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

Bachelorprüfung SS 2022 - MUSTERLÖSUNG

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 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: **[20 Punkte]**

Sie interessieren sich für die Auswirkungen der Gewerkschaftsmitgliedschaft auf die Löhne. Sie verfügen über Querschnittsdaten zu 1260 Personen, die an der Beschäftigungserhebung in dem Jahr 1977 in den USA teilgenommen haben.
Sie beobachten den folgenden Satz von Variablen:

$hourlywage_i$	= Stundenlohn in US-Dollar
$educ_i$	= Jahre der Ausbildung
$female_i$	= 1, wenn weiblich, 0, wenn männlich
$exper_i$	= Jahre der Berufserfahrung
$union_i$	= 1, wenn Mitglied einer Gewerkschaft, 0, wenn nicht

Sie schätzen folgendes Regressionsmodell (= Modell 1) und erhalten untenstehenden Output:

$$hourlywage_i = \beta_0 + \beta_1 \cdot female_i + \beta_2 \cdot union_i + u_i$$

```
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)    7.165      0.176    40.67      0.000
female        -3.007      0.263   -11.43      0.000
union          0.668      0.281     2.38      0.018

Multiple R-squared:  0.102,        Adjusted R-squared:  0.101
```

Der Mittelwert für die abhängige Variable beträgt 6,310.

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Interpretieren Sie den Koeffizientenschätzer von *union* inhaltlich. Ist der Effekt statistisch signifikant? (2 Punkte)

- Wenn die Arbeitnehmer:Innen Mitglied der Gewerkschaft sind, erhalten sie c.p. i.M. einen um 0,668 US-Dollar höheren Stundenlohn.
- Der Koeffizient ist statistisch auf dem 5% Niveau signifikant.

b) Berechnen Sie das 95%-Konfidenzintervall für den Koeffizienten von *union*. Zeigen Sie Ihren Rechenweg und interpretieren Sie das Konfidenzintervall. (4 Punkte)

- t-Wert in Tabelle ablesen: 1,960 (df=1257, 1-a /2 = 0,975).
- Obere Grenze: 0,668+1,96*0,281 = 1,21876 (gerundet: 1,219).
- Untere Grenze: 0,668-1,96*0,281 = 0,11724 (gerundet: 0,117).
- (Das 95%-Konfidenzintervall des Koeffizienten lautet: [0,117; 1,219]).
- Interpretation: Bei wiederholter Stichprobenziehung liegt in 95% der Fälle der wahre Parameter innerhalb der auf diese Weise bestimmten Konfidenzintervalle.

c) Ist der Koeffizientenschätzer von *female* ökonomisch signifikant? Begründen Sie Ihre Antwort. (2 Punkte)

- Da der Koeffizientenschätzer von female ca. 48% des Mittelwerts der abhängigen Variablen beträgt, ist der Effekt ökonomisch signifikant.
- Alternative Antwortmöglichkeiten möglich.

Sie erweitern das Regressionsmodell um die Variablen $exper_i$ und edu_i (= Modell 2):

$$hourlywage_i = \beta_0 + \beta_1 \cdot female_i + \beta_2 \cdot union_i + \beta_3 \cdot educ_i + \beta_4 \cdot exper_i + u_i$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.391	???	-0.578	0.563
female	???	0.257	-9.721	0.000
union	0.784	0.268	2.922	0.004
educ	0.465	0.046	???	0.000
exper	0.083	0.010	7.996	???

Multiple R-squared: 0.193, Adjusted R-squared: ???

d) Berechnen Sie das angepasste $\bar{R}^2$ von Modell 2 und vergleichen Sie es mit Modell 1. Begründen Sie Ihre Beobachtung. (2 Punkte)

- $\bar{R}^2$ = 1- (1-R^2)[(n-1)/(n-k-1)]
- $\bar{R}^2$ = 1- (1-0,193)[(1260-1)/1260-4-1)]
- = 1- 0,807 · [(1259)/1255)]
- = 0,190427888
- ≈ 0,190
- Das angepasste $\bar{R}^2$ steigt im erweiterten Modell an, da der (korrigierte) Erklärungsgehalt im Modell stärker steigt als die Anpassung der Freiheitsgrade.

e) Berechnen Sie se($\hat{\beta}_0$), $\hat{\beta}_1$ und t($\hat{\beta}_3$). Ist der geschätzte Koeffizient für β_4 auf dem 1%-Niveau statistisch signifikant? Begründen Sie kurz. (4 Punkte)

