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

Bachelorprüfung WS 2022 - MUSTERLÖSUNG

Fach: Data Science: Ökonometrie

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: **[19 Punkte]**

Sie möchten die Jahresgehälter der Basketballspieler in der NBA prognostizieren. Folgende Variablen stehen Ihnen aus der Saison 2021/22 für 160 Basketballspieler zur Verfügung:

$salary_i$	= Jahresgehalt von Spieler i in Mio. US-Dollar
PPG_i	= Durchschnittliche Anzahl erzielter Punkte pro Spiel von Spieler i
age_i	= das Alter von Spieler i in Jahren
MPG_i	= Durchschnittliche Anzahl gespielter Minuten pro Spiel von Spieler i
APG_i	= Durchschnittliche Anzahl Vorlagen pro Spiel von Spieler i
PPG_age_i	= Interaktion zwischen PPG und age von Spieler i

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.32379   10.36696   3.118  0.00217
PPG          0.70837    0.82471   0.859  0.39170
age         -1.36718    0.42020  -3.254  0.00140
PPG_age      0.02941    0.03236   0.909  0.36477
---

Residual standard error: 9.107 on 156 degrees of freedom
Multiple R-squared:  0.5533, Adjusted R-squared:  0.5447
```

Sie schätzen folgendes Regressionsmodell (1) und erhalten obenstehenden Output:

$$salary_i = \beta_0 + \beta_1 \cdot PPG_i + \beta_2 \cdot age_i + \beta_3 \cdot PPG\_age_i + u_i \tag{1}$$

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Luka ist 24 Jahre alt. Er erzielt durchschnittlich 36,8 Punkte. Berechnen Sie sein erwartetes Jahresgehalt. (2 Punkte)

- $32,324 + 0,708 \cdot 36,8 - 1,367 \cdot 24 + 0,029 \cdot 24 \cdot 36,8 \approx 51,183$
- Luka hat ein erwartetes Jahresgehalt von ca. 51,183 Mio. Dollar.

b) Leiten Sie den marginalen Effekt der Variable *PPG* auf das Jahresgehalt her. Berechnen Sie den marginalen Effekt von *PPG* für Luka aus der Aufgabe a) und interpretieren Sie das Ergebnis. (3 Punkte)

- Ableitung: $\frac{\Delta \widehat{salary_i}}{\Delta PPG_i} = \hat{\beta}_1 + \hat{\beta}_3 \cdot age_i$
- Einsetzen: $\frac{\Delta \widehat{salary_i}}{\Delta PPG_i} = 0,708 + 0,029 \cdot 24 \approx 1,404$
- Steigt die durchschnittliche Punktezahl pro Spiel um einen Punkt, so steigt das erwartete Jahresgehalt von Luka c.p. i.M. um 1,404 Mio. Dollar.

c) Leiten Sie den marginalen Effekt von Alter her und berechnen Sie auf dieser Basis, bei welchem Wert von *PPG* das Jahresgehalt maximal ist. (4 Punkte)

- Ableitung: $\frac{\widehat{\Delta salary_i}}{\Delta age_i} = \hat{\beta}_2 + \hat{\beta}_3 \cdot PPG_i$
- =0 setzen: $\hat{\beta}_2 + \hat{\beta}_3 \cdot PPG_i = 0$
- Nach PPG auflösen: $PPG^* = -\frac{\hat{\beta}_2}{\hat{\beta}_3}$
- Einsetzen und berechnen: $PPG^* = -\frac{-1{,}367}{0{,}0294} = 46{,}497$

d) Eine Kommilitonin besteht darauf, das Modell durch die Variablen MPG_i und APG_i zu ergänzen. Sie nehmen diese zwei Variablen ins Model auf und schätzen erneut (Modell 2).

$$salary_i = \beta_0 + \beta_1 \cdot PPG_i + \beta_2 \cdot age_i + \beta_3 \cdot PPG\_age_i + \beta_4 \cdot MPG_i + \beta_5 \cdot APG_i + u_i \tag{2}$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.73583   10.08398   2.552  0.01168
PPG          2.20419    0.89165   2.472  0.01452
age         -0.85799    0.42764  -2.006  0.04657
MPG         -0.47286    0.16301  -2.901  0.00427
APG          1.33969    0.42612   3.144  0.00200
PPG_age     -0.02170    0.03364  -0.645  0.51981
---
Residual standard error: 8.736 on 154 degrees of freedom
Multiple R-squared:  0.5943,        Adjusted R-squared:  0.5811
```

Sie möchten nun testen, ob die geschätzten Parameter $\hat{\beta}_4$ und $\hat{\beta}_5$ gemeinsam statistisch signifikant sind. Benennen Sie das Testverfahren und formulieren Sie Null- und Alternativhypothese, berechnen Sie die Teststatistik und bestimmen Sie den kritischen Wert. Kann die Nullhypothese auf dem 1%-Signifikanzniveau abgelehnt werden? (6 Punkte)

