Übung 1: Statistische Grundlagen

Aufgabe 1

Berechnen Sie die Schiefe und Kurtosis folgender Zufallsvariablen:

a) X besitzt die Dichte

$$f_X(x) = \begin{cases} 2x & \text{für } 0 \leq x \leq 1 \\ 0 & \text{sonst} \end{cases} .$$

b) $X \sim \text{Bin}(4; 0{,}3)$, d.h. X besitzt die Wahrscheinlichkeitsfunktion

$$f_X(x) = \begin{cases} \binom{4}{x} 0{,}3^x 0{,}7^{4-x} & \text{für } x \in \{0, 1, 2, 3, 4\} \\ 0 & \text{sonst} \end{cases} .$$

Aufgabe 2

Eine faire Münze wird 1000 mal geworfen. Approximieren Sie mit Hilfe des zentralen Grenzwertsatzes die Wahrscheinlichkeit, dass dabei häufiger als 530 mal Kopf fällt.

Aufgabe 3

Der dreidimensionale Zufallsvektor $\mathbf{X} = (X_1, X_2, X_3)^T$ sei $\text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-verteilt mit

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} \qquad \text{und} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 3 \end{pmatrix} .$$

a) Geben Sie die Verteilung von X_1 an.

b) Geben Sie die Verteilung des zweidimensionalen Zufallsvektors $(X_1, X_3)^T$ an.

c) Ermitteln Sie die Verteilung des zweidimensionalen Zufallsvektors $(X_1, X_2)^T$, bedingt gegeben $X_3 = 3$.

Übung 2: Klassisches lineares Modell 1

Aufgabe 1

Es soll der Zusammenhang zwischen einer abhängigen Variablen y und zwei erklärenden Variablen x_1 und x_2 untersucht werden. Bei x_1 handelt es sich dabei um eine stetige Variable und bei x_2 um eine kategoriale Variable, welche die drei Ausprägungen $1, 2$ und 3 annehmen kann. Für die Variablen y, x_1 und x_2 liegen folgende Beobachtungen vor:

y	x_1	x_2
5	4	1
3	3	1
9	5	2
10	6	2
10	2	3
15	5	3

a) Stellen Sie ein geeignetes Regressionsmodell auf und wählen Sie dabei $x_2 = 1$ als Referenzlevel der kategorialen Variablen x_2.

b) Berechnen Sie den KQ-Schätzer für $\boldsymbol{\beta}$ für das in a) aufgestellte lineare Regressionsmodell und verwenden Sie dafür:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{22}\begin{pmatrix} 60 & -14 & 17 & -11 \\ -14 & 4 & -8 & 0 \\ 17 & -8 & 38 & 11 \\ -11 & 0 & 11 & 22 \end{pmatrix}$$

c) Testen Sie mittels t-Test die Signifikanz der erklärenden Variablen x_2 bei einem Signifikanzniveau von $\alpha = 0{,}01$.

Aufgabe 2

Ein Mitarbeiter in der Marketingabteilung eines Finanzdienstleisters vermutet, dass es einen positiven Zusammenhang zwischen den Werbeausgaben (in Millionen Euro) und dem Absatz (in Millionen Euro) des beworbenen Finanzproduktes gibt. In den letzten drei Jahren wurden folgende Zahlen beobachtet:

Absatz y	Werbeausgaben x
5	2
8	3
9	4

Es wird im Folgenden davon ausgegangen, dass ein linearer Zusammenhang zwischen dem Absatz und den Werbeausgaben vorliegt.

a) Bestimmen Sie die Designmatrix $\mathbf{X}$ und den KQ-Schätzer $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)^T$.

b) Bestimmen Sie eine Schätzung für die Varianz-Kovarianzmatrix des KQ-Schätzers $\widehat{\boldsymbol{\beta}}$.

c) Ermitteln Sie den Wert des Bestimmtheitsmaßes R^2 und des adjustierten Bestimmtheitsmaßes R^2_a.

d) Testen Sie die Signifikanz der erklärenden Variablen x bei einem Signifikanzniveau von $\alpha = 0{,}05$ und $\alpha = 0{,}1$.

e) Ermitteln Sie für den Regressionskoeffizienten β_1 das 95%-Konfidenzintervall.

Übung 3: Klassisches lineares Modell 2

Aufgabe 1

Es wird ein lineares Modell mit sechs zu schätzenden Parametern $\beta_0, \ldots, \beta_5$ betrachtet.

a) Ermitteln Sie den Wert der F-Statistik für die Testsituation

$$H_0 : (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T = \mathbf{0} \quad \text{gegen} \quad H_1 : (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T \neq \mathbf{0}$$

und den Fall, dass $n = 40$ Beobachtungen vorliegen und das Bestimmtheitsmaß $R^2 = 0{,}2$ beträgt. Beurteilen Sie anschließend, ob die Nullhypothese H_0 bei einem Signifikanzniveau von $\alpha = 0{,}05$ zu verwerfen ist.

b) Es liegen nun $n = 400$ Beobachtungen vor. Beurteilen Sie, ob die Nullhypothese H_0 jetzt zu verwerfen ist.

c) Der KQ-Schätzer für $\boldsymbol{\beta}$ sei nun durch $\widehat{\boldsymbol{\beta}} = (2, 2, 3, 3, 4, 1)^T$ gegeben. Berechnen Sie das 95%-Konfidenzintervall für die Linearkombination $\mathbf{c}^T\widehat{\boldsymbol{\beta}}$ mit $\mathbf{c}^T = (1, 2, 1, 4, 5, 3)$ und $\widehat{\sigma}\left(\mathbf{c}^T\widehat{\boldsymbol{\beta}}\right) = 0{,}9$ für den Fall, dass $n = 40$ Beobachtungen vorliegen.

d) Der KQ-Schätzer für $\boldsymbol{\beta}$ sei wieder durch $\widehat{\boldsymbol{\beta}} = (2, 2, 3, 3, 4, 1)^T$ gegeben und $n = 40$. Ferner gelte für den Vektor mit den Werten der erklärenden Variablen $\mathbf{x}_* = (1, 1, 3, 2, 2, 1)^T$ und $\widehat{\sigma}\left(\mathbf{x}_*^T\widehat{\boldsymbol{\beta}}\right) = 0{,}3$. Als Schätzung für den Varianzparameter (mittlerer quadratischer Fehler) σ^2 liegt der Wert $\widehat{\sigma}^2 = 4$ vor. Berechnen Sie mittels diesen Angaben das $(1-\alpha)$-Prognoseintervall für y zum Signifikanzniveau $\alpha = 0{,}05$.

