Masterkurs Statistik – Aktualisierte Übungsbeispiele

In Klammern finden sie teilweise Lösungswerte (ohne Gewähr).

1. Gegeben ist die folgende Messung von zwei Merkmalen X und Y an n Werkstücken:

x_i	2	5	8	3	6
y_i	0.5	3	3.8	1.5	2.8

   (a) Bestimmen sie eine Regressionsgerade für die abhängige Variable Y. $[y = -0.21 + 0.527x]$

   (b) Geben sie das R^2 an. $[R^2 = 0.92]$

   (c) Testen sie die Daten auf Korrelation (z.B. $\alpha = 0.05$) und geben sie das Signifikanzniveau Ihrer Aussage an. [t=5.983]

   (d) Zeigen sie mit einem statistischen Test, dass der Anstieg der Regressionsgeraden größer als 0.25 ist (einseitig!, Fehlerwahrscheinlichkeit= 0.05). $[t = 3.146]$

2. Führen sie für die folgende Zeitreihe eine Trendrechnung mit einem loglinearen und einem halblogarithmischen Modell durch. Prognostizieren sie damit jeweils den Wert für die Periode $t = 9$. $[y = 66.81;\ y = 185.64]$

t_i	1	2	3	4	5	6
y_i	10	11	16	25	38	66

3. Bestimmen sie für die folgenden Daten eine quadratische Trendfunktion und prognostizieren sie den Wert für $t = 6$: $[y = 7.8 - 2.7x + x^2, y = 27.7]$

t_i	0	1	2	3	4
y_i	8	6	6.5	9	13

4. Für die folgenden Quartalsdaten Q_i der Jahre 2007, 2008 und 2009 soll unter Berücksichtigung der Saisonkomponenten eine lineare Trendrechnung durchgeführt werden. Prognostizieren sie damit den Wert für alle Quartale des Jahres 2011.
   $[\ Q(t) = 3.888 + 0.513\,t;\ 14.13, 11.02, 10.66, 17.67\ ]$

i	1	2	3	4	5	6	7	8	9	10	11
Q_i	6	3	2	9	8	5	5	12	10	7	6

5. Eine Buschenschank hat jeweils nur Freitag, Samstag und Sonntag geöffnet. In den Kalenderwochen 1 und 2 wurden die folgenden Umsatzzahlen beobachtet. Führen sie unter Berücksichtigung der Saisonkomponenten eine lineare Trendrechnung durch. Prognostizieren sie damit die Umsatzwerte für alle drei Öffnungstage der 4. Kalenderwoche. $[U(t) = 6 + 0.57t,\ 9.877, 14.948, 12.019]$

	KW 1			KW 2		
Tag i	1	2	3	4	5	6
Umsatz	6	9	6	6	12	9

6. Führen sie für die folgende Zeitreihe eine exponentielle Glättung mit der Methode von Holt-Winters mit den üblichen Startwerten durch. Verwenden sie die Parameter $\alpha = 0.4$, $\beta = 0.8$. $[g_{10}^H = 11.38, b_{10} = 1.6]$

t_i	1	2	3	4	5	6	7	8	9	10
y_i	2	6	2.5	7.5	3.5	10	4	13	6	16

7. Gegeben ist die lineare Funktion $z = 3 - 2.5x_1 + 3x_2 + 0.5x_3$ und folgende Daten einer Stichprobe mit $n = 4$:

i	x_1	x_2	x_3	y
1	1	1	1	1
2	2	0	3	0
3	2	1	1	0
4	3	1	0	0

   (a) Stellen sie eine logistische Regression auf und schätzen sie für alle Werte der Stichprobe das Auftreten des betrachteten Ereignisses Y.
   $[0.982, 0.377, 0.817, 0.182]$

   (b) Bestimmen sie daraus die hit-ratio und das Press'sche Q. $[h = 0.75, Q = 1]$

   (c) Um welchen Faktor ändern sich die Odds, wenn eine Variable x_i um eine Einheit erhöht wird? Bestimmen sie den Faktor für x_1, x_2, x_3. $[0.082, 20.086, 1.6487]$

   (d) Gegeben ist die Beobachtung $x_1 = 1$, $x_2 = -1$, $x_3 = 6$. Wie lauten die Odds für das Ereignis $y = 1$? [1.64] Wie lauten die Odds, wenn x_2 um eine Einheit erhöht wird? [33.12] Bestimmen sie diese neuen Odds sowohl auf direktem Weg als auch unter Verwendung der Ergebnisse von (c).

