

Bachelorprüfung WS 2025/2026

Fach: Data Science: Ökonometrie

Prüferin: Prof. Regina T. Riphahn, Ph.D.

Vorbemerkungen:

- Anzahl der Aufgaben:** Die Klausur besteht aus 4 Aufgaben, die alle bearbeitet werden müssen.
Es wird nur der Klausurbogen eingesammelt. Nur die Angaben auf dem Klausurbogen werden gewertet.
- Bewertung:** Es können maximal 60 Punkte erworben werden. Die maximale Punktzahl ist für jede Aufgabe in Klammern angegeben. Sie entspricht der für die Aufgabe empfohlenen Bearbeitungszeit in Minuten.
- Erlaubte Hilfsmittel:**
- Formelsammlung (ist der Klausur beigelegt)
 - Tabellen der statistischen Verteilungen (sind der Klausur beigelegt)
 - Taschenrechner
 - Fremdwörterbuch
 - Cheat Sheet für R (ist der Klausur beigelegt)
- Wichtige Hinweise:**
- Sollte es vorkommen, dass die statistischen Tabellen, die dieser Klausur beiliegen, den gesuchten Wert der Freiheitsgrade nicht ausweisen, machen Sie dies kenntlich und verwenden Sie den nächstgelegenen Wert.
 - Sollte es vorkommen, dass bei einer Berechnung eine erforderliche Information fehlt, machen Sie dies kenntlich und treffen Sie für den fehlenden Wert eine plausible Annahme.
 - Nutzen Sie die vorgegebenen Antwortkästchen zur Beantwortung der Teilaufgaben.

Name, Vorname	
Matrikelnr.	
Studiengang	
E-Mail-Adresse	
Datum	
Unterschrift	

Aufgabe 1	
Aufgabe 2	
Aufgabe 3	
Aufgabe 4	
Note	

Aufgabe 1:**[19,5 Punkte]**

Sie sind Managerin in einem produzierenden Unternehmen und interessieren sich für die Determinanten der Menge, die Beschäftigte pro Stunde herstellen können. Zur Analyse haben Sie einen Querschnittsdatensatz mit Informationen zu 5.495 Beschäftigten aus Deutschland.

$output_i$ Produzierte Stückzahl pro Stunde der Person i
 $bildung_i$ Bildung von Person i in Jahren
 $alter_i$ Alter von Person i in Jahren
 $\log(inc_i)$ Logarithmiertes Einkommen von Person i in 1.000€
 $west_i$ = 1, wenn Person i in Westdeutschland lebt; = 0 sonst

Sie schätzen folgendes Regressionsmodell und erhalten den unten stehenden Output:

$$output_i = \beta_0 + \beta_1 \cdot bildung_i + \beta_2 \cdot alter_i + \beta_3 \cdot alter_i^2 + \beta_4 \cdot \log(inc_i) + \beta_5 \cdot west_i + u_i \quad (1)$$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.515	4.140	7.372	0.000
bildung	1.530	0.256	?????	?????
alter	5.085	0.102	-0.814	0.413
alter2	-0.063	0.001	5.425	0.000
log(inc)	50.596	0.025	1.975	0.030
west	4.678	0.000	0.000	0.000

R-squared:	0.578			

Hinweis: Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie $\hat{\beta}_4$ inhaltlich und statistisch. (2 Punkte)

- b) Berechnen und interpretieren Sie das 99%-Konfidenzintervall für den geschätzten Koeffizienten der Variable *bildung*. Gehen Sie darauf ein, ob der Koeffizient statistisch signifikant von Null verschieden ist. (6 Punkte)

- c) Eine Kollegin behauptet, dass $\hat{\beta}_4$ vermutlich verzerrt ist, da die individuelle Motivation nicht im Modell enthalten ist. Erläutern Sie, ob der Koeffizient dann unter- oder überschätzt wird und wie man das Problem lösen könnte. (4 Punkte)

d) Bei welchem Alter ist die produzierte Stückzahl ceteris paribus maximal? (2,5 Punkte)

e) Interpretieren Sie das R^2 und nennen Sie zwei Unterschiede zum adjustierten R^2 . (3 Punkte)

