Bachelorprüfung

Fach: Praxis der empirischen Wirtschaftsforschung

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

Name, Vorname	
Matrikelnr.	
E-Mail	
Studiengang	
Semester	
Datum	
Raum, Platznr.	
Unterschrift	

Vorbemerkungen:

Anzahl der Aufgaben: Die Klausur besteht aus 6 Aufgaben, die alle bearbeitet werden müssen.

Bewertung: Es können maximal 90 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 beigefügt)
- Tabellen der statistischen Verteilungen (sind der Klausur beigefügt)
- Taschenrechner
- Fremdwörterbuch

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.

Aufgabe 1: **[14,5 Punkte]**

Für einen Handy-Anbieter sollen Sie die Determinanten der Kundenbindung untersuchen. Der Datensatz beinhaltet folgende Angaben zu 1 000 Kunden:

loyalty	Dauer der Kundenbindung (in Monaten)
age	Alter bei Vertragsabschluss (in Jahren)
male	=1, wenn männlich; =0, sonst
income	Jahreseinkommen (in Tausend $)
married	=1, wenn verheiratet; =0, sonst

Sie unterstellen die Gültigkeit der Gauss-Markov Annahmen und schätzen folgendes Modell:

$$\text{loyalty}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{income}_i + \beta_3 \text{married}_i + u_i$$

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		T	Sig.
		RegressionskoeffizientB	Standardfehler		
1	(Konstante)	-2,116	2,094	-1,010	,313
	age	,782	,049	16,109	,000
	income	,020	,006	3,475	,001
	married	7,053	1,153	6,117	,000

a. Abhängige Variable: loyalty

a) Berechnen und interpretieren Sie den geschätzten Effekt einer Erhöhung des **monatlichen** Einkommens (β_2) um Tausend $. (2 Punkte)

b) Ein Kundenberater behauptet, dass verheiratete Kunden im Durchschnitt mehr als vier Monate länger bleiben als unverheiratete. Testen Sie diese Hypothese auf dem 5%-Niveau. Geben Sie Testverfahren, Null- und Alternativhypothese, Teststatistik, Freiheitsgrade, kritischen Wert und Ihre Testentscheidung an. (4,5 Punkte)

c) Berechnen Sie die erwartete Dauer der Kundenbindung für einen 22-jährigen, unverheirateten Neukunden mit einem Jahreseinkommen von 25 Tausend $. (2 Punkte)

d) Sie möchten prüfen, ob sich der Effekt des Einkommens für weibliche und männliche Kunden statistisch signifikant unterscheidet. Geben Sie die Regressionsgleichung für eine Modellspezifikation an, die Ihnen das gewünschte Ergebnis liefert. (2 Punkte)

e) Berechnen Sie das 99% Konfidenzintervall für den Parameter $\hat{\beta}_1$ (age), zeigen Sie Ihren Rechenweg und interpretieren Sie das Ergebnis. (4 Punkte)

Aufgabe 2: **[8 Punkte]**

a) Leiten Sie den Kleinstquadrateschätzer für den Steigungsparameter β_1 im einfachen linearen Modell **ohne Konstante** her. (6 Punkte)

b) Wann und warum kann es im Allgemeinen sinnvoll sein, ohne Konstante zu schätzen? Begründen Sie und nennen Sie ein Beispiel für einen solchen Sachverhalt. (2 Punkte)

Aufgabe 3: **[8 Punkte]**

Sie analysieren die Arbeitszufriedenheit in Deutschland anhand der Daten des Sozio-oekonomischen Panels (SOEP) des Jahres 2008. Ihr Datensatz umfasst 6 623 Personen und enthält folgende Variablen:

satwork	subjektive Arbeitszufriedenheit, gemessen auf einer Skala von 0=sehr unzufrieden bis 10=sehr zufrieden
female	=1, wenn weiblich; =0, sonst
male	=1, wenn männlich; =0, sonst

