HOCHSCHULE HEILBRONN

Applied Data Science
Sommersemester 2025

Prof. Dr. Florian Kauffeldt

Midterm-Examen Applied Data Science

SPO:	*Alt* ☐	*Neu* ☐
Zeit:	*45 Minuten*	*90 Minuten*

Bei alter SPO nur Frage 1 und 2 beantworten. Andere Antworten werden nicht bewertet.

Name: _____

Matr. Nummer: _____

Hinweise:

1. Zugelassene Hilfsmittel: Open-Book: Aufschriebe, Formelsammlung, Skript, Taschenrechner (keine gespeicherten Formeln etc.!), Notizen.
2. Jede Antwort muss hinreichend begründet werden. Antworten ohne Begründung ergeben 0 Punkte.
3. Unleserliche Ergebnisse werden nicht gewertet. Nutzen Sie bei weiterem Platzbedarf bitte auch die Rückseiten der Klausurblätter!
4. Die geschätzte Bearbeitungszeit (in Minuten) für eine Aufgabe entspricht der Punktzahl. Somit sind die Aufgaben insgesamt 45 Punkte wert.
5. **Viel Glück!!!**

Frage	**Punkte**	**Erreichte Punkte**
1	30	
2	15	
3	20	
4	25	
Gesamt	45 / 90	

Aufgabe 1. Interpretation Lineare Regression (30 Punkte)

Wir möchten untersuchen inwieweit die Marketingausgaben den Umsatz (Sales) eines Unternehmens beeinflussen.

Hierfür führen wir eine lineare Regression mit den folgenden Variablen durch:

- Umsatz (in tausend €)
- Marketingausgaben (in tausend €)
- Reputation des Unternehmens: Low, Medium, High
- Standort des Unternehmens (Location): Asia, Europe, USA

Die Regression ergibt folgende Koeffizienten Tabelle:

Modellkoeffizienten - Sales

Prädiktor	**Schätzung**	**Std.-fehler**	**t**	**p**
Interzept	1080.47	405.619	2.664	0.013
Marketing	2.10	0.758	2.775	0.010
Reputation:				
High – Low	2452.40	464.354	5.281	< .001
Medium – Low	1095.23	413.747	2.647	0.013
Location:				
Asia – USA	129.55	386.306	0.335	0.740
Europe – USA	67.98	377.439	0.180	0.858

a) Schreiben Sie die geschätzte Regressionsgleichung in Bezug auf den Kontext auf.

b) Interpretieren Sie alle Koeffizienten der geschätzten Regressionsgleichung.

c) Berechnen Sie den geschätzten Umsatz eines asiatischen Unternehmens mit Reputation Medium und 40.000 € Marketingausgaben.

d) Schreiben Sie die Null- und die Alternativhypothese des Signifikanztests für einen der Koeffizienten auf (in Bezug auf den Kontext).

e) Welche der unabhängigen Variablen beeinflussen signifikant den Umsatz? Begründen Sie Ihre Antwort.

f) Berechnen Sie wie viel höher die Marketingausgaben eines US-amerikanischen Unternehmens mit Reputation Low sein müssen, um den gleichen Umsatz zu erzielen wie ein US-amerikanisches Unternehmen mit Reputation High.

g) Was besagt die Kennzahl R^2 und wie wird Sie gebildet? Im vorliegenden Fall ist $R^2 = 0{,}708$. Interpretieren Sie diesen Wert. Warum wird manchmal der adjusted R^2 anstatt des einfachen R^2 verwendet?

h) Erläutern Sie, was unter Multikollinearität zu verstehen ist und weshalb sie in einem Regressionsmodell zu verzerrten oder fehlerhaften Ergebnissen führen kann. Wie lässt sich feststellen, ob Multikollinearität vorliegt? Beschreiben Sie außerdem mögliche Maßnahmen, die ergriffen werden können, um dem Problem entgegenzuwirken.

Lösungen:

a) Geschätzte Regressionsgleichung:

$$Umsatz = 1080{,}47 + 2{,}1 \cdot Marketing + 2452.4 \cdot High + 1095{,}23 \cdot Medium + 129{,}55 \cdot Asia + 67{,}98 \cdot Europe$$

b) Interpretationen:

- 1080,47 → Durchschnittlicher Umsatz eines US-amerikanischen Unternehmens mit Low Reputation und 0 Marketingausgaben
- 2,1 → Pro € Marketing steigt der Umsatz um 2,1
- 2452,4 → Eine Firma mit High Reputation macht ca. 2,4 Mio mehr Umsatz als eine mit Low
- 1095,23 → Eine Firma mit Medium Reputation macht ca. 1,1 Mio mehr Umsatz als eine mit Low
- 129,55 → Asiatische Firmen machen ca. 130 tausend mehr Umsatz als US-amerikanische
- 67,98 → Europäische Firmen machen ca. 68 tausend mehr Umsatz als US-amerikanische

c) Geschätzter Umsatz:

$$Umsatz = 1080{,}47 + 2{,}1 \cdot 40 + 1095{,}23 + 129{,}55 = 2389{,}25$$

d) Koeffizient Marketing:

$$H_0: \beta_{Marketing} = 0, H_1: \beta_{Marketing} \neq 0$$

e) Signifikanzen:

