Institut für angewandte Mathematik
Sommersemester 2022
Andreas Eberle, Luis La Rocca

UNIVERSITÄT BONN IAM

2. Klausur „Einführung in die Statistik"

Musterlösung zur 2. Klausur

---

1. (Einige kurze Fragen) [*30 Punkte*]

In der Lösung dieser Aufgabe brauchen Sie (ausnahmsweise) *keine Begründungen anzugeben!*

a) Seien $x = (x_1, \dots, x_n)^T$ und $y = (y_1, \dots, y_n)^T$ Stichproben mit $n \in \mathbb{N}$ und $x_i, y_i \in \mathbb{R}$ für $i = 1, \dots, n$. Was ist in den folgenden Grafiken dargestellt? (mit Skizze)

(i) Boxplot von x

(ii) Streudiagramm von x und y

b) Welche der folgenden Kenngrößen $K(x_1, \dots, x_n)$ sind Lage- bzw. Skalenparameter?

(i) Stichprobenmittelwert

(ii) Schiefe

(iii) Stichprobenmedian

(iv) Spannweite

(v) Interquartilsabstand

(vi) Minimum von $x_1, \dots, x_n$.

c) Wie ist die empirische Verteilungsfunktion von $x_1, \dots, x_n$ definiert?

d) Seien $Z_1, Z_2, Z_3, \dots$ unabhängige standardnormalverteilte Zufallsvariablen. Konstruieren Sie daraus Zufallsvariablen mit den folgenden Verteilungen:

(i) Normalverteilung mit Mittelwert $m \in \mathbb{R}$ und Varianz $v > 0$

(ii) Chiquadratverteilung mit n Freiheitsgraden

(iii) t-Verteilung mit n Freiheitsgraden

(iv) Fisher-Verteilung $F(m, n)$

(v) Multivariate Normalverteilung mit Mittelwert 0 und Kovarianzmatrix $C = \sigma\sigma^T$, wobei σ eine $n \times n$-Matrix ist.

e) Was besagt die Informationsungleichung von Cramér-Rao?

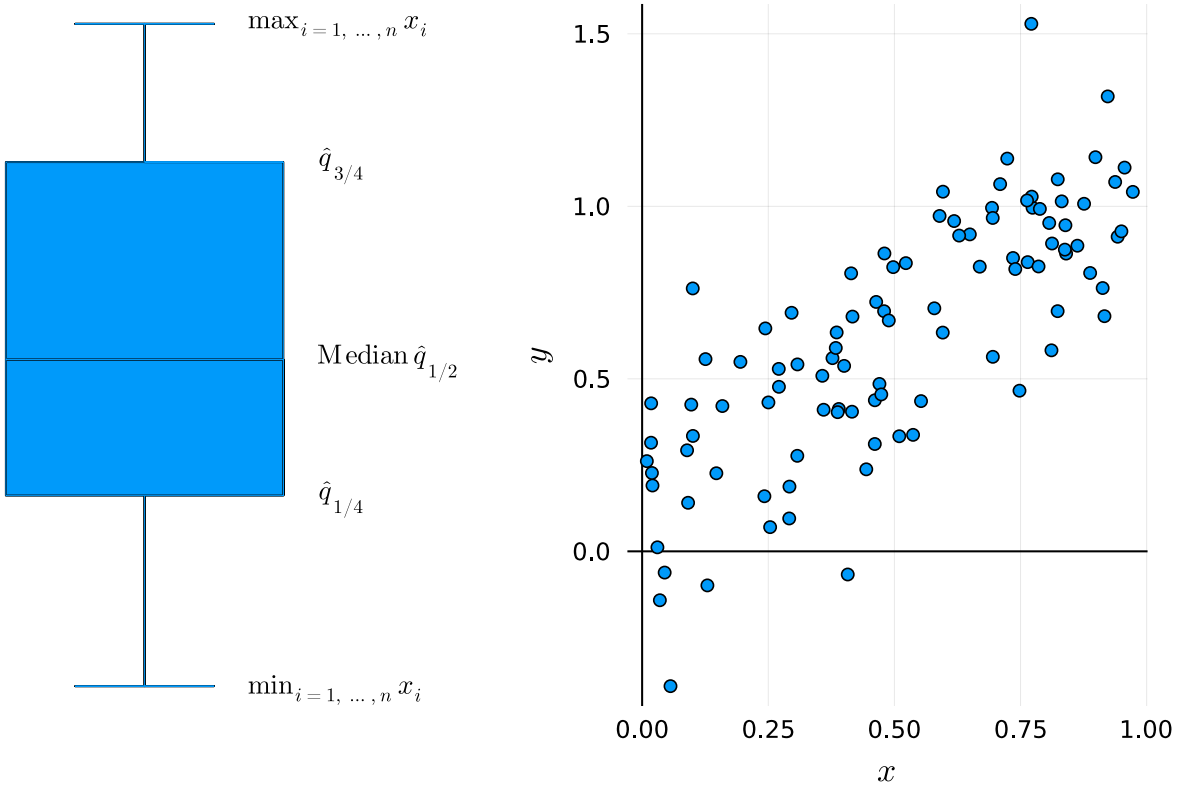
f) Seien μ und ν absolutstetige Wahrscheinlichkeitsverteilungen auf $\mathbb{R}^d$ mit strikt positiven Dichten $f(x)$ bzw. $g(x)$. Wie ist die relative Entropie (Kullback-Leibler-Divergenz) von μ bezüglich ν definiert?