- f) Ein Freund schlägt vor, das Modell um die Variable *ost* (= 1, wenn Person *i* in Ostdeutschland lebt; = 0 sonst) zu ergänzen, um noch besser nach Regionen unterscheiden zu können. Beurteilen Sie diesen Vorschlag. Welche Konsequenz hat das für die Modellschätzung? (2 Punkte)

Aufgabe 2:

[10,5 Punkte]

Sie haben den Auftrag, die Preissensibilität von Kunden einer Getränkefirma zu analysieren. Dazu haben Sie folgende Querschnittsdaten von 560 Individuen gesammelt:

$\log(\text{absatz}_i)$ Gekaufte Flaschen der Person *i* (logarithmiert)
 $\log(\text{preis}_i)$ Verkaufspreis einer Flasche in Euro (logarithmiert)
 einkommen_i Einkommen der Person *i* in 1.000 Euro
 mann_i = 1, wenn Person *i* männlich ist; = 0 sonst

Hinweis: Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

Sie schätzen folgendes Regressionsmodell und erhalten den unten stehenden Output:

$$\log(\text{absatz}_i) = \beta_0 + \beta_1 \cdot \log(\text{preis}_i) + \beta_2 \cdot \text{einkommen}_i + \beta_3 \cdot \text{mann}_i + u_i \quad (2)$$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.515	4.140	7.372	0.000
$\log(\text{preis})$	-0.113	0.001	0.000	0.000
einkommen	10.083	0.102	2.673	?????
mann	5.987	2.950	?????	?????

R-squared:	0.365			

a) Interpretieren Sie $\hat{\beta}_1$ inhaltlich und statistisch. (2 Punkte)

b) Sie möchten testen, ob sich die nachgefragte Menge zwischen Männern und Frauen in diesem Modell signifikant unterscheidet. Führen Sie dazu einen geeigneten Test auf dem 5%-Niveau durch. Geben Sie das Testverfahren, die Null- und Alternativhypothese, die Teststatistik, den kritischen Wert und die Testentscheidung an. (4,5 Punkte)

- c) Erklären Sie kurz den Begriff Heteroskedastie und gehen Sie darauf ein, was das, unter sonst idealen Bedingungen, für die Unverzerrtheit der Koeffizientenschätzung bedeutet. (2 Punkte)

- d) Nennen Sie zwei Gründe, warum es sinnvoll ist, $\log(\text{preis}_i)$ und $\log(\text{absatz}_i)$ zu logarithmieren. (2 Punkte)

Aufgabe 3:

[20 Punkte]

Sie interessieren sich für die Determinanten der Mitgliedschaft bei einem Sportverein. Sie haben Daten zu 4.200 Personen mit folgenden Informationen gesammelt:

$Sport_i$	= 1, wenn Person i Mitglied bei einem Sportverein ist; = 0, sonst
$Alter_i$	Alter von Person i in Jahren
$Bildung_i$	Bildung von Person i in Jahren
$Frau_i$	= 1, wenn Person i eine Frau ist; = 0, sonst
$Stadt_i$	= 1, wenn Person i in einer Stadt mit mehr als 100.000 Einwohnern lebt; = 0, sonst
$BevDichte_i$	Bevölkerungsdichte des Wohnortes von Person i , gemessen in Einwohner pro Quadratkilometer

Sie führen eine Kleinstquadrateschätzung für folgendes Modell (Modell 1) durch und erhalten untenstehenden Output:

$$Sport_i = \beta_0 + \beta_1 \cdot Alter_i + \beta_2 \cdot Bildung_i + \beta_3 \cdot Stadt_i + u_i$$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.126	0.024	5.250	0.000
Alter	-0.006	0.001	-6.000	0.000
Bildung	0.082	0.009	9.111	0.000
Stadt	0.084	0.022	3.818	0.000

