{R}^p, \sigma > 0$ bekannt und vorgegeben sowie dann $X \in \mathbb{R}^{n \times p}, \beta_0 \in \mathbb{R}^p$ vorgegeben, aber $\sigma > 0$ unbekannt.

Die Abgabe (in Zweier-Gruppen) erfolgt **vor** der Vorlesung am Donnerstag, 8.11.18.
Korrektor: Marvin van Stegen, marvin@vanstegen.eu
Sprechzeit: Fr 12:00-12:30 Uhr, Raum 1.104, RUD25

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

3. Übungsblatt

1. Eine physikalische Größe $\mu \in \mathbb{R}$ wird in n verschiedenen Apparaturen gemessen. Für die Messwerte $Y_1, \dots, Y_n$ nehmen wir $\mathbb{E}[Y_i] = \mu$, $\mathrm{Var}(Y_i) = \sigma_i^2$ mit $\sigma_i > 0$ bekannt und $\mathrm{Cov}(Y_i, Y_j) = 0$ für $i \neq j$ an. Schreiben Sie dies als lineares Modell und weisen Sie nach, dass das gewichtete Mittel

$$\hat{\mu} = \frac{\sum_{i=1}^n \sigma_i^{-2} Y_i}{\sum_{i=1}^n \sigma_i^{-2}}$$

   Kleinste-Quadrate-Schätzer ist. Vergleichen Sie die Varianz von $\hat{\mu}$ mit der des Stichprobenmittels $\bar{Y}$ und beweisen Sie dann formal, dass $\hat{\mu}$ kleinste Varianz unter allen linearen erwartungstreuen Schätzern besitzt.

2. Es sei $Y = X\beta + \varepsilon$ ein gewöhnliches lineares Modell mit $\varepsilon \sim N(0, \sigma^2 E_n)$. Wir untersuchen den Bayesansatz unter der a-priori-Verteilung $\beta \sim N(0, \tau^2 E_p)$, wobei $\sigma, \tau > 0$. Zeigen Sie:

   (a) Die a-posteriori-Verteilung des Parameters β ist $\beta \sim N(\mu_\beta, \Sigma_\beta)$ mit $\Sigma_\beta = (\sigma^{-2} X^\top X + \tau^{-2} E_p)^{-1}$, $\mu_\beta = \sigma^{-2} \Sigma_\beta X^\top Y$. Inwiefern ist das Verhalten für $\tau \to 0$ bzw. $\sigma \to 0$ jeweils plausibel?

   (b) Der Bayesschätzer ist ein *penalisierter* Kleinste-Quadrate-Schätzer (auch *ridge regression estimator* genannt)

$$\hat{\beta}_{Bayes} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \Big(\|Y - X\beta\|^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2 \Big).$$

   (c) Sind $\lambda_1 \geqslant \cdots \geqslant \lambda_p > 0$ die Eigenwerte von $X^\top X$, so ist λ_p^{-1} der größte Eigenwert von der Kovarianzmatrix $\mathrm{Cov}(\hat{\beta})$ des Kleinste-Quadrate-Schätzers sowie $\max_i \lambda_i (\lambda_i + \sigma^2/\tau^2)^{-2}$ von $\mathrm{Cov}(\hat{\beta}_{Bayes})$. Was bedeutet das statistisch für die plug-in-Schätzer von $\gamma = \langle \beta, v \rangle$ ($v \in \mathbb{R}^p$ bekannt)?

3. Betrachten Sie das statistische Modell

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

mit n verschiedenen Punkten $x_i \in \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ beliebig und unbekannt sowie $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 E_n$. Unter der polynomiellen Regressionsannahme $f = q_a^d$ mit

$$q_a^d(x) = a_0 + a_1 x + \cdots + a_d x^d$$

sei $\hat{a} = (\hat{a}_0, \dots, \hat{a}_d)$ der Kleinste-Quadrate-Schätzer sowie $q_{\hat{a}}^d(x)$ die Vorhersage des Funktionswerts an der Stelle $x \in \mathbb{R}$. Zeigen Sie unter der Verwendung der *empirischen Norm* $\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^n g(x_i)^2$ einer Funktion $g : \mathbb{R} \to \mathbb{R}$:

   (a) Unter der Annahme $f = q_a^d$ gilt $\mathbb{E}_{q_a^d}[\|q_{\hat{a}}^d - q_a^d\|_n^2] = \frac{\sigma^2}{n}(d+1)$.
   Tipp: Betrachten Sie $\|X\hat{\beta} - X\beta\|^2$ allgemein im linearen Modell.

   (b) Für f beliebig im Modell ergibt sich die Bias-Varianz-Zerlegung

$$\mathbb{E}_f[\|q_{\hat{a}}^d - f\|_n^2] = \min_{q_a^d} \|f - q_a^d\|_n^2 + \frac{\sigma^2}{n}(d+1),$$

   wobei sich der Bias durch den kleinsten Abstand von f zum Raum der Polynome vom maximalen Grad d in empirischer Norm ergibt.

   (c) Gilt für jedes quadratische Polynom f, dass der MSE $\mathbb{E}_f[\|q_{\hat{a}}^d - f\|_n^2]$ für $d = 1$ größer als für $d = 2$ ist?

4. Melanom-Prognose: Lösen Sie Aufgabe 3.5.2 auf Seite 100 des Buchprojekts.

Abgabe **vor** der Vorlesung am Donnerstag, 15.11.18.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

4. Übungsblatt

1. Bestimmen Sie den Likelihood-Quotiententest für die lineare Hypothese $H_0 : K\beta = c$ gegen $H_1 : K\beta \neq c$ im linearen Modell $Y = X\beta + \varepsilon$ mit $\varepsilon \sim N(0, \sigma^2 E_n)$, $\sigma > 0$ unbekannt, und zeigen Sie, dass er mit dem F-Test aus der Vorlesung übereinstimmt.

   Zeigen Sie entsprechend, dass auch der F-Test bei der einfaktoriellen Varianzanalyse einem Likelihood-Quotiententest entspricht.

