Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Prof. Regina T. Riphahn, Ph.D.

Bachelorprüfung WS 2020/2021 - MUSTERLÖSUNG

Fach: Praxis der empirischen Wirtschaftsforschung

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

Vorbemerkungen:

Anzahl der Aufgaben:	Die Klausur besteht aus 3 Aufgaben, die alle bearbeitet werden müssen. **Es wird nur der Lösungsbogen eingesammelt.** Angaben auf dem Aufgabenzettel werden nicht 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 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: **[18 Punkte]**

Sie möchten den durchschnittlichen Kraftstoffverbrauch pro Kilometer vorhersagen. Der Datensatz enthält folgende Variablen für 345 Autos:

$lp100_i$	Kraftstoffverbrauch pro 100 Kilometer in Liter
$weight_i$	Gewicht des Autos in 100 Kilogramm.
hp	Motorleistung in Pferdestärke (PS)
$foreign_i$	=1, wenn das Auto von einem ausländischen Hersteller stammt, =0 sonst

Sie schätzen das folgende Modell:

$$lp100_i = \beta_0 + \beta_1 ln(weight)_i + \varepsilon_i \tag{1}$$

und erhalten die Koeffizientenschätzer: $\hat{\beta}_0 = 4,923$ und $\hat{\beta}_1 = 67,8$ mit $se(\hat{\beta}_0) = 2,267$ und $se(\hat{\beta}_1) = 21,5$

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

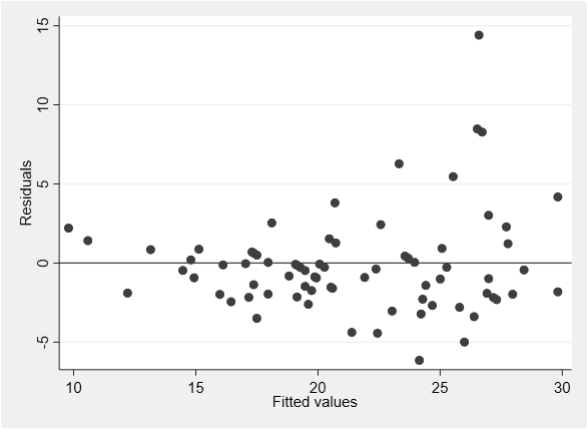
a) Interpretieren Sie den Schätzer $\hat{\beta}_1$ inhaltlich und testen Sie seine statistische Signifikanz am 1% Niveau. (4 Punkte)

- 1% zusätzliches Gewicht ist c.p.i.M. mit 67,8/100 = 0,678 Liter mehr Kraftstoffverbrauch pro 100 Kilometer assoziiert. [1P]
- $t_{emp} = \frac{67,8-0}{21,5} = 3,153$ [1P]
- $c = t_{0,005;345-1-1} \approx t_{0,005;\infty} = 2,57$ [1P]
- Der Koeffizient ist am 1%-Niveau statistisch signifikant von Null verschieden, da $t_{emp} = 3.153 > 2,57 = c$. [1P]

b) Ein Kommilitone macht Sie darauf aufmerksam, dass der Koeffizientenschätzer von $ln(weight)$ überschätzt sein könnte. Sie vermuten daraufhin, dass eine relevante Variable ausgelassen wurde und schätzen die Gleichung erneut mit der Variable hp. Welche Bedingungen muss die Variable hp erfüllen, damit das Auslassen zu einer Überschätzung von $\hat{\beta}_1$ führt? Verdeutlichen Sie jeweils am Beispiel von hp die beiden Bedingungen, die dafür erfüllt sein müssen. (3 Punkte)

- Die ausgelassene Variable muss in einem positiven Zusammenhang mit dem Kraftstoffverbrauch stehen. [1P] Ein Motor mit mehr Leistung führt zu einem höheren Kraftstoffverbrauch. [0,5P]
- Die ausgelassene Variable muss positiv mit $ln(weight)$ korrelieren.[1P] Ein leistungsstärkerer Motor ist größer und führt zu einem höheren allgemeinen Gewicht des Autos. [0,5P] (Andere Antworten möglich. Antwort auch korrekt, wenn zwei negative Zusammenhänge vorliegen.)

c) Eine Betrachtung der Residuen kann Aufschluss über die Störterme geben. Betrachten Sie folgende Grafik. Welche der Gauß-Markov Annahme könnte verletzt sein? Begründen Sie Ihre Antwort kurz. (2 Punkte)