Aufgabe 2

Für ein lineares Modell der Form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

liegen folgende Beobachtungen vor:

y	x_1	x_2
120	3	10
108	5	7
92	-2	3
61	1	-12
198	-5	21
21	-2	-29

a) Prüfen Sie bei einem Signifikanzniveau von $\alpha = 0{,}025$, ob mindestens eine der beiden unabhängigen Variablen einen signifikanten Einfluss auf die Zielvariable hat.

b) Führen Sie einen beidseitigen Test für die Nullhypothese

$$H_0 : \beta_1 = -1 \quad \wedge \quad \beta_2 = 1 \qquad \text{gegen} \qquad \beta_1 \neq -1 \quad \vee \quad \beta_2 \neq 1$$

durch. Kann die Nullhypothese bei einem Signifikanzniveau von $\alpha = 0{,}025$ abgelehnt werden?

Übung 4: Klassisches lineares Modell 3

Aufgabe 1

Es wird wieder die Situation aus Aufgabe 2 des 2. Übungsblatts betrachtet.

a) Ermitteln Sie eine Schätzung für die Varianz-Kovarianzmatrix der Residuen $\widehat{\boldsymbol{\varepsilon}}$.

b) Berechnen Sie die standardisierten Residuen r_i für $i = 1, 2, 3$ und beurteilen Sie damit und unter Verwendung einer gängigen Daumenregel, ob es sich bei den Beobachtungen y_1, y_2, y_3 um Ausreißer handelt.

c) Beurteilen Sie anhand einer gängigen Daumenregel für die Hebelwerte, ob die Beobachtungen y_1, y_2, y_3 selbst-schätzend sind. Erläutern Sie ferner, ob die Anwendung dieser Daumenregel in diesem Fall sinnvoll ist.

Aufgabe 2

Ein Unternehmen möchte mithilfe eines klassischen linearen Modells untersuchen, ob die Ausgaben für Marketing einen signifikanten Einfluss auf den Umsatz des Unternehmens haben. Für die Daten aus den letzten 6 Jahren ergaben sich folgende Werte:

Umsatz y	Marketingkosten x
10	5
13	7
13	6
5	3
30	12
9	4

a) Bestimmen Sie den KQ-Schätzer $\widehat{\boldsymbol{\beta}}$.

b) Das Unternehmen vermutet, dass sich innerhalb der Daten Ausreißer befinden, die das Ergebnis verzerren könnten. Untersuchen Sie, ob man aufrund der standardisierten Residuen eine der Beobachtungen als Ausreißer betrachten sollte.

c) Untersuchen Sie die Hebelwerte der einzelnen Beobachtungen. Prüfen Sie mit Hilfe der Cook-Distanz und einer bekannten Daumenregel, ob es sich bei der Beobachtung mit dem größten Hebelwert, um eine einflussreiche Beobachtung handelt?

d) Berechnen Sie für die im Aufgabenteil c) betrachtete Beobachtung das studentisierte Residuum und vergleichen Sie das Ergebnis mit dem im Aufgabenteil b) berechneten Wert für das standardisierte Residuum.

Übung 5: Allgemeines lineares Modell

Aufgabe 1

Betrachtet wird das lineare Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

für die Beobachtungen:

y	x_1	x_2
120	4	12
150	6	18
180	5	15
160	8	25

a) Berechnen Sie den KQ-Schätzer für $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ und die Residuen $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \widehat{\varepsilon}_3, \widehat{\varepsilon}_4)^T$. Verwenden Sie dafür:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} \frac{77}{6} & -24 & \frac{43}{6} \\ -24 & 62 & -19 \\ \frac{43}{6} & -19 & \frac{35}{6} \end{pmatrix}$$

b) Berechnen Sie den Aitken-Schätzer für $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, die Residuen $\widehat{\boldsymbol{\varepsilon}} = (\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \widehat{\varepsilon}_3, \widehat{\varepsilon}_4)^T$ und $\widehat{\sigma}^2$ unter der Voraussetzung, dass

$$\boldsymbol{\varepsilon} \sim \mathrm{N}\left(0, \sigma^2 \begin{pmatrix} 0{,}5 & 0 & 0 & 0 \\ 0 & 0{,}25 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}\right)$$

gilt. Verwenden Sie dafür:

$$\left(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}\right)^{-1} = \begin{pmatrix} \frac{201}{38} & -\frac{161}{19} & 2{,}5 \\ -\frac{161}{19} & \frac{1001}{19} & -17 \\ 2{,}5 & -17 & 5{,}5 \end{pmatrix}$$

c) Vergleichen Sie die berechneten Residuen aus den Aufgabenteilen a) und b) miteinander.

Aufgabe 2

Die folgende Tabelle enthält für die zwölf Risikoklassen (i, j) eines Autohaftpflichtportfolios die Gesamtschadenhöhen S_{ij}:

Fahrzeuggewicht i	Alter j			
	21-30 ($j=1$)	31-40 ($j=2$)	41-50 ($j=3$)	51-60 ($j=4$)
leicht ($i=1$)	1859	1872	1430	2028
mittel ($i=2$)	2376	1872	1716	2340
schwer ($i=3$)	2772	2184	2002	2730

a) Stellen Sie ein allgemeines lineares Modell für die logarithmierten Gesamtschadenhöhen $\ln(S_{ij})$ auf und wählen Sie dabei die Risikoklasse $(1, 1)$ als Referenzklasse.

b) Bestimmen Sie die angepassten Werte für die erwarteten Gesamtschadenhöhen $\mathbb{E}[S_{ij}]$ der zwölf Risikoklassen und berücksichtigen Sie hierbei nur Koeffizienten, die auf einem Signifikanzniveau von $\alpha = 0{,}05$ signifikant sind. Verwenden Sie dabei die folgenden Ergebnisse:

	Estimate	Std. Error	t value	p value
(Intercept)	7.59625	0.04248	178.834	$2.06e^{-12}$ ***
Fahrzeuggewicht.mittel	0.14270	0.04248	3.360	0.015239 *
Fahrzeuggewicht.schwer	0.29685	0.04248	6.989	0.000427 ***
Alter .x2	−0.15662	0.04905	−3.193	0.018761 *
Alter .x3	−0.30440	0.04905	−6.206	0.000807 ***
Alter .x4	0.01883	0.04905	0.384	0.714344