   (e) Bestimmen sie den Wert der LogLikelihoodfunktion für die Stichprobe. $[-2.395]$

   (f) Testen sie mit Hilfe der Devianz die Güte des logistischen Regressionsmodells ($\alpha = 0.1$). (Verwenden Sie zur Bestimmung der Freiheitsgrade ausnahmsweise $n = 5$.) $[Dev = 4.79,\ K = [2.706, \infty)]$

   (g) Geben sie das Signifikanzniveau ihrer Aussage aus (f) an. $[ca.2.9\%]$

Quantile der χ^2-Verteilungen:

f	$\gamma =$	0,05	0,1	0,9	0,925	0,95	0,97	0,975	0,99	0,995
1		0,004	0,016	2,706	3,17	3,841	4,709	5,024	6,635	7,879
2		0,103	0,211	4,605	5,181	5,991	7,013	7,378	9,210	10,597
3		0,352	0,584	6,251	6,905	7,815	8,947	9,348	11,345	12,838
4		0,711	1,064	7,779	8,496	9,488	10,712	11,143	13,277	14,860

8. Gegeben ist die lineare Funktion $z = 1 + 2x_1 - 0.5x_2 + 2.5x_3$ und folgende Daten einer Stichprobe mit $n = 5$:

i	x_1	x_2	x_3	y
1	1	2	1	1
2	-1	-1	0	0
3	0	2	1	1
4	2	4	-1	1
5	1.5	3	-0.5	1

   (a) Testen sie die Güte des mit z gebildeten logistischen Regressionsmodells über die Devianz ($\alpha = 0.05$). Geben sie genau die Nullhypothese und das Ergebnis des Tests an. $[Dev = 2.58,\ K = [3.841, \infty)]$

   (b) Bestimmen sie die hit-ratio und das Press'sche Q. $[h = 1, Q = 5]$

(c) Es soll der Einfluss der Variablen x_3 auf das Modell untersucht werden. Dazu wurde mit den Variablen x_1 und x_2 ein reduziertes Modell mit $z_R = 2+2.5x_1-x_2$ berechnet. Testen sie den Einfluss von x_3 auf das Modell sowohl mit $\alpha = 0.05$ als auch mit $\alpha = 0.075$. Geben sie genau die Nullhypothese und die jeweiligen Ergebnisse des Tests an.
$[Dev^R = 3.712,\ t = 1.13,\ K = [3.841, \infty), K = [3.17, \infty)]$

9. Gegeben ist eine logistische Regression mit der linearen Funktion $z = -2 + 3x_1 - 2.5x_2 + b_3x_3$. In welchem Bereich muss b_3 liegen, damit der Datensatz $x_1 = -1.5$, $x_2 = -2$, $x_3 = 1/3$, $y = 1$ richtig geschätzt wird? $[b_3 \in [4.5, \infty)\ ]$

10. Bestimmen sie für die folgenden beiden Personen A_1 und A_2 mit 11 binären Merkmalen die Werte von vier verschiedenen Ähnlichkeitsmaßen.