- Es liegt voraussichtlich Heteroskedastie vor [1P], da sich die Varianz der Residuen für verschiedene Werte der abhängigen Variable unterscheidet. [1P]

d) Als nächstes schätzen Sie das Modell

$$lp100_i = \beta_0 + \beta_1 weight_i + \beta_2 hp_i + \varepsilon_i \tag{2}$$

Sie vermuten jedoch, dass sich die Parameter des Modells (2) für deutsche und ausländische Hersteller unterscheiden. Erläutern Sie kurz das Vorgehen und die Entscheidungslogik des Chow-Tests auf Strukturbruch, den Sie mittels der Variable $foreign$ durchführen können. (3 Punkte)

- Alle erklärenden Variablen aus Modell 2 werden mit der Variable $foreign$ interagiert. Diese Interaktionsterme [0.5P] und die Variable $foreign_i$ [0.5P] werden als zusätzliche erklärende Variablen in das Modell (2) aufgenommen (vollständig interagiertes Modell).
- Anschließend wird ein F-Test auf gemeinsame Signifikanz der zusätzlich aufgenommenen Variablen durchgeführt. [1P]
- Ergibt der F-Test Hinweise auf gemeinsame Signifikanz, so wird davon ausgegangen, dass ein Strukturbruch vorliegt und sich die Parameter des Modells für deutsche und ausländische Hersteller unterscheiden. [1P]

e) Um die Vermutung aus Aufgabe d) zu überprüfen, führen Sie nun einen Chow-Test auf Strukturbruch für das Modell (2) am 1%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung an. (6 Punkte)
Hinweise: $SSR_{pooled} = 299,567$, $SSR_1 = 135,674$ (für $foreign = 0$), $SSR_2 = 155,210$ (für $foreign = 1$)

- Hypothesen:
  H_0 : Es besteht kein Strukturbruch zwischen den deutschen und ausländischen Herstellern, oder $\beta_{j,g=1} = \beta_{j,g=2}$ mit $j = 0,\ldots,k$ [0.5P]
  H_1 : Es besteht ein Strukturbruch zwischen den deutschen und ausländischen Herstellern, oder $\beta_{j,g=1} \neq \beta_{j,g=2}$ mit $j = 0,\ldots,k$ [0.5P]
- Teststatistik: $F_{Chow} = \frac{\frac{SSR_p - (SSR_1 + SSR_2)}{k+1}}{\frac{(SSR_1 + SSR_2)}{(n-2(k+1))}} = \frac{\frac{299,567 - (135,674 + 155,210)}{2+1}}{\frac{135,674 + 155,210}{345-2(2+1)}} = \frac{2,894}{0,858} = 3,373$ [3P]
- kritischer Wert: $F_{0,01;3;339} = 3,78$ [1P]
  (krit. Wert nicht tabelliert für df=339 $\rightarrow df = \infty$)
- Testentscheidung: Da $F_{Chow} = 3,373 < 3,78 = c$ kann die Nullhypothese auf dem 1%-Niveau nicht verworfen werden. Das Modell unterscheidet sich nicht für deutsche und ausländische Hersteller. [1P]

Aufgabe 2: **[20 Punkte]**

Sie wollen die Determinanten des monatlichen Bruttoeinkommens untersuchen. Ihnen liegen folgende Variablen für 536 Personen vor:

$income_i$	monatliches Bruttoeinkommen in Euro
$female_i$	=1, wenn Person weiblich; =0,wenn männlich
$educ_i$	Bildung in Jahren
age_i	Alter in Jahren
$experience_i$	Berufserfahrung in Jahren

Sie schätzen das folgende Modell:

$$income_i = \beta_0 + \beta_1 female_i + \beta_2 age_i + \beta_3 educ_i + \beta_4 experience_i + \beta_5 experience \cdot educ_i + u_i$$

Modellzusammenfassung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	0,287(a)	0,136	???	0,415

Koeffizienten[a]

Modell	Nicht standardisierte Koeffizienten			
	RegressionskoeffizientB	Standardfehler	T	Signifikanz
(*Konstante*)	940,473	64,427	14,5970	0.000
female	-167,352	57,195	-2,926	0,012
age	20,512	4,483	4,576	0,000
educ	24,398	7,190	3,161	0,007
experience	31,454	12.364	2,544	0,093
experience · educ	12,955	1,872	6,920	0,002

a. Abhängige Variable: *income*

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie den geschätzten Koeffizienten von *female* inhaltlich. Ist der Effekt statistisch signifikant? (2 Punkte)