Übung 6: Modellwahl und Variablenselektion

Aufgabe 1

Ein Unternehmen möchte ein klassisches lineares Modell für den Umsatz y aufstellen. Dafür wurden der Umsatz der letzten 20 Jahre sowie die Ausprägungen von vier weiteren erklärenden Variablen erhoben, die in der folgenden Tabelle zusammengefasst sind:

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	190	5	17	24	66	8	423	52	46	19	29	15	510	52	54	7	9
2	96	5	8	-13	45	9	405	47	42	57	23	16	525	62	59	63	48
3	255	37	33	5	20	10	185	23	20	42	36	17	380	48	46	9	17
4	382	42	43	17	19	11	70	11	8	43	15	18	30	6	3	16	11
5	270	29	26	4	32	12	123	16	18	19	55	19	210	14	17	-53	37
6	303	42	40	37	41	13	120	7	10	-17	42	20	116	12	13	-35	33
7	303	31	29	-19	13	14	600	41	46	-25	18						

Der Vorstand des Unternehmens zieht einen externen Berater hinzu, der mithilfe der Daten ein geeignetes lineares Regressionsmodell aufstellen soll. Dieser passt ein lineares Regressionsmodell mit Berücksichtigung aller vier erklärenden Variablen an die Daten an sowie alle möglichen Submodelle. Die resultierenden Ergebnisse sind in der nachfolgenden Tabelle angegeben.

a) Welche der erklärenden Variablen würden sich für eine einfache lineare Regression eignen?

b) Welches Modell würde der Experte dem Vorstand empfehlen, wenn er als Auswahlkriterium das adjustierte Bestimmtheitsmaß verwendet? Würde sich auch das gewöhnliche Bestimmtheitsmaß für die Modellauswahl eignen?

c) Wie würde sich die Entscheidung aus Aufgabenteil b) ändern, wenn statt dem adjustierten Bestimmtheitsmaß eines der beiden Informationskriterien AIC und BIC für die Modellauswahl verwendet wird?

d) Führen Sie mit Hilfe der drei Modellauswahlkriterien R_a^2, AIC und BIC jeweils eine Vorwärts-Selektion durch und starten Sie dabei jeweils beim Nullmodell. Für welches Modell würden Sie sich jeweils entscheiden? Stimmen diese Modelle mit den globalen Lösungen aus den Aufgabenteilen b) und c) überein?

e) Führen Sie mit Hilfe der drei Modellauswahlkriterien R_a^2, AIC und BICjeweils eine Rückwärts-Selektion durch und starten Sie dabei mit dem Modell 16 (volles Modell). Für welches Modell würden Sie sich jeweils entscheiden? Stimmen diese Modelle mit den globalen Lösungen aus den Aufgabenteilen b) und c) überein?

f) Vergleichen Sie die p-Werte der Schätzungen $\widehat{\beta}_1$ und $\widehat{\beta}_2$ in den verschiedenen Modellen. Was fällt auf und wodurch werden diese Unregelmäßigkeiten verursacht?

Modell	Intercept		x_1		x_2		x_3		x_4		R^2	R_a^2	AIC	BIC
	$\widehat{\beta}_0$	P-Wert	$\widehat{\beta}_1$	p-Wert	$\widehat{\beta}_2$	p-Wert	$\widehat{\beta}_3$	p-Wert	$\widehat{\beta}_4$	p-Wert				
1	274.80	$4.47 \cdot 10^{-07}$	-	-	-	-	-	-	-	-	0	0	263.7907	265.7822
2	44.2433	0.175	7.9229	$7.63 \cdot 10^{-08}$ ***	-	-	-	-	-	-	0.8069	0.7961	232.903	235.8902
3	10.9296	0.643 *	-	-	9.1305	$1.18 \cdot 10^{-10}$ ***	-	-	-	-	0.9054	0.9001	218.6332	221.6204
4	266.2035	$2.41 \cdot 10^{-06}$ ***	-	-	-	-	0.8597	0.501	-	-	0.0255	−0.02864	265.274	268.2612
5	358.883	0.000312 ***	-	-	-	-	-	-	−2.761	0.258456	0.07033	0.01869	264.3321	267.3193
6	2.318	0.9166	−5.478	0.0571 .	14.944	$8.55 \cdot 10^{-05}$ ***	-	-	-	-	0.924	0.9151	216.2488	220.2317
7	34.8241	0.2486	8.6423	$3.19 \cdot 10^{-08}$ ***	-	-	−1.1514	0.0532 .	-	-	0.846	0.8278	230.3785	234.3615
8	30.1890	0.594	8.0247	$3.18 \cdot 10^{-07}$ ***	-	-	-	-	0.3643	0.760	0.808	0.7854	234.7897	238.7727
9	5.7046	0.7900	-	-	9.6012	$5.23 \cdot 10^{-11}$ ***	−0.8379	0.0374 *	-	-	0.9272	0.9186	215.3902	219.3731
10	26.5287	0.49	-	-	9.0373	$6.37 \cdot 10^{-10}$ ***	-	-	−0.4239	0.60	0.9069	0.896	220.3	224.2829
11	354.809	0.000438 ***	-	-	-	-	1.015	0.424676	−2.961	0.234225	0.1055	0.0003089	265.56	269.5429
12	0.8369	0.968360	−3.8630	0.176062	13.5925	0.000241 ***	−0.6446	0.114080	-	-	0.9353	0.9232	215.0302	220.0088
13	49.3604	0.1531	−7.8339	0.0137 *	17.1417	$3.09 \cdot 10^{-05}$ ***	-	-	−1.3790	0.0867 .	0.9371	0.9253	214.4652	219.4439
14	−4.9162	0.926	9.0080	$1.09 \cdot 10^{-07}$ ***	-	-	−1.2892	0.040 *	1.0008	0.373	0.8536	0.8262	231.3565	236.3352
15	10.3728	0.7713	-	-	9.5667	$3.87 \cdot 10^{-10}$ ***	−0.8252	0.0501 .	−0.1247	0.8676	0.9273	0.9137	217.3544	222.333
16	36.0955	0.323572	−6.1495	0.079963 .	15.6730	0.000272 ***	−0.4268	0.325150	−1.0189	0.240419	0.9412	0.9255	215.1309	221.1053