- Testverfahren: F-Test auf gemeinsame Signifikanz
- H_0: $\hat{\beta}_4 = \hat{\beta}_5 = 0$ und H_1: mindestens ein Parameter $\neq 0$
- Teststatistik: $F = \frac{[(R_U^2 - R_R^2)/q]}{(1-R_U^2)/(n-k-1)} = \frac{(0{,}594-0{,}553)/2}{(1-0{,}594)/(160-5-1)} \approx 7{,}776$
- Kritischer Wert c: $c = F_{(0{,}01;2;160-5-1)} = F_{(0{,}01;2;154)} = 4{,}61$
- Testentscheidung: $F = 7{,}776 > 4{,}61 = c$. Die Nullhypothese kann auf dem 1%-Signifikanzniveau verworfen werden. Die beiden Parameter sind gemeinsam statistisch signifikant von 0 verschieden.

e) Interpretieren Sie $\hat{\beta}_5$ statistisch und inhaltlich. (2 Punkte)

- Steigt die durchschnittliche Anzahl der Vorlagen von Spieler i um eins, so steigt das erwartete Jahresgehalt c.p. i.M. um 1,34 Mio. US. Dollar.
- Der Koeffizient ist Signifikant am 1% Signifikanzniveau.

f) Welchen Wert würde $\hat{\beta}_0$ im Modell (2) annehmen, wenn $salary_i$ nicht in Mio. US-Dollars sondern in 1000 US-Dollars gemessen worden wäre? (2 Punkte)

- $\frac{1}{1000} \cdot 1000 \cdot \widehat{salary_i} = \hat{\beta}_0 + \hat{\beta}_1 \cdot PPG_i + \hat{\beta}_2 \cdot age_i + \hat{\beta}_3 \cdot PPG\_age_i + \hat{\beta}_4 \cdot MPG_i + \hat{\beta}_5 \cdot APG_i$
- $\underbrace{1000 \cdot \widehat{salary_i}}_{\widehat{salary_i}} = \underbrace{\hat{\beta}_0 \cdot 1000}_{\tilde{\beta}_0} + \ldots$
- $\Rightarrow \tilde{\beta}_0 = 25,736 \cdot 1000 = 25736$

Aufgabe 2: **[18 Punkte]**

Sie interessieren sich für das Sparverhalten von Individuen. Sie verfügen über Querschnittsdaten von 830 Personen aus dem Jahr 2020. Ihr Datensatz enthält folgenden Informationen:

$savings_i$	= jährliche Ersparnis von Person i in Euro
$income_i$	= Jahreseinkommen von Person i in Euro
$educ_i$	= Ausbildung von Person i in Jahren
age_i	= Alter von Person i in Jahren

Sie schätzen folgendes Regressionsmodell:

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

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

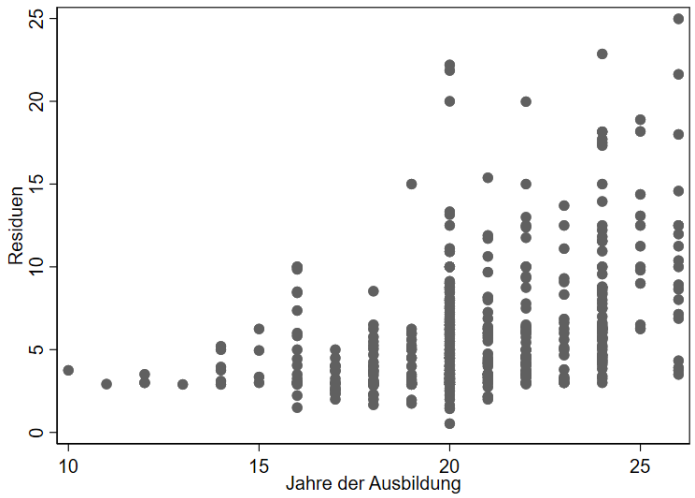
a) Interpretieren Sie $\hat{\beta}_1$ inhaltlich und statistisch. *Hinweis:* $\hat{\beta}_1 = 0,117$ und $p = 0,001$ (2 Punkte)

- Steigt das Jahreseinkommen einer Person um 1 Euro, so steigt ceteris paribus im Mittel die jährliche Ersparnis um 0,117 Euro.
- Der Koeffizient ist statistisch signifikant auf dem 1%-Niveau.

b) Wie lautet Ihre inhaltliche Interpretation von $\hat{\beta}_1$, wenn Sie statt *savings* die logarthmierte Ersparnis $ln(savings)$ als abhängige Variable nutzen. *Hinweis:* $\beta_1^{neu} = 0,005$. (1 Punkt)