g) Sei x eine Stichprobe von einer der beiden Verteilungen μ bzw. ν. Geben Sie einen mächtigsten Test für das folgende Testproblem an:

$$H_0 : \; x \sim \mu \,, \qquad H_1 : \; x \sim \nu \,.$$

Lösung:

a)



(i) Boxplot mit den Größen $\min\limits_{i=1,\dots,n} x_i, \hat{q}_{1/4}$ dem Median, $\hat{q}_{3/4}$ und $\max\limits_{i=1,\dots,n} x_i$

(ii) Streudiagramm von x und y mit den Punkten $(x_i, y_i) \in \mathbb{R}^2$ für $i = 1, \dots, n$

b)
- Lageparameter sind der Mittelwert, der Median und das Minimum von $x_1, \dots, x_n$.
- Skalenparameter sind die Spannweite und der Interquartilsabstand.
- Weder Lage- noch Skalenparameter ist die Schiefe.

c) Die empirische Verteilungsfunktion $F : \mathbb{R} \to [0, 1]$ ist definiert durch

$$F(c) = \frac{\#\{i \leq n \colon x_i \leq c\}}{n}.$$

d)

$$\text{(i)}\ m+\sqrt{v}Z_1 \sim \mathcal{N}(m,v) \qquad \text{(ii)}\ \sum_{i=1}^{n} Z_1^2 \sim \chi^2(n) \qquad \text{(iii)}\ \frac{Z_{n+1}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}} \sim t_n$$

$$\text{(iv)}\ \frac{\frac{1}{m}\sum_{i=1}^{m} Z_i^2}{\frac{1}{n}\sum_{i=m+1}^{m+n} Z_i^2} \sim F(m,n) \qquad \text{(v)}\ \sigma \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim \mathcal{N}(0,C)$$

e) Betrachte ein parametrisches Modell mit Dichte $f_\vartheta(x)$ für $\vartheta \in \Theta$. Ist $\hat{\vartheta} = T(X)$ ein erwartungstreuer und regulärer Schätzer für ϑ (d.h. $\frac{\mathrm{d}}{\mathrm{d}\vartheta}\int T(x) f_\vartheta(x)\mathrm{d}x = \int T(x)\frac{\mathrm{d}}{\mathrm{d}\vartheta} f_\vartheta(x)\mathrm{d}x$) dann gilt $\mathrm{Var}_\vartheta(\hat{\vartheta}) \geq \frac{1}{I(\vartheta)}$, wobei

$$I(\vartheta) = \int \left(\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log f_\vartheta(x)\right)^2 f_\vartheta(x)\mathrm{d}x$$

die Fischer-Information ist.

f) Die relative Entropie (Kullback-Leibler-Divergenz) von μ bezüglich ν ist definiert als

$$H(\mu|\nu) = \int \rho \log \rho \,\mathrm{d}\nu = \int \frac{f}{g}\log\left(\frac{f}{g}\right) g \,\mathrm{d}x$$

wobei $\rho = \frac{f}{g}$ die relative Dichte ist.