Modell	Variablen	R_a^2	AIC	BIC
1	-			

Modell	Variablen	R_a^2	AIC	BIC
2	x_1			

Modell	Variablen	R_a^2	AIC	BIC
3	x_2			

Modell	Variablen	R_a^2	AIC	BIC
4	x_3			

Modell	Variablen	R_a^2	AIC	BIC
5	x_4			

Modell	Variablen	R_a^2	AIC	BIC
6	x_1, x_2			

Modell	Variablen	R_a^2	AIC	BIC
7	x_1, x_3			

Modell	Variablen	R_a^2	AIC	BIC
8	x_1, x_4			

Modell	Variablen	R_a^2	AIC	BIC
9	x_2, x_3			

Modell	Variablen	R_a^2	AIC	BIC
10	x_2, x_4			

Modell	Variablen	R_a^2	AIC	BIC
11	x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
12	x_1, x_2, x_3			

Modell	Variablen	R_a^2	AIC	BIC
13	x_1, x_2, x_4			

Modell	Variablen	R_a^2	AIC	BIC
14	x_1, x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
15	x_2, x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
16	x_1, x_2, x_3, x_4			

			Schritte des Auswahlverfahrens				
		Start-Modell	1	2	3	4	End-Modell
Im Modell vorhandene Variablen	R_a^2	-					
	AIC	-					
	BIC	-					

Modell	Variablen	R_a^2	AIC	BIC
1	-			

Modell	Variablen	R_a^2	AIC	BIC
2	x_1			

Modell	Variablen	R_a^2	AIC	BIC
3	x_2			

Modell	Variablen	R_a^2	AIC	BIC
4	x_3			

Modell	Variablen	R_a^2	AIC	BIC
5	x_4			

Modell	Variablen	R_a^2	AIC	BIC
6	x_1, x_2			

Modell	Variablen	R_a^2	AIC	BIC
7	x_1, x_3			

Modell	Variablen	R_a^2	AIC	BIC
8	x_1, x_4			

Modell	Variablen	R_a^2	AIC	BIC
9	x_2, x_3			

Modell	Variablen	R_a^2	AIC	BIC
10	x_2, x_4			

Modell	Variablen	R_a^2	AIC	BIC
11	x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
12	x_1, x_2, x_3			

Modell	Variablen	R_a^2	AIC	BIC
13	x_1, x_2, x_4			

Modell	Variablen	R_a^2	AIC	BIC
14	x_1, x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
15	x_2, x_3, x_4			

Modell	Variablen	R_a^2	AIC	BIC
16	x_1, x_2, x_3, x_4			

			Schritte des Auswahlverfahrens				
		Start-Modell	1	2	3	4	End-Modell
Im Modell vorhandene Variablen	R_a^2	x_1, x_2, x_3, x_4					
	AIC	x_1, x_2, x_3, x_4					
	BIC	x_1, x_2, x_3, x_4					

Übung 7: Quantil-Regression

Aufgabe 1

Bestimmen Sie die Quantilsfunktion einer Par(α, λ)-Verteilung mit der Verteilungsfunktion

$$F_X(x) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha} \quad \text{für } x > 0$$

mit $\alpha, \lambda > 0$.

Aufgabe 2

Bestimmen Sie für die folgenden Zufallsvariablen jeweils den Median als Lösungen des Minimierungsproblems

$$q_{1/2} = \arg\min_{a \in \mathbb{R}} \mathbb{E}[\rho_{1/2}(X - a)].$$

a) Die Zufallsvariable X sei eine auf der Menge $\{1, 2, \ldots, 9\}$ gleichverteilt.

b) Die Zufallsvariable X sei Exp(λ)-verteilt.

Aufgabe 3

Betrachtet wird das lineare Quantil-Regressionsmodel

$$y = \mathbf{x}^T \boldsymbol{\beta}_\tau + \varepsilon_\tau,$$

wobei zusätzlich angeommen wird, dass die abhängige Variable y eine asymmetrische Laplace-Verteilung, d.h. die Dichtefunktion

$$f(y) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right) \qquad \text{mit } \mu \in \mathbb{R},\ \sigma > 0 \text{ und } \tau \in (0,1)$$

besitzt. Weisen Sie nach, dass der QR-Schätzer $\widehat{\boldsymbol{\beta}}_\tau$ unter dieser zusätzlichen Voraussetzung auch der ML-Schätzer für den Vektor mit den Regressionskoeffizienten $\boldsymbol{\beta}_\tau$ ist.

Hinweis: Verwenden Sie dabei, dass links von μ die Wahrscheinlichkeitsmass der asymmetrischen Laplace-Verteilung genau τ und rechts von μ dementsprechend genau $1 - \tau$ beträgt. D.h. dass das τ-Quantil einer asymmetrischen Laplace-Verteilung bei μ liegt.

Übung 8: Verallgemeinerte lineare Modelle 1

Aufgabe 1

Weisen Sie explizit nach, dass die Geometrische-Verteilung Geo(p) mit der Wahrscheinlichkeitsfunktion

$$f(y;p) = \begin{cases} p(1-p)^{y-1} & \text{für } y \in \mathbb{N} \\ 0 & \text{sonst} \end{cases}$$

mit $p \in (0,1)$ zur Exponential-Dispersions-Familie gehört und berechnen Sie den Erwartungswert und die Varianz der Verteilung.

Aufgabe 2

Ein Kreditinstitut sucht ein geeignetes Modell für die Einschätzung der Kreditwürdigkeit von potenziellen Kreditnehmern. Dabei werden die Merkmale „Alter“ (x_1), „Monatsgehalt“ (x_2), „Laufzeit“ (x_3) und „Höhe des Kredites“ (x_4) erhoben und die abhängige Variable y gibt an, ob der potenzielle Kreditnehmer kreditwürdig ist ($y = 1$) oder nicht ($y = 0$). Mit Hilfe eines Datensatzes bestehend aus $n = 300$ Beobachtungen wurden ein klassisches lineares Modell, ein Logit- und ein Probit-Modell angepasst. Man erhielt folgende Ergebnisse:

	Lineares Modell		Logit-Modell		Probit-Modell	
	Estimate	p value	Estimate	p value	Estimate	p value
$\widehat{\beta}_0$	$-1{,}105$	$1{,}81 \cdot 10^{-4}$	$-8{,}437$	$3{,}34 \cdot 10^{-7}$	$-4{,}627$	$9{,}67 \cdot 10^{-7}$
$\widehat{\beta}_1$	$1{,}328 \cdot 10^{-2}$	$2{,}15 \cdot 10^{-3}$	$6{,}936 \cdot 10^{-2}$	$2{,}40 \cdot 10^{-3}$	$3{,}820 \cdot 10^{-2}$	$4{,}27 \cdot 10^{-3}$
$\widehat{\beta}_2$	$2{,}031 \cdot 10^{-4}$	$1{,}19 \cdot 10^{-4}$	$1{,}068 \cdot 10^{-3}$	$1{,}46 \cdot 10^{-4}$	$5{,}683 \cdot 10^{-4}$	$5{,}24 \cdot 10^{-4}$
$\widehat{\beta}_3$	$2{,}790 \cdot 10^{-2}$	$3{,}49 \cdot 10^{-5}$	$1{,}456 \cdot 10^{-1}$	$7{,}09 \cdot 10^{-5}$	$8{,}455 \cdot 10^{-2}$	$6{,}68 \cdot 10^{-5}$
$\widehat{\beta}_4$	$2{,}303 \cdot 10^{-6}$	$2{,}23 \cdot 10^{-2}$	$1{,}159 \cdot 10^{-5}$	$2{,}46 \cdot 10^{-2}$	$6{,}859 \cdot 10^{-6}$	$2{,}25 \cdot 10^{-2}$

a) Beurteilen Sie für einen Neukunden, der 52 Jahre alt ist, ein Monatsgehalt von 4413 € erhält und einen Kredit in Höhe von 103.361 € mit einer Laufzeit von 17 Jahren beantragt hat die Kreditwürdigkeit mit Hilfe der drei angegebenen Modelle. Interpretieren Sie die Ergebnisse und beurteilen Sie, welches der Modelle am wenigsten geeignet ist.

b) Welchen von den folgenden 10 potenziellen Kreditnehmern sollte das Kreditinstitut einen Kredit gewähren, wenn es für seine Einschätzung das Logit- und Probit-Modell verwendet und einer Person ein Kredit gewährt werden soll, wenn deren Ausfallwahrscheinlichkeit kleiner als 50% ist? Geben Sie den Anteil an richtig klassifizierten Kreditnehmern an.

i	x_1	x_2	x_3	x_4	y_i
1	52	4413	17	103361	1
2	45	3649	17	61044	1
3	37	3548	14	85822	1
4	52	4040	20	84167	1
5	38	5134	11	116721	1
6	40	2609	13	65051	0
7	44	4571	13	67189	0
8	23	2333	6	57439	0
9	32	4041	7	74440	0
10	31	4013	10	62341	0

c) Wie würden sich die Anteile der richtig klassifizierten potenziellen Kreditnehmer jeweils ändern, wenn die Kosten für einen nicht zurückgezahlten Kredit im Durchschnitt 6 mal so hoch sind, wie die Kosten eines irrtümlicherweise nicht vergebenen Kredites? Wie verändert sich der Anteil der vergebenen Kredite?

Übung 9: Verallgemeinerte lineare Modelle 2

Aufgabe 1

In einem verallgemeinerten linearen Modell ist die Nullhypothese

$$H_0 : \beta_j = 0 \qquad \text{gegen} \qquad H_1 : \beta_j \neq 0$$

mit Hilfe der Wald-Statistik zu testen. Für die ML-Schätzung von β_j und deren Varianz gilt

$$\widehat{\beta}_j = 2 \qquad \text{bzw.} \qquad \text{Var}(\widehat{\beta}_j) = 1{,}5.$$

Berechnen Sie den p-Wert und geben Sie an, ob die Nullhypothese bei einem Signifikanzniveau von $\alpha = 0{,}05$ zu verwerfen ist.

Aufgabe 2

Gegeben sei ein verallgemeinertes lineares Modell und zu testen sei die allgemeine lineare Hypothese

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \qquad \text{gegen} \qquad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$$

mit $\text{rang}(\mathbf{C}) = 2$. Für die Log-Likelihoodfunktion resultiert unter den beiden Hypothesen der Wert

$$\ln\left(\text{L}\left(\widehat{\boldsymbol{\beta}}_{H_0}; y_1, \ldots, y_n\right)\right) = -1{,}3 \qquad \text{bzw.} \qquad \ln\left(\text{L}\left(\widehat{\boldsymbol{\beta}}_{H_1}; y_1, \ldots, y_n\right)\right) = 2{,}3.$$

Prüfen Sie, ob die Nullhypothese H_0 zum Signifikanzniveau $\alpha = 0{,}05$ bzw. $\alpha = 0{,}01$ zu verwerfen ist.

Aufgabe 3

Bisher wurde in einem Unternehmen ein verallgemeinertes lineares Modell M_1 mit $k = 3$ erklärenden Variablen eingesetzt. Das Unternehmen steht vor der Entscheidung, ob zukünftig ein erweitertes Modell M_2 verwendet werden sollte, bei dem zu den bisherigen drei erklärenden Variablen zwei weitere hinzugefügt wurden. D.h zu testen ist

$$H_0 : M_1 \qquad \text{gegen} \qquad H_1 : M_2$$

Bei Vorliegen von $n = 20$ Beobachtungen betragen die skalierten Devianzen der beiden Modelle

$$\Delta^*_{M_1} = 7{,}2 \qquad \text{bzw.} \qquad \Delta^*_{M_2} = 1{,}1$$

a) Entscheiden Sie anhand dieser Informationen, ob die Nullhypothese $H_0 : M_1$ zum Signifikanzniveau $\alpha = 0{,}05$ abzulehnen ist.

b) Für die Log-Likelihoodfunktion des Modells M_2 hat man $\ln\left(L\left(\widehat{\boldsymbol{\beta}}_{M_2}; y_1, \ldots, y_n\right)\right) = 5{,}4$ erhalten. Berechnen Sie den Wert der Log-Likelihoodfunktion von M_1 und des saturierten Modells.

c) Wir würde die Entscheidung ausfallen, wenn M_1 kein Submodell von M_2 wäre?

d) Wie groß müsste die Anzahl an Beobachtungen n mindestens sein, damit man sich im Aufgabenteil c) für das kleinere Modell entscheidet? Nehmen Sie dabei an, dass sich die Werte der Log-Likelihoodfunktion der beiden Modelle nicht verändern.