Sie interessieren sich für die Unterschiede zwischen den Geschlechtern und schätzen folgendes lineares Modell:

$$\mathrm{satwork_i} = \beta_1 \mathrm{female_i} + \beta_2 \mathrm{male_i} + \mathrm{u_i}$$

Koeffizienten[a]

	Nicht standardisierte Koeffizienten			
Modell	RegressionskoeffizientB	Standardfehler	T	Sig.
female	6,920	,039	175,72	,000
male	6,938	,040	171,72	,000

a. Abhängige Variable: *satwork*

a) Ist die Annahme MLR.3 in dem so geschätzten Modell verletzt? Begründen Sie. (2 Punkte)

b) Interpretieren Sie statistisch und inhaltlich den geschätzten Parameter für *female* (β_1) (2 Punkte)

c) Wie kann mit dem vorliegenden Datensatz im Rahmen einer Regression die statistische Signifikanz des Geschlechterunterschiedes bestimmt werden? Skizzieren Sie Ihre Vorgehensweise. (4 Punkte)

Aufgabe 4: **[18 Punkte]**

Sie interessieren sich für die Determinanten des Gesundheitszustands. Sie entscheiden sich, den Gesundheitszustand mittels Body-Mass-Index (BMI) zu messen. Ihnen steht ein Datensatz mit Angaben von 4 284 Personen zur Verfügung. Sie betrachten die folgenden Variablen:

BMI	Body-Mass-Index. Einheit: Indexpunkte
lnTV	Natürlicher Logarithmus der Fernsehdauer in Minuten pro Tag
educ	Anzahl der Schuljahre
age	Alter einer Person in Jahren
age2	quadriertes Alter einer Person
Bayern	=1, wenn in Bayern wohnhaft; =0, sonst

Sie schätzen folgendes Modell I:

$$\textbf{Modell I: } \mathrm{BMI}_i = \beta_0 + \beta_1 \ln \mathrm{TV}_i + \beta_2 \mathrm{educ}_i + \beta_3 \mathrm{age}_i + \beta_4 \mathrm{age}_i^2 + u_i$$

Koeffizienten[a]

		Nicht standardisierte Koeffizienten			
Modell		RegressionskoeffizientB	Standardfehler	T	Sig.
1	(Konstante)	16,966	,822	20,646	,000
	lnTV	,701	,110	6,395	,000
	educ	-,166	,025	-6,652	,000
	age	,287	,022	13,277	,000
	age2	-,002	,000	-11,241	,000

a. Abhängige Variable: BMI

a) Interpretieren Sie den geschätzten Effekt der logarithmierten Fernsehdauer *lnTV* (β_1) statistisch und inhaltlich. (2 Punkte)

b) In welchem Alter ist der durchschnittliche Body-Mass-Index maximal? (2 Punkte)

c) Die Fernsehdauer TV wird nicht wie bisher in Minuten, sondern in Stunden pro Tag gemessen.

i. Zeigen Sie die Konsequenzen der Umskalierung für die Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$. (4 Punkte)
Hinweis: $\ln(ab) = \ln a + \ln b$

ii. Berechnen Sie die neuen Koeffizienten, wenn TV in Stunden gemessen wird. (2 Punkte)

Ihr Kommilitone vermutet, dass in Bayern lebende Personen einen höheren BMI haben. Um dies zu untersuchen schätzen sie das folgende Modell II:

$$\textbf{Modell II: } \mathrm{BMI}_i = \beta_0 + \beta_1 \ln \mathrm{TV}_i + \beta_2 \mathrm{educ}_i + \beta_3 \mathrm{age}_i + \beta_4 \mathrm{age}_i^2 + \beta_5 \mathrm{Bayern}_i + u_i$$

ANOVA

Modell		Quadrat-summe	df	Mittel der Quadrate	F	Sig.
1	Regression	7613,386	5	1522,677	82,334	,000[a]
	Nicht standardisierte Residuen	79116,956	4278	18,494		
	Gesamt	86730,343	4283			