- Steigt das Jahreseinkommen einer Person um 1 Euro, so steigt ceteris paribus im Mittel die jährliche Ersparnis um 0,5%.

c) Ihr Kommilitone rät Ihnen, die Variable *birth* mit in das Modell aufzunehmen. Diese Variable enthält Informationen über das Geburtsjahr von Person i. Sollten Sie die Variable *birth* in das Modell aufnehmen? Welche Konsequenzen hätte die Aufnahme für Ihre Schätzung? (4 Punkte)



- Nein, man sollte die Variable *birth* nicht in das Modell aufnehmen.
- Die Aufnahme führt zum Problem der perfekten Multikollinearität, da bei Querschnittsdaten die Variable *birth* perfekt mit der Variable *age* korreliert.
- Problem: Bei perfekter Multikollinearität ist der KQ-Schätzer nicht mehr berechenbar.

d) Der Graph zeigt die Verteilung der Residuen für *educ*. Ist die Gauß-Markov Annahme MLR.5 hier verletzt? Erläutern Sie kurz Ihre Antwort. Welche Konsequenz hat dies für den KQ-Schätzer (3 Punkte)

- Die Gauß-Markov Annahme MLR.5 ist im vorliegenden Beispiel verletzt. Die Residuen sind heteroskedastisch in Bezug auf die Bildungsjahre. Die Varianz der Störterme variiert mit den Bildungsjahren.
- Dies hat zur Folge, dass der Parameterschätzer zwar unverzerrt, die Standardfehler aber falsch berechnet sind.

e) Ihr Kommilitone meint, Sie hätten eine relevante Variable vergessen. Erklären Sie das Problem ausgelassener Variablen und welche Konsequenzen dies prinzipiell für Schätzungen hat. Erläutern Sie ein Beispiel für das Problem ausgelassener Variablen in der vorliegenden Schätzung. Begründen Sie kurz Ihre Antwort. (4 Punkte)

- Das Problem ausgelassener Variablen liegt vor, wenn im Fehlerterm relevante, erklärende Variablen enthalten sind, die sowohl mit der zu erklärenden Variable , als auch mit den unabhängigen Variablen korrelieren .
- Die Konsequenz ausgelassener Variablen ist eine verzerrte Schätzung der Parameterschätzer.
- Im vorliegenden Fall könnte z.B. die Anzahl der Kinder (*kids*) eine ausgelassene, relevante Variable darstellen. *kids* korreliert (vermutlich negativ) mit der jährlichen Ersparnis. Zudem korreliert *kids* (vermutlich positiv) mit dem Alter *age*.
  Alternative Antworten möglich.

f) Berechnen Sie das R^2 der Schätzung und interpretieren Sie dieses. *Hinweis:* Das adjustierte Bestimmtheitsmaß beträgt 0,094. Nennen Sie einen Nachteil des R^2. (4 Punkte)

- $\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k-1} \leftrightarrow R^2 = 1-(1-\bar{R}^2)\frac{n-k-1}{n-1} = 1-(1-0{,}094)\frac{830-3-1}{830-1} = 0{,}097$
- Das Modell erklärt 9,7% der Variation in der abhängigen Variable (= der jährlichen Ersparnis)
- Ein Nachteil des R^2 ist, dass es nicht fallen kann, wenn die Anzahl der erklärenden Variablen steigt. Dies gilt unabhängig davon, ob die zusätzlichen Variablen zum Erklärungsgehalt des Modells beitragen.

Aufgabe 3: **[23 Punkte]**

Sie interessieren sich für die Determinanten von Wohneigentum. Es steht Ihnen dafür ein Datensatz mit 5299 Personen zur Verfügung. Sie beobachten den folgenden Satz von Variablen:

$eigenheim_i$	=1, bei Eigentum von Haus oder Wohnung; =0, sonst
$alter_i$	= Alter von Person i in Jahren
$hheink_i$	= Haushaltsjahreseinkommen von Person i in Tausend Euro
$bildung_i$	= Bildungsjahre von Person i