Übung 10: Ridge-Regression und LASSO

Aufgabe 1

Die Regressionskoeffizienten $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ des linearen Regressionsmodells

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

sollen durch Lösen des Minimierungsproblems

$$\arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 \quad \text{unter der Nebenbedingung} \quad \sum_{j=1}^{k} |\beta_j| \leq t \tag{1}$$

geschätzt werden (LASSO-Schätzer). Der vorhandene Datensatz wird in einen Trainingsdatensatz vom Umfang n_1 und einen Validierungsdatensatz (Testdatensatz) vom Umfang $n_2 = n - n_1$ zerlegt. Erläutern Sie welche Auswirkung eine wachsender Wert von $t \geq 0$ auf die folgenden Größen hat:

a) Geschätzter erwarteter quadrierter Prognosefehler für die Trainingsdaten, d.h.

$$\frac{1}{n_1} \sum_{i=n_1}^{n} (y_i - \widehat{y}_i)^2$$

b) Geschätzter erwarteter quadrierter Prognosefehler für die Testdaten, d.h.

$$\frac{1}{n_n} \sum_{i=n_1+1}^{n} (y_i - \widehat{y}_i)^2$$

c) Varianz, d.h. $\mathrm{Var}(\widehat{y})$

d) Bias, d.h. $\mathbb{E}[\widehat{y}] - \mathbb{E}[y]$

e) Irreduzibler Prognosefehler, d.h. $\mathrm{Var}(y)$

Aufgabe 2

Eine bekannte Eigenschaft der Ridge-Regression ist, dass Koeffizientenschätzungen für korrelierte erklärende Variablen tendenziell ähnlich sind, während bei LASSO teilweise sehr unterschiedliche Koeffizientenschätzungen resultieren können. Diese Eigenschaft soll im Folgenden anhand des sehr einfachen linearen Regressionsmodells

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

und $n = 2$ Beobachtungen untersucht werden. Dabei gelte weiter:

$$\begin{aligned} x_{11} &= x_{12} \\ x_{21} &= x_{22} \\ y_1 + y_2 &= 0 \\ x_{11} + x_{21} &= 0 \\ x_{12} + x_{22} &= 0 \end{aligned}$$

a) Zeigen Sie, dass mit der KQ-Methode, Ridge-Regression und LASSO für den Intercept β_0 jeweils die Schätzung $\widehat{\beta}_0 = 0$ resultiert.

b) Formulieren Sie das zur Ridge-Regression gehörende Optimierungsproblem.

c) Zeigen Sie, dass im Falle der Ridge-Regression $\widehat{\beta}_1 = \widehat{\beta}_2$ gilt.

d) Formulieren Sie das zu LASSO gehörende Optimierungsproblem.

e) Zeigen Sie, dass im Falle von LASSO die Schätzungen $\widehat{\beta}_1$ und $\widehat{\beta}_2$ nicht eindeutig sind. Charakterisieren Sie die Lösungen.

Hinweis: Verwenden Sie hierzu die folgende alternative Formulierung des zu LASSO gehörenden Optimierungsproblems:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^3} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \text{unter der Nebenbedingung} \quad |\beta_1| + |\beta_2| \leq t$$

Aufgabe 3

Betrachtet wird das lineare Regressionsmodell

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad \text{für } i = 1, \ldots, n$$

mit stochastisch unabhängigen und identisch $\mathrm{N}(0, \sigma^2)$-verteilten zufälligen Fehlern $\varepsilon_1, \ldots, \varepsilon_n$.

a) Bestimmen Sie die Likelihoodfunktion $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ mit $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$.

b) Für die Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_k$ wird zusätzlich angenommen, dass sie eine a priori Verteilung besitzen. Genauer wird angenommen, dass die Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_k$ stochastisch unabhängig und identisch-verteilt sind mit der Dichtefunktion

$$f(\beta) = \frac{1}{2b} \exp(-|\beta|/b) \qquad \text{für } \beta \in \mathbb{R}$$

mit $b > 0$ (d.h. bei der a priori Verteilung der Regressionskoeffizienten handelt es sich um eine Laplace-Verteilung (zweiseitige Exponentialverteilung) mit Erwartungswert 0 und Skalierungsparameter $b > 0$). Geben Sie die Dichtefunktion der a posteriori Verteilung von $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ an.

Hinweis: Die Dichtefunktion der a posteriori Verteilung von $\boldsymbol{\beta}$ ist definiert durch

$$f(\boldsymbol{\beta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})}{\int f(\mathbf{y}|\boldsymbol{\beta}) f(\boldsymbol{\beta})\, d\boldsymbol{\beta}}.$$

c) Weisen Sie nach, dass der LASSO-Schätzer $\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda)$ für ein spezielles $\lambda > 0$ der Modus (Modalwert) der a posteriori Verteilung von $\boldsymbol{\beta}$ ist.

d) Es sei nun angenommen, dass die Regressionskoeffizienten $\beta_0, \beta_1, \ldots, \beta_k$ stochastisch unabhängig und identisch $\mathrm{N}(0, c)$-verteilt sind. Geben Sie die Dichtefunktion der a posteriori Verteilung von $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ an.

e) Weisen Sie nach, dass der Ridge-Schätzer $\widehat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\lambda)$ für ein spezielles $\lambda > 0$ der Modus (Modalwert) und der Erwartungswert der a posteriori Verteilung von $\boldsymbol{\beta}$ ist.

Übung 11: Nichtlineare Regression

Aufgabe 1

a) Geben Sie an, welche der folgenden Regressionsmodelle linear oder nichtlinear sind:

1) $y = \beta_1 + \beta_2 e^{2x} + \varepsilon$
2) $y = \beta_1 + \beta_2 x_1 + \beta_3 \ln(x_2) + \varepsilon$
3) $y = \beta_1 + \beta_2 x_1 + \beta_2 x_2^{\beta_3} + \varepsilon$
4) $y = \beta_1 + \beta_2 e^{x_1} + \beta_3 \frac{1}{x_2} + \varepsilon$
5) $y = \beta_1 e^{-\beta_2 x} + \varepsilon$
6) $y = \beta_1 + \beta_2 e^{\beta_3 x} + \sin(\beta_4 x) + \varepsilon$
7) $y = \frac{\beta_1 x}{x+\beta_2} + \varepsilon$
8) $y = \beta_1 + \frac{\beta_2}{\beta_1} x + \varepsilon$
9) $y = \beta_1 x_1^{\beta_2} x_2^{\beta_3} x_3^{\beta_4} + \varepsilon$
10) $y = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}} + \varepsilon$

b) Geben Sie an, bei welchen der nichtlinearen Regressionsmodelle aus Aufgabenteil a) die Regressionsfunktion $\mathbb{E}[y] = f(\mathbf{x};\boldsymbol{\beta})$ durch eine einfache Transformation und/oder Umparameterisierung linearisiert werden kann. Geben Sie gegebenenfalls die Transformation und/oder Umparameterisierung an.

c) Was ist bei linearisierenden Transformationen zu beachten?