2. Untersuchen Sie im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ unter der Normalverteilungsannahme $\varepsilon \sim \mathcal{N}(0, \sigma^2 E_n)$ mit unbekanntem $\beta \in \mathbb{R}^p$ und $\sigma > 0$ folgende lineare Testprobleme:

   (a) Test auf signifikanten Einfluss einer Kovariablen auf die Responsevariable:

   $$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

   für ein fest vorgegebenes $j \in \{1, \dots, p\}$.

   (b) Test eines Subvektors $\beta^* = (\beta_1^*, \dots, \beta_r^*)^\top \in \mathbb{R}^r$ mit $r \leqslant p$:

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

   Stellen Sie diese Hypothesen in der Form $H_0 : K\beta = c$ mit geeigneter Kontrastmatrix K und einem Vektor c dar. Konstruieren Sie die zugehörigen Fisher-Statistiken und bestimmen Sie deren Verteilung.

3. Das Modell der einfaktoriellen Kovarianzanalyse (ANCOVA: analysis of covariance) mit $p \in \mathbb{N}$ Beobachtungsgruppen vom Umfang $n_1, \ldots, n_p$ lautet

$$Y_{ij} = \mu_i + \kappa x_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots p,$$

mit unbekannten Gruppenmittelwerten $\mu_1, \ldots, \mu_p \in \mathbb{R}$, unbekanntem Regressionskoeffizienten $\kappa \in \mathbb{R}$, welcher die Abhängigkeit von einem Regressorvektor $x = (x_{ij}) \in \mathbb{R}^{\sum_{i=1}^p n_i}$ angibt und voneinander unabhängigen Störgrößen $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ mit (unbekannter) Varianz $\sigma^2 > 0$.

   (a) Bestimmen Sie die Designmatrix und den Parametervektor des zugehörigen linearen Modells und stellen Sie durch eine (notwendige und hinreichende) Bedingung an x sicher, dass die Designmatrix vollen Rang hat.

   (b) Bestimmen Sie den Kleinste-Quadrate-Schätzer $\hat{\beta} = (\hat{\mu}_1, \ldots, \hat{\mu}_p, \hat{\kappa})^\top$.

   (c) Verifizieren Sie die Streuungszerlegung

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = &\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{\kappa}x_{ij} - (\bar{Y}_{i\bullet} - \hat{\kappa}\bar{x}_{i\bullet})\right)^2 \\ &+ \sum_{i=1}^{p} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \hat{\kappa}^2 \sum_{i=1}^{p}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\bullet})^2\end{aligned}$$

   in eine Residuenvarianz innerhalb der Gruppen, eine Stichprobenvarianz zwischen den Gruppen, und eine Regressionsvarianz. Bestimmen Sie die Fisher-Statistik für einen Test der Hypothese $H_0 : \mu_1 = \cdots = \mu_p$.

4. Praktische Aufgabe: Klimaentwicklung.
   Arbeiten Sie Beispiel 3.2.20 auf Seite 86 des Buchprojekts zu den jährlichen Temperaturdaten durch. Implementieren Sie die angesprochenen Testprobleme in einer Statistik-Software und bestimmen Sie die p-Werte. Verifizieren Sie die im Buch angegebenen Testentscheidungen.

Abgabe **vor** der Vorlesung am Donnerstag, 22.11.18.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

5. Übungsblatt

1. Im gewöhnlichen linearen Modell $Y = X\beta + \varepsilon$ gelte $X_1 = (1, \dots, 1)^\top$ für die erste Spalte der Designmatrix X (β_1 ist ein *offset*-Parameter). Bezeichnet $\bar{Y}$ den Mittelwert der Beobachtungen und $\hat{Y} = X\hat{\beta}$ den geschätzten Regressanden, so heißt
$$R^2 := \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
Bestimmtheitsmaß. Zeigen Sie, dass R^2 in $[0, 1]$ liegt. Erklären Sie (mit mathematischer Begründung), was die Fälle $R^2 = 0$ und $R^2 = 1$ statistisch aussagen.

2. Beweisen oder widerlegen Sie die Aussage, dass folgende Verteilungen Exponentialfamilien bilden. Bestimmen Sie gegebenenfalls den natürlichen Parameterraum.
   (a) Multinomialverteilung $(M(p_0, \dots, p_s; n))_{0<p_i<1, \sum_{i=1}^s p_i = 1}$;
   (b) p-dimensionale Normalverteilung $(N(\mu, \Sigma))_{\mu \in \mathbb{R}^p}$ mit bekannter Kovarianzmatrix $\Sigma \in \mathbb{R}^{p \times p}$;
   (c) Gleichmäßige Verteilung $(U([0, \vartheta]))_{\vartheta > 0}$;
   (d) Gammaverteilung $(\Gamma(a, b))_{a,b>0}$.

3. Es sei $(P_\vartheta)_{\vartheta \in \Theta}$ eine natürliche Exponentialfamilie, und ein Maximum-Likelihood-Schätzer $\hat{\vartheta}$ liege im Innern von $\Theta \subseteq \mathbb{R}$. Weisen Sie die Identität $\mathbb{E}_{\hat{\vartheta}(x)}[T] = T(x)$ nach. Zeigen Sie ferner, dass $\mathrm{Var}_\vartheta(T) > 0$ für alle $\vartheta \in \Theta$ die Eindeutigkeit von $\hat{\vartheta}$ impliziert,.

4. Geben Sie die Cramér-Rao-Ungleichung samt Voraussetzungen aus der Literatur an. Weisen Sie damit für natürliche Exponentialfamilien nach, dass T erwartungstreuer Schätzer von minimaler Varianz für $\zeta'(\vartheta) = \mathbb{E}_\vartheta[T]$ für ϑ im Innern des natürlichen Parameterbereichs Θ ist. Geben Sie zwei Beispiele dieses Resultats für konkrete Exponentialfamilien.