Koeffizienten[a]

Modell		Nicht standardisierte Koeffizienten		T	Sig.
		RegressionskoeffizientB	Standardfehler		
1	(Konstante)	17,025	,826	20,614	,000
	lnTV	,694	,110	6,309	,000
	educ	-,167	,025	-6,679	,000
	age	,287	,022	13,287	,000
	age2	-,002	,000	-11,254	,000
	Bayern	-,138	,192	-,720	,472

a. Abhängige Variable: BMI

d) Interpretieren Sie den geschätzten Effekt der Variable *Bayern* (β_5) statistisch und inhaltlich. (2 Punkte)

e) Sie wollen den die Güte der geschätzten Modelle beurteilen.

i. Berechnen Sie das Bestimmtheitsmaß R_{II}^2 und das angepasste Bestimmtheitsmaß $\bar{R}_{II}^2$ für das Modell II. (4 Punkte)

ii. Im Modell I beträgt das Bestimmtheitsmaß $\mathrm{R}_\mathrm{I}^2 = 0{,}0877$ und das angepasste Bestimmtheitsmaß $\bar{R}_I^2 = 0{,}0868$. Würden Sie auf Basis der Ergebnisse das Modell I oder II bevorzugen? Geben Sie zwei Gründe an. (2 Punkte)

Aufgabe 5: **[27.5 Punkte]**

Wahr oder falsch? Tragen Sie für jede der folgenden Aussagen ein „w“ für „wahr“ oder ein „f“ für „falsch“ ein. Für jede richtige Antwort gibt es 0,5 Punkte, für jede falsche Antwort werden 0,5 Punkte abgezogen. Die Gesamtpunktzahl in dieser Aufgabe kann nicht negativ werden.