Aufgabe 2

Die untenstehende Tabelle zeigt die Kursentwicklung der Kryptowährung „Statistikcoin“ in den letzten 8 Jahren.

Jahr x	1	2	3	4	5	6	7	8
Kurs y in Euro	20	25	36	48	64	86	114	168

An diese Daten soll das nichtlineare Regressionsmodell

$$y = \beta_1 e^{\beta_2 x} + \varepsilon \tag{1}$$

angepasst werden.

a) Ermitteln Sie analytisch die zu diesem nichtlinearen Regressionsmodell gehörende Funktionalmatrix (Jacobi-Matrix) $\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\beta})$.

b) Geben Sie die Matrix $\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\beta})^T \mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\beta})$ und die asymptotische Varianz-Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}$ an.

c) Stellen Sie die Rekursionsvorschrift des Newton-Raphson-Verfahrens zur Bestimmung des KQ-Schätzers für $\boldsymbol{\beta}$ auf.

d) Bestimmen Sie mit Hilfe von Excel oder `R` und der Rekursionsvorschrift aus Aufgabenteil c) Schätzungen für die beiden Regressionskoeffizienten β_1 und β_2 des nichtlinearen Regressionsmodells (1) (auf vier Nachkommastellen genau). Verwenden Sie dabei die beiden Startwerte $\boldsymbol{\beta}_0 = (15; 0{,}25)$. Geben Sie ferner die Funktionsgleichung der an die Daten angepassten Regressionsfunktion $\widehat{y} = f(x;\widehat{\boldsymbol{\beta}})$ an.

e) Stellen Sie die Daten y und die angepasste Regressionsfunktion $\widehat{y} = f(x;\widehat{\boldsymbol{\beta}})$ in einem Streudiagramm dar.

f) Bestimmen Sie mit Hilfe von Excel oder `R` und den Ergebnissen aus den Aufgabenteilen b) und d) Schätzungen für den Varianzparameter σ^2 (auf fünf Nachkommastellen genau) und für die asymptotische Varianz-Kovarianzmatrix von $\widehat{\boldsymbol{\beta}}$.

g) Ermitteln Sie mit Hilfe des Ergebnisses aus Aufgabenteil f) Schätzungen für die Varianzen der KQ-Schätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$. Beurteilen Sie damit auf Basis einer Daumenregel, ob die beiden KQ-Schätzungen bei einem Signifikanzniveau von $\alpha = 5\%$ statistisch signifikant von 0 verschieden sind.

i	x_i	y_i
1	0,02	76
2	0,02	47
3	0,06	97
4	0,06	107
5	0,11	123
6	0,11	139
7	0,22	159
8	0,22	152
9	0,56	191
10	0,56	201
11	1,10	207
12	1,10	100

Aufgabe 3

An die folgenden Daten soll das nichtlineare Regressionsmodell

$$y = \frac{\beta_1 x}{x + \beta_2} + \varepsilon \tag{2}$$

angepasst werden.

a) Das nichtlineare Regressionsmodell (2) wurde durch eine einfache Transformation und Umparameterisierung linearisiert. Für die beiden Regressionskoeffizienten des resultierenden linearen Regressionsmodells

$$\widetilde{y} = \widetilde{\beta}_1 + \widetilde{\beta}_2 \widetilde{x} + \widetilde{\varepsilon} \tag{3}$$

erhält man die KQ-Schätzungen

$$\widehat{\widetilde{\beta}}_1 = 0{,}005107 \qquad \text{bzw.} \qquad \widehat{\widetilde{\beta}}_2 = 0{,}0002472.$$

Geben Sie die Transformation und Umparameterisierung an, die zur Linearisierung (3) geführt hat und ermitteln Sie damit Schätzungen für die Regressionskoeffizienten des nichtlinearen Regressionsmodells (2).

b) Stellen Sie die transformierten Daten und das daran angepasste lineare Regressionsmodell (3) sowie die nichttransformierten Daten und das daran angepasste nichtlineare Regressionsmodell (2) jeweils in einem Streudiagramm dar. Beurteilen Sie damit anschließend die Güte des angepassten nichtlinearen Regressionsmodells?

c) Ermitteln Sie analytisch die zum nichtlinearen Regressionsmodell (3) gehörende Funktionalmatrix (Jacobi-Matrix) $\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\beta}_0)$ für die Startwerte $\boldsymbol{\beta}_0 = (205; 0{,}08)$. Berechnen Sie damit anschließend die Näherung $\boldsymbol{\beta}_1$.

d) Geben Sie anstelle von $\boldsymbol{\beta}_0 = (205; 0{,}08)$ einen alternativen Satz von sinnvollen Startwerten für das Newton-Raphson-Verfahren an.

e) Die Anpassung des nichtlinearen Regressionsmodells (2) an die Daten mittels `R` liefert die KQ-Schätzungen $\widehat{\beta}_1 = 212{,}7$ und $\widehat{\beta}_2 = 0{,}06412$, d.h. das angepasste Modell lautet:

$$\widehat{y} = \frac{212{,}7x}{x + 0{,}06412} \tag{4}$$

Ferner gilt:

$$\left(\mathbf{J}_{\boldsymbol{\mu}}(212{,}7; 0{,}06412)^T \mathbf{J}_{\boldsymbol{\mu}}(212{,}7; 0{,}06412)\right)^{-1} = \begin{pmatrix} 0{,}4037 & 36{,}82 \times 10^{-5} \\ 36{,}82 \times 10^{-5} & 57{,}36 \times 10^{-8} \end{pmatrix}$$

Stellen Sie die Daten und das angepasste nichtlineare Regressionsmodell (4) in einem Streudiagramm dar und beurteilen Sie die Anpassungsgüte. Bestimmen Sie ferner die Residual-Standardabweichung $\widehat{\sigma}$, die geschätzten Standardfehler für die KQ-Schätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$ sowie eine Schätzung für die Korrelation zwischen $\widehat{\beta}_1$ und $\widehat{\beta}_2$.

f) Beurteilen Sie mit Hilfe der Ergebnisse aus Aufgabenteil e) auf Basis einer Daumenregel, ob die beiden KQ-Schätzungen $\widehat{\beta}_1 = 212{,}7$ und $\widehat{\beta}_2 = 0{,}06412$ bei einem Signifikanzniveau von $\alpha = 5\%$ statistisch signifikant von 0 verschieden sind. Bestimmen Sie ferner für die beiden Regressionskoeffizienten β_1 und β_2 die (asymptotischen) 95%-Konfidenzintervalle.