A_1	1	0	0	0	1	1	1	0	1	1	0
A_2	0	1	1	0	0	0	1	1	1	0	0

11. Bestimmen sie für die folgenden eindimensionalen Daten eine Clusterung. Verwenden sie dazu drei hierarchisch-agglomeratives Verfahren bis zum Ende, und zwar (i) single linkage, (ii) complete linkage, (iii) Centroid Clustering.
[(i): $(6, 7), (4, 5), (3, 4, 5), (1, 2), (1, 2, 3, 4, 5)$]

i	1	2	3	4	5	6	7	8
x_i	0,5	2,5	5,5	6,5	7,2	11	11,1	18

12. Gegeben sind die folgenden 9 Objekte mit eindimensionalen Messwerten:

i	1	2	3	4	5	6	7	8	9
x_i	3	5	7	10	14	1	6	10	17

Durch ein Clusterungsverfahren wurden die Objekte bereits in 4 Cluster eingeteilt:

$$C_1 = \{1, 6\} \quad C_2 = \{2, 3, 7\} \quad C_3 = \{4, 8\} \quad C_4 = \{5, 9\}$$

(d.h. C_1 besteht aus den Objekten mit den Index-Nummern 1 und 6, etc.)
Führen sie das hierarchisch-agglomerative Clusterungsverfahren dreimal mit unterschiedlichen Cluster-Distanzmaßen wie in Bsp. 11 aber ausgehend von der gegebenen Clusterung jeweils bis zum Ende (nur mehr ein Cluster) durch.
[Centroid: $d(C_1, C_2) = 4, d(C_2, C_3) = 4, d(C_3, C_4) = 5.5$] usw.

13. Gegeben sind 6 Messwerte im $\mathbb{R}^2$ mit den Koordinanten (x_i, y_i). Führen sie das hierarchisch-agglomerative Clusterungsverfahren mit single linkage einmal mit der L1-Norm und einmal mit der ∞-Norm bis zum Ende (nur mehr ein Cluster) durch.
$[L_1 : (2, 5), (2, 5, 6), (1, 4), (1, 2, 4, 5, 6)]$

i	1	2	3	4	5	6
x_i	2	2	-3	4	2.5	1.5
y_i	3	-2	4	2	-2	-3

14. Gegeben sind 8 Messwerte im $\mathbb{R}^2$ mit den Koordinanten (x_i, y_i).

i	1	2	3	4	5	6	7	8
x_i	-1	2	3	-2	3	5	8	11
y_i	-3	1	-4	3	-2	-3	7	12

(a) Betrachten sie die Cluster $C_1 = \{2, 4, 6\}$ und $C_2 = \{1, 8\}$. Bestimmen sie die Cluster-Distanz mit average linkage und mit dem Centroid Verfahren unter Verwendung der L1-Norm. [13.8, 7.5]

(b) Bestimmen sie den Abstand des Clusters $C_3 = \{7\}$ von C_1 mit average linkage unter Verwendung der ∞-Norm. [8.66]

(c) Fusionieren sie C_2 und C_3 und bestimmen sie den Abstand des neuen Clusters von C_1 mit single und mit complete linkage bei Verwendung der L2-Norm. [5, 16.16]

15. Gegeben sind 7 Messwerte im $\mathbb{R}^3$ mit den Koordinanten (x_i, y_i, z_i).

i	1	2	3	4	5	6	7
x_i	2	-2	3	-6	4	8	0
y_i	134	256	333	187	266	177	408
z_i	0.03	-0.06	0.08	0.1	-0.04	-0.2	0.09

Führen Sie eine Standardisierung der Daten durch.

16. Gegeben sind 5 Objekte mit den folgenden 8 binären Merkmalen:

O_1	1	0	0	1	1	1	1	1
O_2	1	1	1	0	0	0	1	0
O_3	0	0	1	1	1	0	1	1
O_4	1	1	0	1	0	0	0	1
O_5	0	0	1	0	1	0	0	0

(a) Bestimmen sie eine Ähnlichkeitsmatrix unter Verwendung des SMC-Koeffizienten und wandeln sie diese in eine Distanzmatrix um.

(b) Führen sie das hierarchisch-agglomerative Clusterungsverfahren mit complete linkage bis zum Ende durch.

(c) Was ist die Distanz der Cluster $\{O_1, O_2\}$ und $\{O_3, O_5\}$ mit average linkage? [0.5625]