	Bei der Alternativhypothese $H_1: \beta_1 \neq 2$ wird ein einseitiger t-Test durchgeführt.
	$\frac{1}{n}\sum_i u_i^2$ ist ein unter den Gauss-Markov Annahmen unverzerrter Schätzer für die Varianz der Störterme u_i im Modell $y_i = \beta_0 + \beta_1 x_i + u_i$.
	Ordinale Informationen können nicht mittels mehrerer Dummy-Variablen abgebildet werden.
	Eine Nullhypothese, die am Signifikanzniveau $\frac{\alpha}{2}$ verworfen wurde, wird am Niveau α ebenfalls verworfen.
	Der Typ 2 Fehler wird begangen, wenn eine unzutreffende H_0 verworfen wird.
	Wenn als abhängige Variable das Einkommen betrachtet wird, ist eine Logarithmierung der unabhängigen Variablen für eine sinnvolle Modellspezifikation zwingend notwendig.
	Das Signifikanzniveau beschreibt die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, obwohl sie zutrifft.
	Der KQ-Schätzer minimiert die Summe der quadrierten vertikalen Abweichungen von der Regressionsgeraden.
	Wenn für den KQ-Schätzer die Bedingungen für Unverzerrtheit erfüllt sind, ist dieser konsistent.
	Signifikanztests testen, ob geschätzte Parameter signifikant von Null verschieden sind.
	Der KQ-Schätzer maximiert das R^2.
	Die F-Verteilung ist asymmetrisch.
	Die geschätzte Varianz des KQ-Schätzers für die Konstante ist eine Zufallsvariable.
	Unter der Alternativhypothese H_1 ist die Teststatistik beim t-Test t-verteilt.
	Bei einem t-Test braucht man Zähler- und Nennerfreiheitsgrade.
	Das Auslassen relevanter Variablen kann zu Verzerrungen der geschätzten Parameter führen.
	Im linearen Modell ist der Erwartungswert der Konstante gleich Null.
	Unter den Gauss-Markov-Annahmen MLR.1-MLR.4 ist der KQ-Schätzer konsistent.
	Perfekte Multikollinearität kann durch größere Stichproben behoben werden.
	Durch Logarithmierung der abhängigen Variable verlieren Ausreißerbeobachtungen an Bedeutung.
	Effiziente Schätzverfahren sind immer normalverteilt.
	Bei Multikollinearität steigen die Standardfehler.
	Die Nennerfreiheitsgrade des F-Tests hängen von der Anzahl der Beobachtungen ab.
	Ein Steigungsparameter kann kausal interpretiert werden, wenn alle anderen Einflussgrößen der abhängigen Variable im Modell berücksichtigt sind.
	Wenn die Varianz des Störterms mit einer der erklärenden Variablen sinkt, liegt Heteroskedastie vor.
	Schätzt man eine Regressionsgleichung mit einem Interaktionsterm, dann muss der Interaktionsterm signifikant sein, um unverzerrte Schätzergebnisse zu erhalten.
	Für eine stetige Variable x gilt immer $P(x = 2) = 0$.
	Große Standardfehler implizieren breite Konfidenzintervalle.
	Eine Saisonbereinigung für vier Jahreszeiten kann mit Hilfe von 3 Dummy-Variablen erfolgen.
	Wenn zwei Variablen x und z statistisch unabhängig sind, dann gilt $\mathrm{Cov}(x, z) = 0$.
	Bei perfekter Multikollinearität ist KQ-Schätzung möglich, aber unpräzise.
	Ein vollständig interagiertes Modell erlaubt es, Unterschiede zwischen den Steigungsparametern für zwei Gruppen abzubilden.
	Das angepasste Bestimmtheitsmaß $\bar{R}^2$ kann nicht verwendet werden, um genestete Modelle miteinander zu vergleichen.
	Schätzt man fälschlicherweise mit Konstante, so sind die Steigungsparameter verzerrt.
	Auch beim Schätzen ohne Konstante können die Schätzer unverzerrt sein.
	Das Bestimmtheitsmaß R^2 im Modell $y = \beta_0 + \beta_1 x + u$ ist der quadrierte Korrelationskoeffizient von x und y.
	Die t-Verteilung ist um die Eins symmetrisch.
	Bei einem Chow-Test benötigt man Zähler- und Nennerfreiheitsgrade.
	Konsistenz und asymptotische Effizienz sind für $n \to \infty$ definiert.
	Ein niedriges R^2 im multiplen Regressionsmodell deutet auf eine unpräzise Schätzung der Parameter hin.
	Wenn binäre Variablen mit den Werten -1 und 1 kodiert werden, wird der Schätzer verzerrt, gewinnt aber an Effizienz.

	Die Varianz einer Konstanten ist immer gleich 1.
	Nach einer KQ-Schätzung eines Modells ohne Konstante ist der Stichprobendurchschnitt der Residuen per Definition gleich Null.
	Ein Regressor ist endogen, wenn er mit dem Störterm unkorreliert ist.
	Ein Schätzverfahren für β_j ist konsistent, falls $\text{p}\lim_{n\to\infty} \hat{\beta}_j = \beta_j$.
	Wird $\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 + u$ geschätzt, so ergeben sich vier Bedingungen erster Ordnung.
	Mit Hilfe der KQ-Methode lässt sich der Schnittpunkt der Regressionsgeraden mit der x-Achse berechnen.
	Berechnet man für eine Variable aus einem selbst erstellten Datensatz sowohl Varianz als auch Stichprobenvarianz, so ist die Varianz immer größer als die Stichprobenvarianz.
	Im Modell $y = \beta_0 + \beta_1(x_1 - x_2) + \beta_2 x_3 + u$ werden drei Steigungsparameter geschätzt.
	Im einfachen linearen Regressionsmodell gibt die Konstante den Mittelwert der abhängigen Variable an.
	Beim natürlichen Logarithmus gilt: $\ln(e^x) = x - 1$.
	$y = \beta_0 + \beta_1 \sqrt[3]{x} + u$ ist nicht-linear in Parametern.
	Durch einen Interaktionsterm zwischen einer Dummy-Variable und einer stetigen Variable wird ein gruppenspezifischer Steigungsparameter modelliert.
	Wenn ein perfekter linearer Zusammenhang vorliegt, nimmt der Korrelationskoeffizient ρ den Wert 0 an.
	Im multiplen Regressionsmodell lassen sich auch nicht-lineare Zusammenhänge abbilden.