- Marketing → p-Wert = 1% < 5%
- Reputation High → p-Wert = 0,1% < 5%
- Reputation Low → p-Wert = 1,3% < 5%

f) Berechnung:

$$2{,}1 \cdot Marketing_{High} + 2452{,}4 = 2{,}1 \cdot Marketing_{Low}$$

$$\Leftrightarrow Marketing_{Low} - Marketing_{High} = 2452{,}4/2{,}1$$

g) Die Kennzahl R2 gibt Aufschluss über die Modellgüte. Sie ist das Verhältnis von erklärter zur Gesamtvarianz. 0,708 bedeutet, dass 70,8% der Unterschiede im Umsatz durch die Variablen Location, Marketing und Reputation erklärt werden

können. Der adjusted R2 wird verwendet, wenn es sehr viele unabhängigen Variablen gibt und daher eine Überanpassung des Modells nicht ausgeschlossen werden kann.

h) [Erklärung Multikollinearität in eigenen Worten]. Kann über Variance Inflation Factors festgestellt werden, wenn die Werte über 10 sind. Wenn Multikollinearität vorliegt kann man z.B. die Stichprobengröße erhöhen, Variablen ausschließen oder kombinieren.

Aufgabe 2. Interpretation Chi2-Test (15 Punkte)

Wir möchten wissen, ob der Beziehungsstatus (single, taken) von Studierenden unabhängig davon ist, ob diese bei ihren Eltern wohnen (parents_home yeah/nope).

Wir erhalten die folgende Kreuztabelle von beobachteten und erwarteten Häufigkeiten:

Kreuztabellen

		relationship		
parents_home		**single**	**taken**	**Insgesamt**
nope (jipiii)	Beobachtet	28	36	64
	Erwartet	33.3	A	64.0
yeah (unfortunately)	Beobachtet	34	21	55
	Erwartet	B	26.3	55.0
Insgesamt	Beobachtet	62	57	119
	Erwartet	62.0	57.0	119.0

a) Berechnen Sie die ausgeblendeten erwarteten Häufigkeiten A und B.

b) Schreiben Sie kontextbezogen die Null- und die Alternativhypothese eines Chi2-Tests für die Kreuztabelle auf.

c) Berechnen Sie die Teststatistik (Chi2) und die Freiheitsgrade des Chi2-Tests.

d) Berechnen Sie die Kennzahl Cramers V für den Chi2-Test. Interpretieren Sie den Wert von Cramers V. Warum wird Cramers V neben dem p-Wert betrachtet?

e) Erklären Sie warum ein Chi2-Test nicht geeignet ist, wenn parents_home eine numerische Variable wäre (z.B. Gewicht). Welchen Test würden Sie durchführen, um einen Zusammenhang zwischen Gewicht und Beziehungsstatus zu testen?

Lösungen:

a) $A = \frac{64 \times 57}{119} \approx 30{,}7; B = \frac{55 \cdot 62}{119} \approx 28{,}7$

b) Hypothesen:

H0: Relationship und parents_home sind unabhängig

H1: Relationship und parents_home sind abhängig

c) Berechnungen:

- Freiheitsgrade: (Z-1)(S-1) = 1
- Teststatistik:

$$\frac{(28-33{,}3)^2}{33{,}3} + \frac{(36-30{,}7)^2}{30{,}7} + \frac{(34-28{,}7)^2}{28{,}7} + \frac{(21-26{,}3)^2}{26{,}3} \approx 3{,}8$$

d) Cramers V (vereinfacht, da Freiheitsgrade = 1):

$$V = \sqrt{\mathrm{X}^2/n} = \sqrt{3{,}8/119} \approx 0{,}18$$

- Ein Cramers V von 0,18 wird als schwacher Zusammenhang interpretiert
- Der p-Wert gibt Aufschluss über die Signifikanz (Verallgemeinerbarkeit), Cramers V über die Stärke des Effekts

e) Der Chi2 Test wäre dann nicht geeignet, weil Gewicht sehr viele Werte annehmen kann, was zu einer sehr großen Kreuztabelle führt, bei der in jeder Zelle kaum Beobachtungen erfasst werden.