g) Ein mächtigster Test für das Testproblem ist der Likelihood-Quotiententest: Verwerfe H_0 falls

$$\frac{L(\nu;x)}{L(\mu;x)} = \frac{g(x)}{f(x)} > c$$

wobei der Schwellenwert $c \in \mathbb{R}$ so gewählt wird, dass

$$\mathbb{P}_0[\text{verwerfe}] = \int \mathbb{1}_{\frac{g(x)}{f(x)}>c}\, f(x)\,\mathrm{d}x = \alpha.$$

2. (Hypothesentests) *[25 Punkte]*

Formulieren Sie in den folgenden Situationen jeweils ein statistisches Modell und ein Testproblem. Geben Sie die Entscheidungsregeln für Hypothesentests an, und berechnen Sie die p-Werte (falls nötig approximativ). Welche Entscheidung liefern die Tests zum Signifikanzniveau 5%?

a) Mit Hilfe eines Zweifach-Wahlapparats soll festgestellt werden, ob ein Käfer in der Lage ist, eine an einem von zwei Ausgängen angebrachte chemische Substanz zu orten. Bei 10 Versuchen nimmt der Käfer 2 mal den ersten Ausgang, und 8 mal den zweiten Ausgang.

b) Ein Spieler wirft bei 180 Würfen eines Würfels 40 mal eine ,,Sechs".

c) In einem Praktikumsversuch ergeben sich bei 10 Messungen einer reellwertigen Größe die folgenden Abweichungen vom theoretisch prognostizierten Wert:

$$4 \quad 8 \quad -1 \quad 1 \quad 0 \quad 1 \quad -1 \quad 6 \quad 3 \quad -1$$

Sie gehen davon aus, dass die Messwerte normalverteilt sind.

Lösung:

a) Sei $X \sim \mathrm{Bin}(n,p)$ die Häufigkeit, dass der Käfer Ausgang 1 wählt, mit unbekanntem Parameter $p \in [0,1]$ und $n = 10$. Betrachte das Testproblem

$$H_0 \colon p = \frac{1}{2} \quad H_1 \colon p < \frac{1}{2}.$$

Betrachte die Teststatistik X und berechne den p-Wert

$$\begin{aligned}\mathbb{P}_0[X \leq 2] &= \mathrm{Bin}\left(10, \frac{1}{2}\right)[\{0,1,2\}] \\ &= 2^{-10}\left(\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right) = \frac{1 + 10 + \frac{90}{2}}{1024} \\ &= \frac{56}{1024} > 5\%\end{aligned}$$

H_0 kann zum Niveau von 5% nicht verworfen werden.

b) Sei $X \sim \mathrm{Bin}(n,p)$ die Häufigkeit, dass der Spieler eine 6 würfelt, mit unbekanntem Parameter $p \in [0,1]$ und $n = 180$. Betrachte das Testproblem

$$H_0 \colon p = \frac{1}{6} \quad H_1 \colon p > \frac{1}{6}.$$

Betrachte die Teststatistik X und berechne den p-Wert

$$\begin{aligned}\mathbb{P}_0[X \geq 40] &= \mathbb{P}_0\left[\frac{X - \mathbb{E}[X]}{\sigma[X]} \geq \frac{40 - \mathbb{E}[X]}{\sigma[X]}\right] \\ &\overset{\text{ZGS}}{\approx} 1 - \Phi\left(\frac{40 - \mathbb{E}[X]}{\sigma[X]}\right)\end{aligned}$$

Unter H_0 ist $\mathbb{E}[X] = np = 30, \mathrm{Var}[X] = np(1-p) = 30 \cdot \frac{5}{6} = 25$ und $\sigma[X] = \sqrt{25} = 5$. Also ist

$$\frac{40 - \mathbb{E}[X]}{\sigma[X]} = \frac{40 - 30}{5} = 2 \quad \text{und} \quad \Phi(2) \approx 0,977.$$

Damit ist

$$\mathbb{P}_0[X \geq 40] \approx 2,3\%.$$

H_0 kann zum Niveau von 5% verworfen werden.

c) Seien $X_i \sim \mathcal{N}(m, v)$ die unabhängigen Messwerte mit unbekannten Parametern m, v für $i = 1, \ldots, 10$. Betrachte das Testproblem

$$H_0 \colon m = 0 \quad H_1 \colon m > 0$$

und führe einen t-Test durch. Betrachte dazu die Teststatistik

$$T(X) = \frac{\bar{X}}{\sqrt{n/V_\star}} \sim t(\underbrace{n-1}_{=9}) \quad \text{mit } n = 10.$$

Verwerfe H_0, falls $T(X) > q_{0,95;t(9)}$. Hier ist $\bar{X} = 2$, $\sum_{i=1}^{10}(X_i - \bar{X})^2 = 36+16+27+8+3 = 90$, $V_\star = \frac{90}{9} = 10$, $n/V_\star = \frac{10}{10} = 1$ und somit $T(X) = \bar{X} = 2$. Berechne

$$\mathbb{P}_0[T(X) \geq 2] = 1 - \underbrace{F_{t(9)}(2)}_{\approx 0,96} \approx 0,04 < 5\%$$

und verwerfe H_0 zum Niveau von 5%.