Aufgabe 6: **[15 Punkte]**

Welche Antwort ist richtig? Kreuzen Sie nur **eine Antwort** pro Aufgabe an. Falls mehrere Aussagen korrekt sind, kreuzen Sie **nur** die entsprechende **Antwortkombination** an. Für jede richtige Antwort gibt es 1 Punkt. Für falsche Antworten werden in dieser Aufgabe keine Punkte abgezogen.

1.	Im Modell $\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$	
a	❑	ändert sich y c.p. im Durchschnitt um β_1% wenn x_1 um eine Einheit steigt.
b	❑	ändert sich y c.p. im Durchschnitt um $100\beta_1$% wenn x_1 um 1% steigt.
c	❑	ändert sich y im Durchschnitt um β_1% gegeben, dass x_1 um eine Einheit steigt und x_2 konstant bleibt.
d	❑	ändert sich y c.p. im Durchschnitt um $100(\beta_1 + \beta_2)$% wenn x_1 und x_2 gleichzeitig jeweils um eine Einheit steigen.
e	❑	c und d
f	❑	keine der genannten Antworten

2.	Ergibt sich nach einer KQ-Schätzung ein Residuum für die 4. Beobachtung $\hat{u}_4 = 5$,	
a	❑	so wird y_4 unterschätzt.
b	❑	so wird y_4 überschätzt.
c	❑	so liegt y_4 auf der Regressionsgerade.
d	❑	so ist die Annahme $E(u\|x) = E(u) = 0$ verletzt.
e	❑	a und c
f	❑	a und d

3.	Trifft $E(u\|x_1, x_2) = 0$ im Modell $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ **nicht** zu,	
a	❑	so ist mindestens eine erklärende Variable endogen.
b	❑	so ist mindestens eine erklärende Variable exogen.
c	❑	so sind die Parameter verzerrt.
d	❑	so werden die Parameter unpräzise geschätzt.
e	❑	a und c
f	❑	a und d

4.	Wird nach einer KQ-Schätzung ein 95%-Konfidenzintervall (KI) für β_2 berechnet,	
a	❑	so liegt $\hat{\beta}_2$ mit 95% Wahrscheinlichkeit innerhalb der Intervallgrenzen.
b	❑	so liegt β_2 mit 95% Wahrscheinlichkeit innerhalb der Intervallgrenzen.
c	❑	so wird die Breite des KI durch die Varianz von $\hat{\beta}_2$ beeinflusst.
d	❑	so ändern sich bei wiederholter Stichprobenziehung und Berechnung des KI die Intervallgrenzen.
e	❑	b und c
f	❑	c und d

5.	Wenn der p-Wert für den Test der Nullhypothese $H_0: \beta_1 = 0$ gleich 0.08 ist, dann kann der p-Wert für den Test der Nullhypothese	
a	❑	$H_0: \beta_1 \geq 0$ gleich 0.16 sein.
b	❑	$H_0: \beta_1 \leq 0$ gleich -0.08 sein.
c	❑	$H_0: \beta_1 \geq 0$ gleich 0.08 sein.
d	❑	$H_0: \beta_1 \geq 0$ gleich 0.04 sein.
e	❑	a und b
f	❑	keine der genannten Antworten

6.	Wird das Modell $\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ auf Gesamtsignifikanz getestet,	
a	❑	so lautet das restringierte Modell log $(y) = \beta_0$+ u.
b	❑	so kann die Nullhypothese als $H_0: \beta_1 = \beta_2 = 0$ formuliert werden.
c	❑	so ist die Teststatistik F-verteilt.
d	❑	so kann das Modell statistisch signifikant sein, obwohl die individuellen Koeffizienten insignifikant sind.
e	❑	a und c.
f	❑	alle genannten Antworten.