Abgabe **vor** der Vorlesung am Donnerstag, 29.11.18.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

6. Übungsblatt

1. Beweisen Sie den Satz aus der Vorlesung:
   Im verallgemeinerten linearen Modell mit kanonischer Linkfunktion und den Zeilen $x_1, \dots, x_n \in \mathbb{R}^p$ der Designmatrix X gilt für die Loglikelihoodfunktion
   $$\nabla_\beta \ell(\beta, \varphi) = \tfrac{1}{\varphi} \textstyle\sum_{i=1}^n (Y_i - \zeta'(\langle x_i, \beta \rangle)) x_i.$$
   Ist die Fisher-Informationsmatrix $I(\beta) = \frac{1}{\varphi} \sum_{i=1}^n \zeta''(\langle x_i, \beta \rangle) x_i x_i^\top$ positiv-definit für alle β und existiert ein $\hat\beta$ mit $\nabla_\beta \ell(\hat\beta, \varphi) = 0$ (für irgendein $\varphi > 0$), so ist $\hat\beta$ eindeutig bestimmter Maximum-Likelihood-Schätzer von β.

2. Die Poissonverteilung mit Parameter $\lambda > 0$ bildet eine Exponentialfamilie in $T(k) = k$ mit natürlichem Parameter $\vartheta = \log \lambda \in \mathbb{R}$. Überprüfen Sie die Identitäten $\mathbb{E}_\vartheta[T] = \zeta'(\vartheta)$, $\mathrm{Var}_\vartheta(T) = \zeta''(\vartheta)$.
   Betrachten Sie nun das Modell der Poissonregression mit $\log \lambda_i = a x_i + b$ für gegebene Designpunkte $x_1, \dots, x_n \in \mathbb{R}$. Stellen Sie gemäß Aufgabe 1 eine Gleichung für den MLE auf und untersuchen Sie, ob der MLE existiert und eindeutig ist. Wie sieht die Iteration in Fishers Scoring-Methode aus?

3. Betrachten Sie im Modell der logistischen Regression mit $p_i = \frac{\exp(\langle x_i, \beta \rangle)}{1+\exp(\langle x_i, \beta \rangle)}$, d.h. Zeilen $x_1, \dots, x_n$ der Designmatrix X, den MLE $\hat\beta$ und die Vorhersage $\hat p(x) := \frac{\exp(\langle x, \hat\beta \rangle)}{1+\exp(\langle x, \hat\beta \rangle)}$ für $x \in \mathbb{R}^p$. Bestimmen Sie die geometrische Form der Klassifikationsgrenze $\hat B := \{x \in \mathbb{R}^p \mid \hat p(x) = 1/2\}$.
   Simulieren Sie unabhängige $N(0, E_2)$-verteilte Zufallsvektoren $X_1, \dots, X_n \in \mathbb{R}^2$ als Zeilen der Designmatrix X und damit Beobachtungen $Y_1, \dots, Y_n$ der logistischen Regression mit Parametervektor $\beta = (2, 1)^\top$ und $n = 100$. Stellen Sie Daten, die Funktion $\hat p : \mathbb{R}^2 \to (0, 1)$ sowie die Klassifikationsgrenze $\hat B$ gemeinsam in einem Koordinatensystem dar.

4. Geben Sie das *Probit*-Modell aus der Literatur an und zeigen Sie, dass es sich um ein verallgemeinertes lineares Modell mit nicht-kanonischer Linkfunktion handelt.
   Freiwillig: Vergleichen Sie für die Daten in 3. die Klassifikationsgrenzen, die für logistische Regression und Probit-Modell erhalten werden.

Abgabe **vor** der Vorlesung am Donnerstag, 6.12.18.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

7. Übungsblatt

1. Betrachten Sie für Klassifizier $C : \mathcal{X} \to \{0,1\}$ und $\alpha \in (0,1)$ den Klassifikationsfehler

$$R_\alpha(C) = \alpha P\big(C(X) = 0, Y = 1\big) + (1-\alpha) P\big(C(X) = 1, Y = 0\big).$$

   Bestimmen Sie unter Benutzung von $\eta(x) = P(Y = 1 \,|\, X = x)$ den zugehörigen Bayes-Klassifizierer $C_{B,\alpha}$, der $R_\alpha(C_{B,\alpha}) = \min_C R_\alpha(C)$ erfüllt.

2. Für $w_1, \dots, w_k \geqslant 0$ mit $\sum_{i=1}^k w_i = 1$ sei $\hat{C}_w$ ein gewichteter k-nächste Nachbar-Klassifizierer der Form

$$\hat{C}_w(x) = \mathbf{1}(\hat{\eta}_w(x) \geqslant 1/2) \text{ mit } \hat{\eta}_w(x) = \textstyle\sum_{i=1}^n W_i(x) Y_i,$$

   wobei $W_i(x) = w_j$, falls $X_i = X_{(j)}(x)$, d.h. der j-nächste Nachbar von x wird mit w_j gewichtet.

   Nehmen Sie nun $k = k_n \to \infty$ für $n \to \infty$ mit $k_n/n \to 0$ an und finden Sie eine möglichst allgemeine Bedingung an die Gewichtfolge $(w_1^{(n)}, \dots, w_{k_n}^{(n)})$, unter der Sie zeigen können, dass $\hat{C}_{w^{(n)}}$ konsistent ist.

3. Ein Klassifizierer bei K Klassen ist gegeben durch eine messbare Funktion $C : \mathcal{X} \to \{1, \dots, K\}$ und sein symmetrischer Klassifikationsfehler wiederum durch $R(C) = P(C(X) \neq Y)$, wobei (X, Y) ein $\mathcal{X} \times \{1, \dots, K\}$-wertiger Zufallsvektor ist. Zeigen Sie, dass der Klassifizierungsfehler durch C_B mit

$$C_B(x) = \operatorname{argmax}_{k=1,\dots,K} \eta_k(x), \quad \eta_k(x) = P(Y = k \,|\, X = x)$$

   minimiert wird. Wie groß ist $R(C_B)$?