- $\hat{\beta}_1 = -167,352$
- `Eine Frau verdient c.p. im Mittel monatlich 167,352€ weniger als ein Mann. [1P]`
- `Der Koeffizient ist statistisch signifikant auf dem 5%-Niveau. [1P]`

b) Interpretieren Sie das R^2 der Schätzung und berechnen Sie das korrigierte Bestimmtheitsmaß $\bar{R}^2$. (3 Punkte)

- `Das Modell erklärt 13,6 % der Variation im monatlichen Einkommen. [1P]`
- $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$
  $= 1 - (1 - 0,136)\frac{535}{530} = 0,128$ `[2P]`

c) Welchen Wert würde der Koeffizient von *experience* (β_4) annehmen, wenn man Berufserfahrung statt in Jahren in 10 Jahren misst? (2 Punkte)

- Umskalierung einer unabhängigen Variable *experience*
- Experience in 10 Jahren: $\widetilde{experience} = \frac{experience}{10}$ [1P]
- $\widehat{income}_i = ... + \hat{\beta}_4 \cdot 10 \cdot \frac{experience_i}{10}$
- $\tilde{\beta}_4 = 10 \cdot 31,454 = 314,54$ [1P]

d) Berechnen Sie den marginalen Effekt der Berufserfahrung für eine Person mit 15 Jahren Bildung. Erläutern Sie das Ergebnis. (2 Punkte)

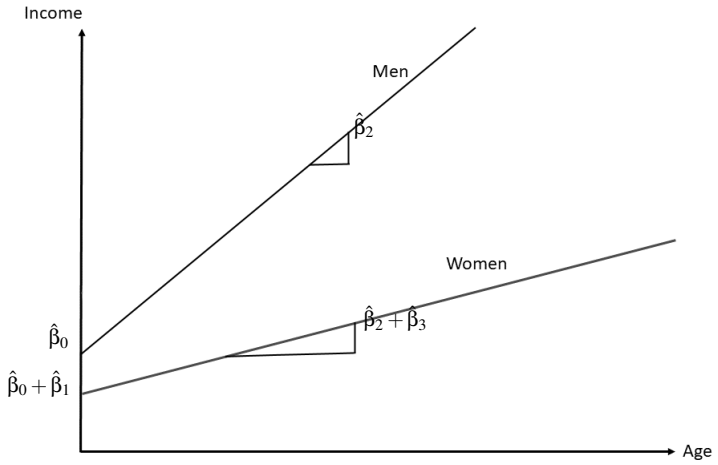
- Partielle Ableitung: $\frac{\Delta \widehat{income}}{\Delta experience} = \hat{\beta}_4 + \hat{\beta}_5 \cdot educ_i = 31,454 + 12,955 \cdot 15 = 225,779$ [1P]
- Für eine Person mit 15 Bildungsjahren steigt das monatliche Einkommen c.p. im Mittel um 225,779 €, wenn die Berufserfahrung um 1 Jahr ansteigt. [1P]

e) Sie vermuten, dass der Alterseffekt vom Geschlecht abhängt. Sie schätzen folgendes Modell:

$$income_i = \beta_0 + \beta_1 female_i + \beta_2 age_i + \beta_3 female_i \cdot age_i + u_i$$

Skizzieren Sie graphisch den Zusammenhang zwischen *age* und *income* für Frauen und Männer. Beschriften Sie die Achsen, die Achsenabschnitte und die Steigungen der Geraden. *Hinweis:* $\hat{\beta}_0 > 0,\ \hat{\beta}_1 < 0,\ \hat{\beta}_2 > 0,\ \hat{\beta}_3 < 0$. (4 Punkte)

- Jeweils 0.5 Punkte auf die Achsen. 0.5 Punkte auf den unterschiedlichen Achsenabschnitt. Je 1 Punkt auf die Steigung.
- Mit den Angaben sind andere Verläufe auch möglich.



f) Sie vermuten, dass der Einfluss des Alters auf das monatliche Einkommen nicht linear verläuft. Sie nehmen daher das quadrierte Alter mit in Ihr Modell auf und schätzen folgendes Modell:

$$income_i = \beta_0 + \beta_1 female_i + \beta_2 age_i + \beta_3 age^2{}_i + \beta_4 educ_i + \beta_5 experience_i + \beta_6 experience \cdot educ_i + u_i$$

Berechnen Sie das Alter, welches das geschätzte monatliche Einkommen maximiert.
Hinweis: $\hat{\beta}_2 = 20,512$ *und* $\hat{\beta}_3 = -0,249$. (2 Punkte)