Aufgabe 3. Interpretation Logistische Regression (20 Punkte)

Wir möchten die Größe des Wohnorts vorhersagen. Hierfür führen wir eine logistische Regression mit den folgenden Variablen durch:

- Größe Wohnort (village / city)
- Interesse an Religion: Nicht interessiert 1-2-3-4-5 Sehr interessiert
- Anzahl Geschwister (siblings): natürliche Zahl (0, 1, 2,…)

Die Regression ergibt folgende Koeffizienten Tabelle:

Modellkoeffizienten - Village - town

Prädiktor	**Schätzung**	**Std.-fehler**	**Z**	**p**
Interzept	-1.632	0.1634	-9.99	< .001
Number of siblings	0.188	0.0675	2.79	0.005
Religion	0.220	0.0524	4.20	< .001

Modellkoeffizienten - Village - town

Prädiktor	Schätzung	Std.-fehler	Z	p

Anmerkung. Schätzungen sind die Log-Odds-Raten von „Village - town = village“ gegenüber „Village - town = city“

a) Schreiben Sie die geschätzte Regressionsgleichung in Bezug auf den Kontext auf.

b) Erklären Sie was Odds und was Log-Odds sind. Interpretieren Sie den Koeffizienten von „Religion“ anhand dieser Konzepte.

c) Berechnen Sie die Wahrscheinlichkeit, dass der Wohnort „village“ ist bei einer Person mit Religionsinteresse = 3 und Geschwister = 2.

d) Ab wie viel Geschwistern ist die Wahrscheinlichkeit, dass der Wohnort „village“ ist immer größer als 50%.

e) Geben Sie Gründe an, warum ein lineares Regressionsmodell in der vorliegenden Fragestellung zu fehlerhaften Ergebnissen führen kann.

Lösungen:

a) Geschätzte Gleichung:

$$P(Y = village) = \frac{e^z}{e^z + 1},$$

mit $z = -1{,}632 + 0{,}188\ siblings + 0{,}22 religion$

b) Odds = Wahrscheinlichkeit durch Gegenwahrscheinlichkeit / Log-Odds = Logarithmus der Odds. Interesse für Religion erhöht die Log-Odds von „village“ um 0,22.

c) P = 35,5%

d) Ab 8 Geschwistern

e) Geschätzte Wahrscheinlichkeiten können größer 1 oder kleiner 0 sein.

Aufgabe 4. Interpretation Korrelation (25 Punkte)

Die folgende Tabelle zeigt die paarweisen Korrelationen (r) der Ländervariablen Fertilitätsrate, Freiheitsindex und Bruttoinlandsprodukt (BIP) pro Kopf sowie die p-Werte der paarweisen Korrelationstests:

	Fertilitätsrate	Freiheit	BIP pro Kopf
Fertilitätsrate	—		

	Fertilitätsrate	Freiheit	BIP pro Kopf
Freiheit	-0.344 (p = 0,03)	—	
BIP pro Kopf	-0.475 (p = 0,014)	0.441 (p = 0,021)	—

a) Schreiben Sie die Null- und die Alternativhypothese des zweiseitigen Korrelationstests für die Variablen Fertilitätsrate und Freiheit auf.

b) Interpretieren Sie den quadrierten Wert des Korrelationskoeffizienten (r^2) von BIP pro Kopf.

c) Berechnen Sie den partiellen Korrelationskoeffizienten zwischen Fertilitätsrate und Freiheit indem Sie für BIP pro Kopf kontrollieren.

d) Erklären Sie warum bei multiplen Tests eine p-Wert-Anpassung vorgenommen werden sollte. Berechnen Sie die Holm-Bonferroni-Korrektur für die drei p-Werte aus der oberen Tabelle. Sind alle Korrelationen weiterhin signifikant?

Die folgende Tabelle zeigt das Interesse dreier Personen an Geschichte und Mathe (Nicht interessiert 1-2-3-4-5 Sehr interessiert):

Geschichte	Mathe
3	2
4	1
5	3

e) Berechnen Sie den Kendall Tau Korrelationskoeffizienten für die beiden Variablen. Interpretieren Sie diesen Wert.

f) Welche Annahmen des Pearson Korrelationstests verletzen beide Variablen. Begründen Sie Ihre Antwort kurz.

Lösungen:

a) Hypothesen: H0: Korrelationskoeffizient = 0 und H1: Korrelationskoeffizient ≠ 0.

b) $-0{,}475^2 \approx 0{,}23 \rightarrow$
$23\%\ der\ Unterschiede\ in\ der\ Fertilität\ lassen\ sich\ durch\ unterschiedliches\ BIP\ erklären$

c) Berechnung:

$$Partieller\ Korrelationskoeffizient = \frac{-0{,}344 - (-0{,}475 \times 0{,}441)}{\sqrt{1-0{,}23} \times \sqrt{1-0{,}19}} \approx -0{,}17$$

d) Eine p-Wert Anpassung ist erforderlich, da bei multiplen Tests der Alpha Fehler inflationiert wird.

Anpassung Werte: 0,014 → 0,014*3 | 0,021 → 0,021*2 | 0,03 → 0,03

Alle Korrelationen sind weiterhin < 5% und damit signifikant.

e) Kendall Tau:

Diskordante Paare: 1

Konkordante Paare: 2

$$Kendall\ Tau = \frac{2-1}{3} = \frac{1}{3}$$

f) Die Variablen sind nicht numerisch und nicht gleichverteilt.