4. Betrachten Sie für K Klassen mit Klassenwahrscheinlichkeiten $\pi_k \in [0,1]$, $\sum_{k=1}^K \pi_k = 1$, Mittelwerten $\mu_k \in \mathbb{R}^d$ und invertierbaren Kovarianzmatrizen $\Sigma_k \in \mathbb{R}^{d\times d}$ die Normalverteilungsmischungsdichte

$$f(x) = \sum_{k=1}^{K} \pi_k \varphi_{\mu_k,\Sigma_k}(x), \quad x \in \mathbb{R}^d\,.$$

Weisen Sie nach, dass die *decision boundary* zwischen Klasse k und ℓ (bei Standard-Klassifikationsfehler) gegeben ist durch $\{x \in \mathbb{R}^d \mid \delta_k(x) = \delta_\ell(x)\}$ mit quadratischen Diskrimanten

$$\delta_k(x) = -\tfrac{1}{2}\log(\det(\Sigma_k)) - \tfrac{1}{2}\langle \Sigma_k^{-1}(x-\mu_k), x-\mu_k\rangle + \log \pi_k, \quad k = 1,\dots,K.$$

Skizzieren Sie eine typische *decision boundary* im Fall $K = d = 2$.

5. *Freiwillig:* (praktische Aufgabe)

   (a) Lesen Sie Kapitel 4.6 "Lab: Logistic Regression, LDA, QDA and KNN" im Buch *Introduction to Statistical Learning*.

   (b) Bearbeiten Sie Aufgabe 4.7.10(a,d-h) aus demselben Buch.

   (c*) Probieren Sie, bessere Klassifikationsmethoden gemäß Aufgabe 4.7.10(i) zu finden.

Abgabe **vor** der Vorlesung am Donnerstag, 13.12.18.

Vorlesung *Methoden der Statistik*
Wintersemester 2017/18
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

8. Übungsblatt

1. Für einen Klassifizierer C und Beobachtungen $(X_i, Y_i)_{1\leqslant i\leqslant n}$ bezeichnet

$$R_n(C) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \neq C(X_i))$$

das *empirische Risiko*. $\hat{C}$ ist *ERM-Klassifizierer* in einer Klasse $\mathscr{C}$ von Klassifizierern, falls $R_n(\hat{C}) = \min_{C\in\mathscr{C}} R_n(C)$ gilt. Zeigen Sie für das Risiko R (Klassifizierungsfehler)

$$R(\hat{C}) \leqslant \inf_{C\in\mathscr{C}} R(C) + 2\sup_{C\in\mathscr{C}} |R_n(C) - R(C)|.$$

2. Beweisen Sie schrittweise die Hoeffding-Ungleichung (1963): Es seien $Z_1, \dots, Z_n$ unabhängige, reellwertige Zufallsvariablen mit $\mathbb{E}[Z_i] = 0$, $|Z_i| \leqslant R$, $i = 1, \dots, n$. Dann gilt

$$\mathbb{P}(|\textstyle\sum_{i=1}^n Z_i| \geqslant \kappa) \leqslant 2\exp(-\kappa^2/(2nR^2)), \quad \kappa > 0.$$

   (a) Für $\lambda > 0$, $|\delta| \leqslant R$ gilt $e^{\lambda\delta} \leqslant \frac{R-\delta}{2R}e^{-\lambda R} + \frac{R+\delta}{2R}e^{\lambda R}$.

   (b) Dies impliziert $\mathbb{E}[e^{\lambda Z_i}] \leqslant (e^{-\lambda R} + e^{\lambda R})/2 < e^{\lambda^2 R^2/2}$.

   (c) Mit der verallgemeinerten Markov-Ungleichung erhalten wir für $\lambda > 0$

$$\mathbb{P}(\textstyle\sum_{i=1}^n Z_i \geqslant \kappa) \leqslant \exp(-\lambda\kappa + \lambda^2 nR^2/2).$$

   (d) Die letzte Schranke ist minimal für $\lambda = \kappa/(nR^2)$ und mit einem symmetrischen Argument für $-\sum_{i=1}^n Z_i$ folgt die Hoeffding-Ungleichung.

3. Betrachten Sie eine endliche Familie $\mathscr{C} = \{C_1, \dots, C_M\}$ von Klassifizierern und den zugehörigen ERM-Klassifizierer $\hat{C}$. Verwenden Sie Aufgabe 1 und die Hoeffding-Ungleichung, um für alle $\delta \in (0, 1)$ zu zeigen, dass mit Wahrscheinlichkeit $1 - \delta$

$$R(\hat{C}) \leqslant \min_{1\leqslant m\leqslant M} R(C_m) + \frac{\sqrt{8\log(2M/\delta)}}{\sqrt{n}}.$$

4. Betrachten Sie das Optimierungsproblem

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d} \left(\tfrac{1}{2}|\beta|^2 + C \sum_{i=1}^{n} \big(1 - y_i(\langle x_i, \beta \rangle + \beta_0)\big)_+ \right)$$

für gegebene $y_i \in \{-1, +1\}$, $x_i \in \mathbb{R}^d$, $C > 0$. Beweisen Sie:
Die Lösung für β (deren Existenz vorausgesetzt sei) besitzt die Form

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i \text{ mit } \alpha_i \in [0, C].$$

Punkte x_i mit $\alpha_i > 0$ heißen *Stützvektoren* (*support vectors*). Punkte x_i mit $y_i(\langle x_i, \beta \rangle + \beta_0) > 1$ sind keine Stützvektoren (sie liegen jenseits des *margin*). Für Punkte x_i mit $y_i(\langle x_i, \beta \rangle + \beta_0) < 1$ gilt $\alpha_i = C$ und diese x_i sind Stützvektoren.
Hinweis: $f(z) := z_+ = \max(z, 0)$ ist nicht differenzierbar bei $z = 0$, aber für den Differenzenquotienten gilt $\frac{f(h)-f(0)}{h} \in [0, 1]$ für alle $h \neq 0$.

Abgabe **vor** der Vorlesung am Donnerstag, 20.12.18.

Wir wünschen Ihnen frohe Weihnachtstage und einen guten Start ins Jahr 2019!