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

- a) Interpretieren Sie den geschätzten Koeffizienten $\hat{\beta}_1$ inhaltlich und statistisch. (2 Punkte)

- b) Basierend auf der Schätzung von Modell (1), wie hoch ist die Wahrscheinlichkeit der Mitgliedschaft in einem Sportverein für eine Person die 46 Jahre alt ist, 10 Jahre Bildung hat und in einer Stadt mit mehr als 100.000 Einwohnern lebt? (3 Punkte)

- c) Berechnen Sie, welchen Wert $\hat{\beta}_2$ annimmt, wenn die Bildung nicht in Jahren sondern in Monaten gemessen wird und zeigen Sie ihren Rechenweg. Interpretieren Sie den resultierenden Koeffizienten. Welche Auswirkung hat diese Umskalierung auf die t-Statistik des Koeffizienten? (4 Punkte)

- d) Sie erweitern das Modell um die Variable $BevDichte_i$ (Modell 2):

$$Sport_i = \beta_0 + \beta_1 \cdot Alter_i + \beta_2 \cdot Bildung_i + \beta_3 \cdot Stadt_i + \beta_4 \cdot BevDichte_i + \varepsilon_i$$

Welche Größen beeinflussen die Varianz eines Koeffizientenschätzers im multiplen Regressionsmodell und in welche Richtung? Durch die Hinzunahme der Variable $BevDichte_i$ ist $\hat{\beta}_3$ nun statistisch insignifikant. Geben Sie eine Begründung hierfür an. (5 Punkte)

- e) Sie vermuten, dass sich die Parameter von Modell 2 für Frauen und Männer unterscheiden. Um diese Vermutung zu überprüfen, führen Sie nun einen Chow-Test auf Strukturbruch für das Modell am 5%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung an. (6 Punkte)

Hinweise: $SSR_{pooled} = 299$, $SSR_1 = 134$ (für $Frau = 0$), $SSR_2 = 155$ (für $Frau = 1$)

Aufgabe 4:**[10 Punkte]**

Ein Hotel möchte seine Buchungen analysieren. Erfasst wurden der Preis pro Nacht (in Euro), die Anzahl der gebuchten Nächte, die Buchungsart (Geschäftsreise = 1, Urlaubsreise = 0) und das Alter des Kunden. Eine Mitarbeiterin hat einen Beispieldatensatz mit vier Beobachtungen unter dem Namen 'buchungen' erstellt.

Tabelle 1: buchungen

preis	naechte	reiseart	alter
200	2	1	33
150	3	1	48
170	3	0	22
220	2	1	42

a) Erklären Sie knapp, was mit folgendem R-Code ausgegeben wird. (1,5 Punkte)

```
buchungen[3, 2]
```

b) Erklären Sie knapp, was mit folgendem R-Code ausgegeben wird. (1,5 Punkte)

```
mean(buchungen$naechte[buchungen$alter>30])
```

- c) Für die weitere Analyse wurde nun der vollständige Datensatz 'buchungen' mit 522 Beobachtungen verwendet. Erläutern Sie, was durch diesen R-Code programmiert ist. (1,5 Punkte)

```
buchungen$preis[buchungen$preis<50] <- NA
```

- d) Erklären Sie, was in den einzelnen Zeilen (1-2) des folgenden R-Codes programmiert ist. (3,5 Punkte)

```
Zeile 1: model <- lm(preis ~ reiseart + alter, data=buchungen, subset=naechte>=3)  
Zeile 2: summary(model)
```

e) Erläutern Sie, was dieser R-Code ausführt. (2 Punkte)

```
buchungen <- buchungen %>%  
mutate(lang = ifelse(naechte > 3, 2, 1))
```