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output:

$$eigenheim_i = \beta_0 + \beta_1 \cdot alter_i + \beta_2 \cdot hheink_1 + \beta_3 \cdot bildung_i + u_i$$

```
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  -1.842     0.457    -4.031   0.000
alter         0.044     0.004    11.000   0.000
hheink        0.014     0.001     ???      ???
bildung      -0.017     0.002    -8.500   0.000
---

Residual standard error: 0.461 on 5295 degrees of freedom
Multiple R-squared:  0.09813, Adjusted R-squared:  0.09762
F-statistic:   192 on 3 and 5295 DF,  p-value: < 2.2e-16
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

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

- Mit jedem zusätzlichem Lebensjahr steigt die Wahrscheinlichkeit für Haus- bzw. Wohnungseigentum um 4,4 Prozentpunkte c.p. im Mittel.
- Der Koeffizient ist auf dem 1%-Niveau statistisch signifikant.

b) Berechnen und interpretieren Sie das 99%-Konfidenzintervall für den geschätzten Koeffizienten der Variable *hheink*. Gehen Sie darauf ein, ob der Koeffizient statistisch signifikant von Null verschieden ist. (5 Punkte)

- t-Wert in Tabelle ablesen: 2,576 (df=5255, 1-α /2 = 0,995).
- Obere Grenze: $0{,}014 + 2{,}576 * 0{,}001 = 0{,}017$.

- Untere Grenze: $0{,}014 - 2{,}576 * 0{,}001 = 0{,}011$.
- (Das 99%-Konfidenzintervall des Koeffizienten von *hheink* lautet: [0,011; 0,017]).
- Interpretation: Mit wiederholten Stichproben liegt das wahre β_{hheink} in 99% der Fälle im auf diese Weise berechneten Konfidenzintervall.
- Da der Wert 0 nicht im 99%-Konfidenzintervall enthalten ist, ist der Koeffizient auf dem 1%-Niveau statistisch signifikant von Null verschieden.

c) Wie hoch ist die Wahrscheinlichkeit, dass eine 31 Jahre alte Person, mit einem Haushaltsjahreseinkommen von 55.000 Euro und 12 Jahren Schulbildung Wohneigentum besitzt? (3 Punkte)

- $\widehat{eigenheim}_i = -1{,}842 + 0{,}044 \cdot 31 + 0{,}014 \cdot 55 - 0{,}017 \cdot 12 = 0{,}088$
- Die vorhergesagte Wahrscheinlickeit für Wohneigentum für eine Person mit diesen Eigenschaften liegt bei 8,8 Prozent.

d) Sie überlegen, dass β_2 unterschätzt sein könnte, da die Höhe des Einkommens im Zusammenhang mit der regionalen Bevölkerungdichte steht. Unter welchen Bedingungen würde sich die Vermutung bestätigen? (5 Punkte)

- $Cov(Bevölkerungsdichte, hheink) > 0$, d.h. die Bevölkerungsdichte korreliert positiv mit der Höhe des Einkommens.
- $Cov(Bevölkerungsdichte, eigenheim) < 0$, d.h. die Bevölkerungsdichte korreliert negativ mit der Wahrscheinlichkeit von Wohneigentum.
- (Andere Lösungen denkbar)

e) Nennen Sie zwei Nachteile des linearen Wahrscheinlichkeitsmodells. (2 Punkte)

- Die vorhergesagten Werte für die abhängige Variable können Werte außerhalb (0,1) annehmen.
- Die Varianz der Störterme ist nicht konstant, es liegt Heteroskedastie vor. Alternativ: Die Schätzung ist nicht effizient.
- Die Störterme sind nicht normalverteilt, daher sind t- und F-Tests nicht exakt gültig.
- Ein Punkt pro korrekter Antwort, maximal zwei Punkte.

f) Sie vermuten, dass sich die Parameter für ländliche und städtische Gebiete unterscheiden und führen einen Chow-Test auf Strukturbruch am 1%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung an. Wird Ihre Vermutung bestätigt? (6 Punkte)
Hinweise: $SSR_{pooled} = 320$, $SSR_1 = 158$ (für *ländlich=0*), $SSR_2 = 160$ (für *ländlich=1*)

- Hypothesen:
  H_0 : Es besteht kein Strukturbruch zwischen städtischen und ländlichen Regionen, oder $\beta_{j,g=1} = \beta_{j,g=2}$ mit $j = 0, \ldots, k$
  H_1 : Es besteht ein Strukturbruch zwischen städtischen und ländlichen Regionen, oder $\beta_{j,g=1} \neq \beta_{j,g=2}$ mit $j = 0, \ldots, k$

- Teststatistik: $F_{Chow} = \frac{\frac{SSR_p - (SSR_1 + SSR_2)}{k+1}}{\frac{(SSR_1 + SSR_2)}{(n-2(k+1))}} = \frac{\frac{320-(158+160)}{3+1}}{\frac{158+160}{5299-2(3+1)}} = \frac{0,5}{0,06} = 8,3$
- kritischer Wert: $F_{0,01;4;5291} = 3,32$
  (krit. Wert nicht tabelliert für df=5291 $\rightarrow df = \infty$)
- Testentscheidung: Da $F_{Chow} = 8,3 > 3,32 = c$ kann die Nullhypothese auf dem 1%-Niveau verworfen werden. Das Modell unterscheidet sich signifikant für städtische und ländliche Regionen.