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

9. Übungsblatt

1. Zeigen Sie für die Kullback-Leibler-Divergenz folgende Eigenschaften:

   (a) $\mathrm{KL}(\mathbb{P}^{\otimes n} \mid \mathbb{Q}^{\otimes n}) = n\,\mathrm{KL}(\mathbb{P} \mid \mathbb{Q})$;

   (b) Für naürliche Exponentialfamilien mit Dichten

   $$\frac{d\,\mathbb{P}_\vartheta}{d\mu}(x) = c(x)\exp\left(\langle T(x),\vartheta\rangle_{\mathbb{R}^k} - \zeta(\vartheta)\right), \quad x \in \mathcal{X},\ \vartheta \in \Theta,$$

   und $\vartheta_0 \in \mathrm{int}(\Theta)$ gilt

   $$\mathrm{KL}(\mathbb{P}_{\vartheta_0} \mid \mathbb{P}_\vartheta) = \zeta(\vartheta) - \zeta(\vartheta_0) - \langle \nabla_\vartheta \zeta(\vartheta_0), \vartheta - \vartheta_0\rangle_{\mathbb{R}^k}, \quad \vartheta \in \Theta.$$

2. Bestimmen Sie die Kullback-Leibler-Divergenzen:

   (a) $\mathrm{KL}(N(\vartheta_0,\Sigma)\,|\,N(\vartheta,\Sigma))$ für $\vartheta_0,\vartheta \in \mathbb{R}^k$ und $\Sigma \in \mathbb{R}^{k\times k}$ symmetrisch, positiv-definit;

   (b) $\mathrm{KL}(\mathrm{Poiss}(\lambda)\,|\,\mathrm{Poiss}(\lambda'))$ für $\lambda,\lambda' > 0$;

   (c) $\mathrm{KL}(\mathrm{Bin}(n,p)\,|\,\mathrm{Bin}(n,p'))$ für $p,p' \in (0,1)$.

3. Der *Totalvariationsabstand* zweier Wahrscheinlichkeitsmaße $\mathbb{P}$ und $\mathbb{Q}$ auf einem messbaren Raum $(\mathcal{X},\mathscr{F})$ ist definiert als

   $$\mathrm{TV}(\mathbb{P},\mathbb{Q}) := \sup_{A\in\mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

   (a) Weisen Sie $\mathrm{TV}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int |p-q|\,\mathrm{d}\nu$, nach für ein dominierendes Maß ν und ν-Dichten $p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\nu}$, $q = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\nu}$.

   (b) Zeigen Sie die *Pinsker-Ungleichung* $\mathrm{TV}(\mathbb{P},\mathbb{Q}) \leqslant \sqrt{\mathrm{KL}(\mathbb{P}|\mathbb{Q})/2}$.
   Hinweis: Betrachten Sie die Funktion $h(z) := z\log(z) - z + 1, z > 0$, mit stetiger Fortsetzung in Null. Zeigen Sie für alle $z \geqslant 0$:

   $$\left(\tfrac{4}{3} + \tfrac{2}{3}z\right)h(z) \geqslant (z-1)^2.$$

   Benutzen Sie anschließend die Cauchy-Schwarz-Ungleichung.

(c) Prüfen Sie ob die Folgen $(\mathbb{P}_n)_{n\in\mathbb{N}}$ auf $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ schwach, im Totalvariationsabstand oder in der Kullback-Leibler-Divergenz gegen einen geeigneten Grenzwert $\mathbb{P}$ konvergieren (betrachten Sie sowohl $\mathrm{KL}(\mathbb{P}_n \,|\, \mathbb{P})$ als auch $\mathrm{KL}(\mathbb{P} \,|\, \mathbb{P}_n)$):

(i) $\mathbb{P}_n = \delta_{1/n}$, Dirac-Maß in $1/n$;

(ii) $\mathbb{P}_n = (1 - \frac{1}{n})\nu + \frac{1}{n}\delta_1$, wobei ν das Maß der $\mathcal{N}(0,1)$-Verteilung ist.

4. Praktische Aufgabe: AIC.

(a) Vollziehen Sie das Beispiel 6.1.15 des Buchprojekts nach, um die Verwendung von AIC bei der Polynomregression zu verstehen.

(b) Simulieren Sie ein Polynomregressionsmodell mit $x_i = i/n$, $i = 1, \ldots, n$, $\varepsilon_i \sim N(0,1)$ i.i.d., und Polynomfunktion $p(x) = x^3 - 2x^2 + x$. Bestimmen Sie für $n = 10$, $n = 100$ und $n = 1000$ in jeweils 100 Monte-Carlo-Iterationen:

i. den *mittleren Vorhersage-Fehler* $E_k := \frac{1}{n}\sum_{i=1}^n (\hat{p}_k(x_i) - p(x_i))^2$ des geschätzten Polynoms $\hat{p}_k(x) = \sum_{j=0}^k \hat{\beta}_j x^j$ vom Grad k für $k = 0, 1, \ldots, 5$,

ii. den von AIC gewählten Polynomgrad $\hat{k}^{AIC}$ und den zugehörigen Fehler $E_{\hat{k}^{AIC}}$ zusammen mit den Orakelfehlern $\min_k E_k$.

Stellen Sie Ihre Ergebnisse mittels Boxplots über die Monte-Carlo-Iterationen dar. Wieso wird $\min_k E_k$ nicht stets bei $k = 3$ angenommen?

(c) Freiwillig: Welche Ergebnisse in (b) ergeben sich, wenn die Daten mittels $Y_i = \sin(\pi x_i) + \varepsilon_i$ erzeugt werden?

Abgabe *vor* der Vorlesung am Donnerstag, 24.1.19.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

Elaboration on the Akaike Information Criterion (AIC)

Let $(\mathcal{X}, \mathscr{F}, (\mathbb{P}_\vartheta^{(k)})_{\vartheta\in\Theta_k})$, $k = 1, \dots, K$, be a collection of K statistical models and let $\mathbb{P}$ be the true underlying probability on $(\mathcal{X}, \mathscr{F})$. Assume that all probabilities $\mathbb{P}_\vartheta^{(k)}$ have likelihoods $L_k(\vartheta)$ with respect to some measure μ and that $\mathbb{P}$ has likelihood L with respect to μ. Moreover, we suppose that we dispose of an i.i.d. sample $X_1, \dots, X_n$, generated under $\mathbb{P}$, for which an MLE $\hat\vartheta_n^{(k)}$ exists within every model $(\mathbb{P}_\vartheta^{(k)})_{\vartheta\in\Theta_k}$, $k = 1, \dots, K$.