3. (Robuste Schätzung des Bereichs von Zufallszahlen) *[25 Punkte]*

Ein Zufallszahlengenerator erzeugt n Zufallszahlen aus einem Interval $(0, \theta)$, wobei $\theta \in (0, \infty)$ ein unbekannter Parameter ist.

a) Formulieren Sie ein statistisches Modell.

b) Sei $\hat{q}_\alpha$ das α-Stichprobenquantil von $x_1, \ldots, x_n$. Welche der folgenden Statistiken liefern erwartungstreue Schätzer für θ? (ohne Beweis)

(i) $T_1(x) = 2\overline{x}$

(ii) $T_2(x) = 2\hat{q}_{1/2}$

(iii) $T_3(x) = x_{(n)} - x_{(1)}$

(iv) $T_4(x) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$ mit $k = \lfloor \tau n \rfloor$ für einen festen Wert $\tau \in (0, 1/2)$.

c) Welche dieser Schätzer sind robust? Begründen Sie kurz, und geben Sie ohne Beweis die asymptotischen Bruchpunkte der Schätzer an.

d) Bestimmen Sie für $n = 8$ ein (natürlich möglichst kleines) Konfidenzintervall für θ zum Konfidenzniveau $\alpha = 0,9$, dass durch einen einzelnen Ausreißer nicht beliebig stark verändert werden kann.

Lösung:

a) Seien $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$ unabhängige Zufallsvariablen mit dem unbekannten Parameter $\theta \in \mathbb{R}_+$

b) (i) T_1 ist erwartungstreu.

(ii) T_2 ist erwartungstreu.

(iii) T_3 ist nicht erwartungstreu

(iv) T_4 ist nicht erwartungstreu.

c) (i) T_1 ist nicht robust: ein einzelner Wert kann $\overline{x}$ beliebig stark verändern. Bruchpunkt = 0.

(ii) T_2 ist robust: der Median verschiebt sich bei Austauschen von $k < \frac{n}{2}$ Werten nur zu einem der anderen Werte. Bruchpunkt = $\frac{1}{2}$.

(iii) T_3 ist nicht robust: Austausch eines Werts kann $x_{(n)}$ beliebig stark verändern. Bruchpunkt = 0.

(iv) T_4 ist robust: Man muss mindestens den Anteil τ der Werte austauschen um T_4 zu verändern. Bruchpunkt = τ.

d) Berechne das Ordnungsintervall

$$\begin{aligned}\mathbb{P}_\theta\left[\frac{\theta}{2} < X_{(2)}\right] &= \mathbb{P}_\theta\left[X_i \leq \frac{\theta}{2} \text{ für höchstens ein } i \in \{1,\dots,8\}\right] \\ &= \mathrm{Bin}\left(8, \frac{1}{2}\right)(\{0,1\}) = 2^{-8}\left(\binom{8}{0} + \binom{8}{1}\right) = \frac{1+8}{256}.\end{aligned}$$

Analog ist

$$\mathbb{P}_\theta\left[\frac{\theta}{2} > X_{(7)}\right] = \mathbb{P}_\theta\left[X_i \geq \frac{\theta}{2} \text{ für höchstens ein } i \in \{1,\dots,8\}\right] = \frac{9}{256}.$$

Also ist

$$\mathbb{P}_\theta\left[\frac{\theta}{2} \notin \left[X_{(2)}, X_{(7)}\right]\right] = \frac{18}{256} < 10\%$$

das heißt $[2X_{(2)}, 2X_{(7)}]$ ist ein Konfidenzintervall zum Niveau 90%, dass durch einen einzelnen Ausreißer nicht beliebig stark verändert werden kann.