- $\frac{\Delta \widehat{income}}{\Delta age} = \hat{\beta}_2 + 2 \cdot \hat{\beta}_3 \cdot age_i = 0$ [0.5P]
- $\Leftrightarrow 2 \cdot \hat{\beta}_3 \cdot age_i = -\hat{\beta}_2$ [0.5P]
- $\Leftrightarrow age_i = \frac{-\hat{\beta}_2}{2 \cdot \hat{\beta}_3} = \frac{-20{,}512}{2 \cdot (-0{,}249)} = 41{,}189$ [1P]
- Mit ca. 41 Jahren wird das geschätze monatliche Einkommen maximiert.

g) Statt des Modells, welches das Alter in Jahren enthält, entscheiden Sie sich für ein *neues* Modell. Sie bilden insgesamt 5 Dummy Variablen für einzelne Alterskategorien und nehmen alle in Ihr neues Modell mit auf. Welches Problem entsteht hierbei? Wie lässt es sich lösen? (3 Punkte)

- Es kommt zu perfekter Multikollinearität (MLR.3 ist verletzt). Der KQ-Schätzer ist nicht mehr bestimmbar (dummy variable trap). [2P]
- Lösung: Modell ohne Konstante schätzen oder eine Dummy-Variable weg lassen. [1P]

h) Anstatt des monatlichen Bruttoeinkommens als abhängige Variable wählen Sie nun eine Dummy-Variable als abhängige Variable. Nennen Sie zwei Nachteile des linearen Wahrscheinlichkeitsmodells. (2 Punkte)

- Es ist möglich, dass vorhergesagte Werte außerhalb des $(0,1)$ Intervalls liegen. [1P]
- Es ist oft unplausibel, dass einzelne Variablen über ihren gesamten Wertebereich linear mit der abhängigen Variable zusammen hängen. [1P]
- Oder: Das Modell ist heteroskedastisch. Unverzerrtheit bleibt erhalten, aber übliche Standardfehler fehlerhaft und t- und F-Test nicht anwendbar. [1P]

Aufgabe 3: **[22 Punkte]**

Sie interessieren sich dafür, ob die Schulklassengröße einen Einfluss auf die Löhne im Erwachsenenalter hat. Sie verfügen über einen US-amerikanischen Datensatz mit den folgenden Variablen für 764 Schüler:

$lohn_i$	Stundenlohn von Person i im Alter 27 (in US$)
$groß_i$	=1, wenn Person i in einer großen Klasse war ($>$ 17 SchülerInnen); =0 sonst
$frau_i$	=1, wenn Frau; =0 sonst
$eltern_i$	Gesamteinkommen der Eltern zum Zeitpunkt der Einschulung von Person i (in 1000 US$)

Sie schätzen folgendes lineares Regressionsmodell mit SPSS:

$$log(lohn_i) = \beta_0 + \beta_1 groß_i + \beta_2 frau_i + \beta_3 eltern_i + u_i$$

Koeffizienten[a]

Modell	Nicht standardisierte Koeffizienten			
	RegressionskoeffizientB	Standardfehler	T	Signifikanz
$(Konstante)$	6,013	0,088	68,32	0,000
$groß$	-0,065	0,016	???	???
$frau$	-0,362	0,063	???	???
$eltern$	0,118	0,029	4,07	0,000

a. Abhängige Variable: $log(lohn)$

Das R^2 der Schätzung beträgt $0,234$.
Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie den geschätzten Koeffizienten von *eltern* sowohl inhaltlich als auch statistisch. (2 Punkte)

- Ein Anstieg des Einkommens der Eltern bei Einschulung um 1000 US$ ist c.p. im Mittel mit einem um ca. 11,8% höherem Stundenlohn assoziiert. [1P]
- Der Koeffizient ist auf dem 1%-Signifikanzniveau statistisch signifikant von Null verschieden. [1P]

b) Berechnen und interpretieren Sie inhaltlich den genauen (!) Effekt der Variable *frau* auf den Stundenlohn. (2 Punkte)

- $\hat{\beta}_2 = -0,362$
- Der genaue Effekt beträgt $e^{\hat{\beta}_2} - 1 = -0,304$. [1P]
- Frauen haben c.p. im Mittel ein um $-30,4$ % niedrigeres Einkommen pro Stunde als Männer. [1P]

c) Testen Sie auf dem 5%-Signifikanzniveau, ob die Koeffizienten der Variablen *groß* und $frau$ gemeinsam signifikant sind. Bei einer erneuten Schätzung des Modells ohne diese beiden Variablen erhalten Sie ein R^2 von 0,186. Geben Sie Null- und Alternativhypothese, Freiheitsgrade, Teststatistik, kritischen Wert und Testentscheidung an. (5 Punkte)