Then under suitable regularity conditions $\hat\vartheta_n^{(k)}$ converges $\mathbb{P}-a.s.$ for $n \to \infty$ to the Kullback-Leibler projection (also called information projection) $\operatorname{argmin}_{\vartheta\in\Theta_k} \mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta^{(k)})$. Noting

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{P}_\vartheta^{(k)}) = \mathbb{E}_\mathbb{P}[\log(L/L_k(\vartheta))] = \mathbb{E}_\mathbb{P}[\log(L)] - \mathbb{E}_\mathbb{P}[\log(L_k(\vartheta))],$$

provided all expectations exist, we can equivalently minimise the *Kullback-Leibler discrepancy*

$$d_k(\vartheta) := \mathbb{E}_\mathbb{P}[-2\log(L_k(\vartheta))].$$

The idea of AIC is now to select the model $\hat k_{AIC}$ minimising the KL discrepancy $d_k(\hat\vartheta_n^{(k)})$ between the MLE and $\mathbb{P}$ over $k = 1, \dots, K$. Note that for larger models k the MLE $\hat\vartheta_n^{(k)}$ will have a larger variance and might thus be further away from the KL projection $\operatorname{argmin}_{\vartheta\in\Theta_k} d_k(\vartheta)$. On the other hand, as the model k becomes smaller, the minimal KL discrepancy $\min_{\vartheta\in\Theta_k} d_k(\vartheta)$ will increase (i.e., there is a larger modeling bias). The model $\hat k_{AIC}$ should thus exhibit an MLE which minimises a combination of modeling bias and stochastic variability, resulting in an (almost) optimal error, measured in terms of d_k.

A first (naive) approach to estimate $d_k(\hat\vartheta_n^{(k)})$, which is not known due to the dependence of d_k on $\mathbb{P}$, consists in estimating $d_k(\vartheta)$ by its empirical counterpart $-\frac{2}{n}\sum_{i=1}^n \log(L_k(\vartheta, X_i))$ and then to plug in $\vartheta = \hat\vartheta_n^{(k)}$. The problem is that $\hat\vartheta_n^{(k)}$ depends on $X_1, \dots, X_n$ and thus the estimation yields wrong results. In particular, if the models are nested, then we would always select the largest model $\hat k_{AIC} = K$ because the log-likelihood at the MLE, which is the maximal log-likelihood, increases in k.

Example. Let Θ_k be a k-dimensional subspace of $\mathbb{R}^K$, $\mathbb{P}_\vartheta^{(k)} = N(\vartheta, \sigma^2 E_K)$ for $\vartheta \in \Theta_k$ and $\mathbb{P} = N(\mu, \sigma^2 E_K)$ for some unknown $\mu \in \mathbb{R}^K$. We obtain, using Lebesgue measure μ,

$$-2L_k(\vartheta, x) = K\log(2\pi\sigma^2) + \|x-\vartheta\|^2/\sigma^2, \quad d_k(\vartheta) = K(\log(2\pi\sigma^2)+1) + \|\mu-\vartheta\|^2/\sigma^2$$

and with the orthogonal projection Π_k of $\mathbb{R}^K$ onto Θ_k and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$

$$\hat{\vartheta}_n^{(k)} = \Pi_k \bar{X}, \quad -\frac{2}{n}\sum_{i=1}^n \log(L_k(\hat{\vartheta}_n^{(k)}, X_i)) = K\log(2\pi\sigma^2) + \frac{1}{n\sigma^2}\sum_{i=1}^n \|X_i - \Pi_k\bar{X}\|^2.$$

To see that this is not a good estimate of $d_k(\hat{\vartheta}_n^{(k)})$, let us compare the expectations, writing $X_i = \mu + \sigma\varepsilon_i$ with $\varepsilon \sim N(0, E_K)$ under $\mathbb{P}$:

$$\begin{aligned}
\mathbb{E}_\mathbb{P}[d_k(\hat{\vartheta}_n^{(k)})] &= K(\log(2\pi\sigma^2)+1) + \sigma^{-2}\,\mathbb{E}_\mathbb{P}[\|\mu - \Pi_k\bar{X}\|^2] \\
&= K(\log(2\pi\sigma^2)+1) + \sigma^{-2}\|(E_K - \Pi_k)\mu\|^2 + \mathbb{E}[\|\Pi_k\bar{\varepsilon}\|^2] \\
&= K(\log(2\pi\sigma^2)+1) + \sigma^{-2}\|(E_K - \Pi_k)\mu\|^2 + kn^{-1},
\end{aligned}$$

whereas

$$\begin{aligned}
&\mathbb{E}_\mathbb{P}\Big[-\frac{2}{n}\sum_{i=1}^n \log(L_k(\hat{\vartheta}_n^{(k)}, X_i))\Big] = K\log(2\pi\sigma^2) + \sigma^{-2}\,\mathbb{E}_\mathbb{P}[\|X_1 - \Pi_k\bar{X}\|^2] \\
&= K\log(2\pi\sigma^2) + \sigma^{-2}\,\mathbb{E}_\mathbb{P}[\|(E-\Pi_k)X_1\|^2 + \|\Pi_k(X_1 - \bar{X})\|^2] \\
&= K\log(2\pi\sigma^2) + \sigma^{-2}\Big(\|(E-\Pi_k)\mu\|^2 + \sigma^2(K-k) + \sigma^2\,\mathbb{E}_\mathbb{P}[\|\Pi_k(\varepsilon_1 - \bar{\varepsilon})\|^2]\Big) \\
&= K\log(2\pi\sigma^2) + \sigma^{-2}\|(E-\Pi_k)\mu\|^2 + (K-k) + k(n-1)/n \\
&= K(\log(2\pi\sigma^2)+1) + \sigma^{-2}\|(E-\Pi_k)\mu\|^2 - kn^{-1}.
\end{aligned}$$