4. (Lineare Regression) *[35 Punkte]*

Seien $x_1, \ldots, x_n \in \mathbb{R}$ feste Werte und $Y_1, \ldots, Y_n$ reelle Beobachtungswerte. Sie vermuten einen linearen Zusammenhang

$$Y_i \;=\; a + bx_i + \xi_i \qquad \text{mit } a, b \in \mathbb{R} \text{ und } \xi_i \sim \mathcal{N}(0,1) \text{ unkorreliert.}$$

a) Wie könnte man überprüfen, ob die Normalverteilungsannahme gerechtfertigt ist?

b) Zeigen Sie, dass der Maximum-Likelihood-Schätzer $\hat{\theta} = (\hat{a}, \hat{b})$ für $\theta = (a, b)$ mit dem Kleinste-Quadrate-Schätzer übereinstimmt.

c) Zeigen Sie, dass mit $\tilde{x}_i = x_i - \overline{x}$ und $\tilde{Y}_i = Y_i - \overline{Y}$ gilt

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - a - bx_i)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - b\tilde{x}_i)^2 + \left(\overline{Y} - b\overline{x} - a\right)^2 .$$

d) Folgern Sie, dass der Maximum-Likelihood-Schätzer gegeben ist durch

$$\hat{b} \;=\; \frac{\sum_{i=1}^{n}\tilde{x}_i\tilde{Y}_i}{\sum_{i=1}^{n}\tilde{x}_i^2}, \qquad \hat{a} \;=\; \overline{Y} - b\overline{x}\,.$$

e) Zeigen Sie, dass $\hat{b}$ ein erwartungstreuer Schätzer für die Steigung b der Regressionsgeraden mit Varianz $v = 1\,/\sum_{i=1}^{n}(x_i - \overline{x})^2$ ist.

f) Welche Verteilung hat $\hat{b}$?

g) Bestimmen Sie ein rechtsseitiges Konfidenzintervall für b zum Konfidenzniveau 98%.

h) Welche Realisierung des Konfidenzintervalls ergibt sich für $n = 10$ bei den Beobachtungswerten $\hat{b} = 2$ und $s_X = 1$? Was können Sie daraus über den Zusammenhang der zugrundeliegenden Merkmale X und Y schließen ?

Lösung:

a) Zur Überprüfung der Annahme, dass die Residuen normalverteilt sind kann entweder ein Kolmogorov-Smirnov-Test durchgeführt werden oder ein Normal-QQ-Plot angelegt werden. Im Normal-QQ-Plot werden die empirischen Quantile gegen die theoretischen Quantile der Standardnormalverteilung aufgetragen. Liegen die Punkte näherungsweise auf einer Geraden, dann kann von einer Normalverteilung ausgegangen werden.

b) Seien $Y_i \sim \mathcal{N}(a+bx_i, 1)$ unabhängig für $i = 1, \ldots, n$. Dann ist die Likelihoodfunktion gegeben durch

$$L(a, b; y) = (2\pi)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2}\sum\limits_{i=1}^{n}(y_i - a - bx_i)^2}.$$

Damit ist der Maximum-Likelihood Schätzer das globale Minimum $(a, b) \in \mathbb{R}^2$ der Funktion

$$\sum_{i=1}^{n}(Y_i - a - bx_i)^2 = \|Y - a\mathbf{1} - bx\|^2$$

Dies entspricht somit dem Kleinste Quadrate Schätzer (LSE).

c) Es gilt

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - a - bx_i)^2 &= \sum_{i=1}^{n}\left(\tilde{Y}_i - b\tilde{x}_i + \overline{Y} - a - b\overline{x}\right)^2 \\ &= \sum_{i=1}^{n}\left(\tilde{Y}_i - b\tilde{x}_i\right)^2 + n\left(\overline{Y} - a - b\overline{x}\right)^2 + 2\left(\overline{Y} - a - b\overline{x}\right)\underbrace{\sum_{i=1}^{n}\left(\tilde{Y}_i - b\tilde{x}_i\right)}_{=0 \text{ da zentriert}}\end{aligned}$$

d) Für jeden festen Wert von b ist $n\left(\overline{Y} - a - b\overline{x}\right)^2$ minimal für $a = \overline{Y} - b\overline{x}$. Außerdem gilt

$$\begin{aligned}\sum_{i=1}^{n}\left(\tilde{Y}_i - b\tilde{x}_i\right)^2 &= \|\tilde{Y} - b\tilde{x}\|^2 \\ &= \|\tilde{Y}\|^2 - 2b\tilde{x}\cdot\tilde{Y} + b^2\|\tilde{x}\|^2\end{aligned}$$