7.	Beim Umskalieren der abhängigen Variable im Modell $y = \beta_0 + \beta_1 x + u$ ändern sich	
a	❑	die geschätzten Koeffizienten nicht.
b	❑	die Standardfehler der Koeffizienten.
c	❑	die t-Statistiken.
d	❑	die Konfidenzintervalle für beide Koeffizienten.
e	❑	b und d
f	❑	a und c

8.	Wenn $Cov(x, u) = 0$, dann	
a	❑	ist $E(u\|x) = E(u) = 0$.
b	❑	ist der KQ-Schätzer unverzerrt.
c	❑	ist die ceteris-paribus Interpretation legitim.
d	❑	besteht kein linearer Zusammenhang zwischen den erklärenden Variablen und den Störtermen.
e	❑	b und d
f	❑	a und c

9.	Die Stichprobenvarianz	
a	❑	beschreibt die Summe der quadrierten Abweichungen einer Variable um ihren Mittelwert.
b	❑	ist kein Dispersionsmaß.
c	❑	ist ein erwartungstreuer Schätzer für das R^2
d	❑	nimmt nie den Wert Null an.
e	❑	a und c
f	❑	keine der genannten Antworten

10.	Gepoolte Querschnittsdaten	
a	❑	enthalten für eine Beobachtungseinheit eine Messung.
b	❑	enthalten wiederholte Messungen für jede Beobachtungseinheit.
c	❑	sind Kombinationen von Zeitreihenerhebungen zu verschiedenen Erhebungszeitpunkten.
d	❑	werden auch als Paneldaten bezeichnet.
e	❑	b und e
f	❑	keine der genannten Antworten

11.	Das Bestimmtheitsmaß R^2	
a	❑	kann nicht zum Vergleich von Modellen herangezogen werden.
b	❑	ist umso kleiner je mehr Erklärungsgehalt das Modell besitzt.
c	❑	gibt den Anteil der durch das Modell erklärten Variation in der abhängigen Variable an.
d	❑	kann genauso interpretiert werden wie das adjustierte Bestimmtheitsmaß $\bar{R}^2$.
e	❑	a und c
f	❑	c und d

12.	Bei konsistenten Schätzverfahren	
a	❑	ist der Schätzer immer unverzerrt.
b	❑	sinkt die Varianz des Schätzers mit steigender Beobachtungszahl n.
c	❑	ist die Anzahl der Beobachtungen n nicht relevant.
d	❑	liegt der Schätzer umso näher am wahren Wert, desto größer die Stichprobengröße n.
e	❑	a und c
f	❑	b und d

13.	Die Annahme $x \sim \text{i.i.d.}(\mu, \sigma^2)$ impliziert unter anderem, dass	
a	❑	$E(\mu) = 0$
b	❑	$E(\sigma^2) = 0$
c	❑	$E(\mu) = x$
d	❑	$Var(x) = \sigma^2$
e	❑	a und d
f	❑	keine der Antworten

14.	Der Typ 1 Fehler	
a	❑	wird begangen, wenn eine falsche H_0 nicht verworfen wird.
b	❑	wird begangen, wenn eine wahre H_0 verworfen wird.
c	❑	tritt ein mit der Wahrscheinlichkeit 1-α.
d	❑	tritt ein mit der Wahrscheinlichkeit α.
e	❑	a und d
f	❑	b und d

15.	Für eine unverzerrte Schätzung des wahren Bevölkerungsparameters mittels KQ ist erforderlich,	
a	❑	dass der Erwartungswert der Konstante Null ist.
b	❑	dass die Störterme normalverteilt sind.
c	❑	dass der Erwartungswert des Störterms abhängig von den erklärenden Variablen ist.
d	❑	dass die Stichprobe gegen unendlich geht.
e	❑	b und d
f	❑	keine der genannten Antworten