So, the two expectations differ by the value $2kn^{-1}$, which does not depend on any unknown quantity. For AIC this bias is now corrected and the model selection rule is

$$\begin{aligned}
\hat{k}_{AIC} &:= \mathrm{argmin}_{k=1,\dots,K}\Big(-\frac{2}{n}\sum_{i=1}^n \log(L_k(\hat{\vartheta}_n^{(k)}, X_i)) + 2kn^{-1}\Big) \\
&= \mathrm{argmin}_{k=1,\dots,K}\Big(-2\sum_{i=1}^n \log(L_k(\hat{\vartheta}_n^{(k)}, X_i)) + 2k\Big) \\
&= \mathrm{argmin}_{k=1,\dots,K}\Big(\sum_{i=1}^n \|X_i - \Pi_k\bar{X}\|^2 + 2k\sigma^2\Big) \\
&= \mathrm{argmin}_{k=1,\dots,K}\Big(n\|(E_K - \Pi_k)\bar{X}\|^2 + 2k\sigma^2\Big),
\end{aligned}$$

where we used $\sum_i \|X_i - \bar{X} + \bar{X} - \Pi_k\bar{X}\|^2 = \sum_i \|X_i - \bar{X}\|^2 + \sum_i \|(E_K - \Pi_K)\bar{X}\|^2$ because the sum over the cross term in the binomial formula vanishes. Later we shall interpret the first term as the residual sum of squares of the MLE/LSE $\Pi_k\bar{X}$ and the second term as twice the variance in the corresponding linear model.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

10. Übungsblatt

1. Folgen Sie dem Ansatz der *unbiased risk estimation (URE)*, um im linearen Modell das AIC-Modellwahlkriterium herzuleiten: es gelte $Y = X^{(k)}\beta^{(k)} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 E_n)$ mit bekanntem $\sigma^2 > 0$ und unbekanntem $\beta^{(k)} \in \mathbb{R}^k$. Unter der wahren Verteilung P gelte $Y = \mu + \varepsilon$ mit einem $\mu \in \mathbb{R}^n$. Zeigen Sie

   (a) Der empirische Vorhersagefehler $\hat{R}_k := \|Y - X^{(k)}\hat{\beta}^{(k)}\|^2$ des KQ-Schätzers $\hat{\beta}^{(k)}$ erfüllt $\mathbb{E}_P[\hat{R}_k] = \|(E_n - \Pi_{X^{(k)}})\mu\|^2 + \sigma^2(n-k)$.

   (b) Für den wahren Vorhersagefehler $R_k := \mathbb{E}_P[\|\mu - X^{(k)}\hat{\beta}^{(k)}\|^2]$ von $\hat{\beta}^{(k)}$ gilt

   $$R_k = \mathbb{E}_P[\hat{R}_k] + \sigma^2(2k - n).$$

   (c) Durch $\hat{R}_k + \sigma^2(2k-n)$ wird R_k erwartungstreu geschätzt, und es gilt $\hat{k}^{AIC} = \operatorname{argmin}_k(\hat{R}_k + \sigma^2(2k-n)) = \operatorname{argmin}_k(\hat{R}_k + 2\sigma^2 k)$.

2. Betrachten Sie für lineare Modelle $Y = X^{(p)}\beta^{(p)} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 E_n)$, $X^{(p)} \in \mathbb{R}^{n\times p}$ von vollem Rang, die AIC-Modellwahl $\hat{p}_{AIC}(\sigma^2)$ bei bekanntem $\sigma^2 > 0$.

   (a) Im Fall geordneter Modelle, d.h. $\operatorname{range}(X^{(p)}) \subseteq \operatorname{range}(X^{(p+1)})$ für alle p, untersuchen Sie, ob $\hat{p}_{AIC}(\sigma^2)$ in σ^2 wachsend oder fallend ist. Erklären Sie, warum Ihr Ergebnis plausibel ist.

   (b) Mit $\tilde{k}_{AIC}$ sei die AIC-Modellwahl bei unbekanntem $\sigma^2 > 0$ mit Parameterdimension $k = p+1$ bezeichnet. Vergleichen Sie $\tilde{k}_{AIC} - 1$ mit $\hat{p}_{AIC}(\hat{\sigma}_p^2)$, wobei der MLE $\hat{\sigma}_p^2 = RSS_p/n$ eingesetzt wird.

3. Betrachten Sie für lineare Modelle $Y = X^{(p)}\beta^{(p)} + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 E_n)$, $X^{(p)} \in \mathbb{R}^{n\times p}$ von vollem Rang, das AIC-Modellwahlkriterium $AIC(k)$ und die KL-Diskrepanz d_k bei unbekanntem Parameter $\vartheta = (\beta^{(p)}, \sigma^2) \in \mathbb{R}^k$, $k = p+1 \leqslant n-2$. Weisen Sie nach

   $$\mathbb{E}_P[AIC(k)] = \mathbb{E}_P[d_k(\hat{\vartheta}_k)] - 2\frac{k(k+1)}{n-k-1}$$

   für den Fall, dass unter P gilt $Y = X^{(p)}\beta_0^{(p)} + \varepsilon$ für ein $\beta_0^{(p)} \in \mathbb{R}^p$ (die wahre Verteilung liegt also im p-ten Modell, aber es wird nicht exakt erwartungstreu geschätzt).
   Hinweis: Für $Z \sim \chi^2(n-p)$ gilt $\mathbb{E}[1/Z] = (n-p-2)^{-1}$, $n \geqslant p+3$.

4. Klären Sie, ob die Modellwahl mittels AIC oder BIC ein kleineres Modell wählt. Wiederholen Sie Aufgabe 4 vom 9. Übungsblatt mit BIC anstelle von AIC und vergleichen Sie sowohl die Fehler $\hat{E}_{AIC}$ und $\hat{E}_{BIC}$ als auch die relativen Häufigkeiten der Ereignisse $\{\hat{k}^{AIC} = k\}$, $\{\hat{k}^{BIC} = k\}$ für $k = 0, \ldots, 5$.

Abgabe *vor* der Vorlesung am Donnerstag, 31.1.19.