Übung 12: Nichtparametrische Regression

Aufgabe 1

Von einem linearen Spline-Polynom $s : [a, b] \longrightarrow \mathbb{R}$ zu den Knoten $1, 2$ und 3 sei bekannt, dass $s(0) = 1$, $s(1) = 1{,}3$, $s(2) = 5{,}5$, $s(4) = 6$ und $s(5) = 6$ gilt. Berechnen Sie die folgenden Werte:

a) $s(0{,}5)$

b) $s(3)$

c) $\int_2^4 s(x)\,dx$

Aufgabe 2

Gegeben sei das quadratische Spline-Polynom $s : [a, b] \longrightarrow \mathbb{R}$ mit

$$s(x) = 1 + 0{,}65x + x^2 + (x-1)_+^2 + 0{,}6(x-2)_+^2.$$

Berechnen Sie die folgenden Werte:

a) $s(1{,}5)$

b) $s'(1{,}5)$

c) $s''(2{,}2)$

Aufgabe 3

Zeigen Sie, dass sich der der Nadaraya-Watson-Schätzer für die Regressionsfunktion f eines nichtparametrischen Regressionsmodells $y = f(x) + \varepsilon$ als Schätzer für den bedingten Erwartungswert $\mathbb{E}[Y|X = x]$ herleiten lässt.

Hinweis: Verwenden Sie $\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y \frac{f_{(X,Y)}(x,y)}{f_X(x)}\,dy$ und Schätzen Sie die beiden dafür benötigten Dichten mittels der Kerndichteschätzer

$$\widehat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad \text{und} \qquad \widehat{f}_{(X,Y)}(x,y) = \frac{1}{nhh'}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h'}\right).$$

Aufgabe 4

Zeigen Sie, dass eine Funktion $s : [a, b] \longrightarrow \mathbb{R}$ genau dann ein kubischer Polynom-Spline zu den Knoten $a < t_1 < t_2 \ldots < t_K < b$ ist, wenn sie sich in der Form

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + b_1(x - t_1)_+^3 + \ldots + b_K(x - t_K)_+^3$$

darstellen lässt.

Hinweis: Es genügt den Nachweis für den Spezialfall eines Polynom-Splines s mit nur einem Knoten $a < t < b$ zu erbringen, da sich dieser Nachweis leicht auf den allgemeinen Fall mit einer beliebigen Anzahl an Knoten verallgemeinern lässt.

Aufgabe 5

Die Funktion $s : [a, b] \longrightarrow \mathbb{R}$ sei ein natürlicher kubischer Polynom-Spline zu den Knoten $a < x_1 < x_2 < \ldots < x_K < b$ mit $K \geq 2$ und der Eigenschaft

$$s(x_i) = z_i \quad \text{für} \quad i = 1, \ldots, K. \tag{1}$$

D.h. der natürliche kubische Polynom-Spline s interpoliert die K Stützpunkte $(x_1, z_1), \ldots, (x_K, z_K)$. Ferner sei die Funktion $g : [a, b] \longrightarrow \mathbb{R}$ eine weitere differenzierbare Funktion, welche die Stützpunkte $(x_1, z_1), \ldots, (x_K, z_K)$ ebenfalls interpoliert.

a) Erläutern Sie, weshalb es zu K beliebig gewählten Stützpunkten $(x_1, z_1), \ldots, (x_K, z_K)$ stets einen natürlichen kubischen Polynom-Spline s gibt, der diese Stützpunkte interpoliert, d.h. für den (1) gilt.

b) Zeigen Sie mit Hilfe von partieller Integration, dass für $h(x) := g(x) - s(x)$ gilt:

$$\int_a^b s''(x)h''(x)\,dx = -\sum_{i=1}^{K-1} s'''(x_i)\big(h(x_{i+1}) - h(x_i)\big) = 0$$

c) Weisen Sie die Gültigkeit der Ungleichung

$$\int_a^b g''(x)^2\,dx \geq \int_a^b s''(x)^2\,dx$$

nach sowie dass die Gleichheit genau dann gilt, wenn $h(x) = 0$ für alle $x \in [a, b]$ erfüllt ist.

d) Zeigen Sie mit Hilfe des Ergebnisses aus Aufgabenteil c), dass das Minimierungsproblem

$$\min_f \sum_{i=1}^{K}\big(y_i - f(x_i)\big)^2 + \lambda \int_a^b f''(t)^2\,dt \tag{2}$$

nur durch einen natürlichen kubischen Polynom-Spline zu den Knoten $a < x_1 < \ldots < x_K < b$ gelöst wird.

Aufgabe 6

Es gelte:

$$\widehat{f} = \arg\min_f \sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \int_a^b f^{(m)}(x)^2 dx$$

Erläutern Sie für die folgenden Fälle, welche Schätzung $\widehat{f}$ für die Regressionsfunktion f jeweils resultiert:

a) $\lambda \to \infty$ und $m = 0$

b) $\lambda \to \infty$ und $m = 1$

c) $\lambda \to \infty$ und $m = 2$

d) $\lambda \to \infty$ und $m = 3$

e) $\lambda = 0$ und $m = 3$

Aufgabe 7

Es gelte

$$\widehat{f_1} = \arg\min_f \sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \int_a^b f^{(3)}(x)^2 dx$$

und

$$\widehat{f_2} = \arg\min_f \sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 + \lambda \int_a^b f^{(4)}(x)^2 dx.$$

Als Gütemaß wird der geschätzte mittlere quadrierte Prognosefehler

$$\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{f}(x_i)\big)^2$$

verwendet.

a) Geben Sie an ob f_1 oder f_2 für $\lambda \to \infty$ den kleineren geschätzten mittleren quadrierten Prognosefehler für den Trainingsdatensatz aufweist.

b) Geben Sie an ob f_1 oder f_2 für $\lambda \to \infty$ den kleineren geschätzten mittleren quadrierten Prognosefehler für den Validierungsdatensatz (Testdatensatz) aufweist.

c) Geben Sie an ob f_1 oder f_2 für $\lambda = 0$ den kleineren geschätzten mittleren quadrierten Prognosefehler für den Trainings- und Validierungsdatensatz aufweist.