Durch quadratische Ergänzung oder Ableiten nach b findet man das Minimum

$$b = \frac{\tilde{x}\cdot\tilde{Y}}{\|\tilde{x}\|^2}$$

Also ist der Maximum Likelihood Schätzer für $\theta = (a, b)$ gegeben durch

$$\hat{b} = \frac{\tilde{x}\cdot\tilde{Y}}{\|\tilde{x}\|^2} \quad \text{und} \quad \hat{a} = \overline{Y} - \hat{b}\overline{x}.$$

e) Da $Y_i = a + bx_i + \xi_i$ und $\overline{Y} = a + b\overline{x} + \overline{\xi}$ ist, gilt

$$\tilde{Y}_i = Y_i - \overline{Y} = b(x_i - \overline{x}) + \xi_i - \overline{\xi} = b\tilde{x}_i + \xi_i - \overline{\xi}$$

und folglich

$$\hat{b} = \frac{\tilde{x}\cdot\tilde{Y}}{\|\tilde{x}\|^2} = b + \frac{\tilde{x}\cdot(\xi - \overline{\xi}\mathbf{1})}{\|\tilde{x}\|^2} = b + \frac{\tilde{x}\cdot\xi}{\|\tilde{x}\|^2}$$

wobei im letzten Schritt benutzt wurde, dass

$$\tilde{x}\cdot\mathbf{1} = \sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Damit ergibt sich der Erwartungswert des Schätzers $\hat{b}$ als

$$\mathbb{E}\left[\hat{b}\right] = \mathbb{E}\left[b + \frac{\tilde{x}\cdot\xi}{\|\tilde{x}\|^2}\right] = b + \frac{\tilde{x}\cdot\mathbb{E}\left[\xi\right]}{\|\tilde{x}\|^2} = b,$$

da die Residuen zentriert sind. Als Varianz des Schätzers $\hat{b}$ erhalten wir

$$\mathrm{Var}\left[\hat{b}\right] = \frac{\mathrm{Var}\left[\tilde{x}\cdot\xi\right]}{\|\tilde{x}\|^4} = \frac{1}{\|\tilde{x}\|^4}\mathrm{Var}\left[\sum_{i=1}^{n}\tilde{x}_i\xi_i\right] = \frac{1}{\|\tilde{x}\|^4}\sum_{i,j=1}^{n}\tilde{x}_i\tilde{x}_j\mathrm{Cov}\left[\xi_i,\xi_j\right] = \frac{1}{\|\tilde{x}\|^2}$$

da die Residuen unkorreliert sind und Varianz 1 haben.

f) Da die Residuen ξ_i für $i = 1,\ldots,n$ standard normalverteilt und unkorreliert sind, ist der Vektor $\xi = (\xi_1,\ldots,\xi_n) \sim \mathcal{N}(0,\mathbb{1})$ ein n-dimensionaler, standord normalverteilter Zufallsvektor. Durch die lineare Transformation

$$\hat{b} = b + \frac{\tilde{x}\cdot\xi}{\|\tilde{x}\|^2}$$

bleibt die Normalverteilung erhalten und es gilt $\hat{b} \sim \mathcal{N}\left(b, \frac{1}{\|\tilde{x}\|^2}\right)$.

g) Nach f) ist $(\hat{b}-b)\|\tilde{x}\|$ standardnormalverteilt. Damit gilt

$$\mathbb{P}\left[b > \hat{b} - c\|\tilde{x}\|^{-1}\right] = \mathbb{P}\left[(\hat{b}-b)\|\tilde{x}\| < c\right] = \Phi(c) > 0,98$$

für $c = 2.1$, siehe Tabelle der Verteilungsfunktion der Standardnormalverteilung. Also ist

$$\left[\hat{b} - 2.1\|\tilde{x}\|^{-1}, \infty\right)$$

ein rechtsseitiges Konfidenzintervall für b zum Konfidenzniveau 98%.

h) In diesem Fall ist $\|\tilde{x}\|^2 = (n-1)s_X^2 = 9$, also $\|\tilde{x}\| = 3$. Damit ergibt ich das Konfidenzintervall $[1.3,\infty)$ für die Steigung der Regressionsgeraden. Wir können also mit hoher Signifikanz von einer positiven Korrelation ausgehen.