Vorlesung *Methoden der Statistik*
Wintersemester 2018/19
Humboldt-Universität zu Berlin

Prof. Dr. Markus Reiß

Probeklausur

1. Es seien $X_1, \dots, X_n \sim N(\vartheta_0, \sigma^2)$, $Y_1, \dots, Y_n \sim N(\vartheta_1, \sigma^2)$ unabhängige Beobachtungen zu unbekannten Parametern $\vartheta_0, \vartheta_1 \in \mathbb{R}$ für $\sigma > 0$ bekannt.

   (a) Geben Sie das zugehörige statistische Modell und die Likelihoodfunktion an.

   (b) Beschreiben Sie kurz die allgemeine Form eines Likelihood-Quotiententests und erklären Sie, in welchem Sinne er für einelementige ('einfache') Nullhypothese und Alternative optimal ist.

   (c) Weisen Sie nach, dass $\varphi = \mathbf{1}(\bar{Y} - \bar{X} \geqslant c_\alpha)$ für $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n$ und $c_\alpha \in \mathbb{R}$ ein Likelihood-Quotienten-Test für das Testproblem $H_0 : \vartheta_1 \leqslant \vartheta_0$ gegen $H_1 : \vartheta_1 > \vartheta_0$ ist.

   (d) Finden Sie für $\alpha \in (0,1)$ den kritischen Wert $c_\alpha \in \mathbb{R}$, so dass φ Niveau α besitzt.

2. In einem Krankenhaus soll zur Hebammenplanung mit 95% Sicherheit eine Obergrenze für die Verteilung der Geburtenzahl pro Tag angegeben werden. Bekannt sind die Geburtenzahlen $N_1, \dots, N_n$ der vergangenen n Tage.

   (a) Begründen Sie, weshalb $N_1, \dots, N_n$ näherungsweise als unabhängig und Poiss(λ)-verteilt mit unbekanntem Parameter $\lambda > 0$ angesehen werden können. Geben Sie das entsprechende statistische Modell an.

   (b) Bestimmen Sie in dem Modell den Maximum-Likelihood-Schätzer $\hat{\lambda}$.

   (c) Ist λ erwartungstreu und/oder konsistent für $n \to \infty$? Berechnen Sie den MSE.

   (d) Zeigen Sie für $n \to \infty$ und $\tilde{I}_n = [0, \hat{\lambda} + \sqrt{\hat{\lambda}}\, q_{0,95}/\sqrt{n}]$, dass $\tilde{I}_n$ ein asymptotisches 95%-Konfidenzintervall ist.
   Hinweis: q_α bezeichnet das α-Quantil von $N(0,1)$ und Sie dürfen das Lemma von Slutsky verwenden.

3. Logistische Regression.

   (a) Es seien $Y_1, \dots, Y_n$ unabhängige $\{0,1\}$-wertige Beobachtungen mit $Y_i \sim \mathrm{Bin}(1, p_i(\vartheta))$ für $\vartheta \in \Theta$ unbekannt und bekanntes $p = (p_1, \dots, p_n) : \Theta \to (0,1)^n$. Geben Sie das zugehörige statistische Modell an und zeigen Sie, dass die Verteilungen $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$, der Beobachtungen eine Exponentialfamilie bilden.

(b) Nun sei $\Theta \subseteq \mathbb{R}^k$. Für welche Funktionen p ist ϑ natürlicher Parameter der Exponentialfamilie? Zeigen Sie, dass dies im Fall $\Theta = \mathbb{R}^k$, $k \leqslant n$, genau auf das Modell der logistischen Regression mit $\operatorname{logit}(p_i(\vartheta)) = (X\vartheta)_i$, $i = 1, \ldots, n$, und $X \in \mathbb{R}^{k\times n}$ führt.

(c) Beschreiben Sie, wie logistische Regression für die Klassifikation (bei zwei Klassen) verwendet wird. Welcher Klassifizierer ergibt sich unter symmetrischem Missklassifikationsfehler?

(d) Nennen Sie zwei weitere Klassifikationsverfahren und geben Sie kurz ihre Konstruktionsidee an.

4. Gegeben sei das lineare Modell $Y = X\beta+\varepsilon$ mit deterministischer Designmatrix $X \in \mathbb{R}^{n\times k}$, die orthogonal ($X^\top X = E_k$) sei, $\beta \in \mathbb{R}^k$ unbekannt sowie $\varepsilon \sim N(0, \sigma^2 E_n)$, $\sigma > 0$ bekannt. Zeigen Sie:

(a) Der Kleinste-Quadrate-Schätzer ist $\hat{\beta}_{KQ} = X^\top Y$ und dies ist auch der Maximum-Likelihood-Schätzer.

(b) Im Bayesschen Sinne sei β gemäß einer Lebesguedichte $f^\beta(b) = \prod_{i=1}^k \varphi(b_i)$, $b \in \mathbb{R}^k$, a-priori-verteilt. Dann ist

$$f^{\beta|Y}(b) = C \exp\Big(\sum_{j=1}^k \Big(\frac{-b_j^2 + 2(X^\top Y)_j b_j}{2\sigma^2} + \log(\varphi(b_j))\Big)\Big), \quad b \in \mathbb{R}^k,$$

mit einer Normierungskonstanten $C > 0$ (unabhängig von b) die a-posteriori-Lebesguedichte von β gegeben Y.
Hinweis: Geben Sie zunächst die Bayesformel für Dichten an.

(c) Die MAP-Schätzer $\hat{\beta}_\varphi = \operatorname{argmax}_{b\in\mathbb{R}^k} f^{\beta|Y}(b)$ in (b) entsprechen für $\varphi(b) = (2\pi\tau^2)^{-1/2} e^{-b^2/(2\tau^2)}$ einem Ridge-Regression-Schätzer und für $\varphi(b) = (2\tau)^{-1} e^{-|b|/\tau}$ einem LASSO-Schätzer ($\tau > 0$ jeweils fest).

(d) Erklären Sie allgemein, in welchen Fällen Ridge-Regression- und LASSO-Schätzer Vorteile gegenüber $\hat{\beta}_{KQ}$ haben.