- H_0 : $\beta_3 = \beta_4 = 0$ [0,5P] und H_1 : mindestens ein Parameter $\neq 0$. [0,5P]
- Freiheitsgrade (aus unrestringiertem Modell): $n-k-1 = 764-3-1 = 760$ [0,5P] und $q = 2$. [0,5P]
- Teststatistik: $F_{empirisch} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)} = \frac{(0,234-0,186)/2}{(1-0,234)/(764-3-1)} = \frac{0,024}{0,001} = 24$ ($23,812$ ohne Zwischenrunden). [1P]
- Kritischer Wert: $F_{kritisch} = F_{0;05;2;760} = 3,00$. [1P]
- Testentscheidung: Da $F_{empirisch} > F_{kritisch}$ wird die Nullhypothese auf dem 5%-Niveau verworfen. [1P]
- Die Variablen tragen gemeinsam signifikant zum Erklärungsgehalt bei.

d) Berechnen und interpretieren Sie das 99%-Konfidenzintervall für den geschätzten Koeffizienten der Variable $frau$. Gehen Sie darauf ein, ob der Koeffizient statistisch signifikant von Null verschieden ist. (4 Punkte)

- t-Wert in Tabelle ablesen: 2,576 (df=760, 1-α /2 = 0,995). [0,5P]
- Obere Grenze: $-0,362+2,576*0,063=-0,200$. [1P]
- Untere Grenze: $-0,362-2,576*0,063=-0,542$. [1P]
- (Das 99%-Konfidenzintervall des Koeffizienten von *frau* lautet: [-0,542; -0,200]).
- Interpretation: Mit wiederholten Stichproben liegt das wahre β_{frau} in 99% der Fälle im auf diese Weise berechneten Konfidenzintervall. [0,5P]
- Da der Wert 0 nicht im 99%-Konfidenzintervall enthalten ist, ist der Koeffizient auf dem 1%-Niveau statistisch signifikant von Null verschieden. [1P]

e) Testen Sie, ob die Variable *groß* einen signifikant negativen Zusammenhang mit dem späteren Einkommen hat. Geben Sie ein konkretes Testverfahren, Null- und Alternativhypothese, Teststatistik, Freiheitsgrade, kritischen Wert und Ihre Testentscheidung für das 5%-Signifikanzniveau an. (5 Punkte)

- Testverfahren: Einseitiger, linksseitiger t-Test. [1P]
- Hypothesen: $H_0:\beta_1\geq 0,\ H_1:\beta_1<0$. [1P]
- Teststatistik: $t=\dfrac{\hat{\beta}_1-0}{se(\hat{\beta}_1)}=\dfrac{-0,065-0}{0,016}=-4,063$. [1P]
- Freiheitsgrade: $n-k-1=764-3-1=760$. [0,5P]
- Kritischer Wert auf dem 5%-Signifikanzniveau: $-c=t_{\alpha;n-k-1}=t_{0,05;760}=-1,645$. [0,5P]
- Testentscheidung: Da $t_{empirisch} = -4,063 < -1,645 = -c$ wird die Nullhypothese auf dem 5%-Niveau verworfen. [0,5P]
- Die Variable *groß* hat einen auf dem 5%-Niveau negativen und statistisch signifikanten Zusammenhang mit dem späteren Einkommen. [0,5P]

f) Wie muss ein einfaches lineares Modell spezifiziert sein, damit der geschätzte Steigungsparameter i) eine Semielastizität und ii) eine Elastizität angibt? (2 Punkte)

- i) Wenn nur die abhängige und nicht die unabhängige Variable logarithmiert ist. [1P]
- ii) Wenn sowohl die abängige als auch die unabhängige Variable logarithmiert ist. 1P]

g) Welchen Effekt hat Heteroskedastie auf Unverzerrtheit und Effizienz des KQ-Schätzers? (2 Punkte)

- Heteroskedastie hat keinen Einfluss auf die Unverzerrtheit der KQ-Schätzers. [1P]
- Im Fall von Heteroskedastie ist der KQ-Schätzers nicht mehr der effizienteste unter allen linearen Schätzern. Die KQ-Standardfehler sind falsch. [1P]