- $se(\hat{\beta}_0) = \frac{\hat{\beta}_0}{t(\hat{\beta}_0)}$= -0,391/-0,578 ≈ 0,676
- $\hat{\beta}_1 = t(\hat{\beta}_1) \cdot se(\hat{\beta}_1)$ = -9,721 · 0,257 ≈ -2,498
- $t(\hat{\beta}_3) = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)}$ = 0,465/0,046 ≈ 10,109
- Da t($\hat{\beta}_4$) = 7,996 > 2,576 (kritischer Wert der t-Verteilung bei 1,255 Freiheitsgraden) ist der geschätzte Koeffizient statistisch signifikant auf dem 1%-Niveau.
  Auch richtig, wenn der t-Wert nicht abgelesen, sondern berechnet wurde: Da t($\hat{\beta}_4$) = 8,3 > 2,576 (kritischer Wert der t-Verteilung bei 1,255 Freiheitsgraden) ist der geschätzte Koeffizient statistisch signifikant auf dem 1%-Niveau.

f) Sie möchten testen, ob die geschätzten Parameter $\hat{\beta}_3$ und $\hat{\beta}_4$ in Modell 2 gemeinsam 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 10%-Signifikanzniveau abgelehnt werden? (6 Punkte)

- Testverfahren: F-Test auf gemeinsame Signifikanz
- H_0: $\hat{\beta}_3 = \hat{\beta}_4 = 0$ und H_1: mindestens ein Parameter $\neq 0$
- Teststatistik: $F = [(R_U^2 - R_R^2)/q]/[(1-R_U^2)/(n-k-1)$ F = [(0,193- 0,102)/2]/ [1 - 0,193) / (1260 - 4-1) $\approx 70,759$
- Kritischer Wert c: c = $F_{(0,1;2;1260-4-1)} = F_{(0,1;2;1255)}$ = 2,3
- Testentscheidung: F = 70,759 > 2,3 = c. Die Nullhypothese kann auf dem 10%-Signifikanzniveau verworfen werden. Die beiden Parameter sind gemeinsam statistisch signifikant von 0 verschieden.

Aufgabe 2: **[10 Punkte]**

a) Sie möchten den Zusammenhang zwischen dem Mietpreis für ein WG-Zimmer in Euro (y_i) und der Nähe zur Uni in km (x_i) mittels des Modells $y_i = \beta_0 + \beta_1 x_i + u_i$ schätzen. Hierzu befragen Sie 3 Student:Innen zu (x,y) und erhalten die Beobachtungen (0, 450), (4, 400) und (2, 350). Berechnen Sie $\hat{\beta}_0$ und $\hat{\beta}_1$. (6 Punkte)

- $\bar{x} = (0+4+2)/3 = 2$
- $\bar{y} = (450+400+350)/3 = 400$
- $Var(x) = 1/(n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2$
- $Var(x) = 1/(3-1)[(0-2)^2 + (4-2)^2 + (2-2)^2]$
- $Var(x) = 1/2[4+4]$
- $Var(x) = 4$
- $Cov(x) = 1/(n-1)\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- $Cov(x) = 1/(3-1)[(0-2)(450-400) + (4-2)(400-400) + (2-2)(350-400)]$
- $Cov(x) = 1/2[(-2)(50)]$
- $Cov(x) = -50$
- $\hat{\beta}_1$= Cov(x, y)/ Var(x)
- $\hat{\beta}_1$ = -12,5
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$
- $\hat{\beta}_0$ = 425

b) Erläutern Sie kurz zwei der Annahmen, die erfüllt werden müssen, damit der Kleinste-Quadrate (KQ)-Schätzer ein unverzerrter Schätzer ist. (2 Punkte)

- 1 Punkt pro Nennung einer der folgenden Annahmen:
- MLR.1: Modell in der Grundgesamtheit(linearer Zusammenhang).
  $\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- MLR.2: Zufallsstichprobe, die dem Modell der Grundgesamtheit folgt (repräsentativ).

- MLR.3: Keine der erklärenden Variablen ist konstant; keine der erklärenden Variablen ist Linearkombination der anderen erklärenden Variablen (perfekte Multikollinearität).
- MLR.4: $E(u|x1,x2,\ \ldots\ ,xk) = 0 \Rightarrow$ mittlere bedingte Unabhängigkeit des Störterms und der erklärenden Variablen (vgl. ceteris paribus Annahme).

c) Erklären Sie den Typ 1- Fehler sowie den Typ 2-Fehler bei Hypothesentests. (2 Punkte)

- Bei Hypothesentests kann man zwei Arten von Fehlern machen:
- Verwirft man H_0, obwohl H_0 zutrifft, spricht man vom Typ 1-Fehler.
- Verwirft man H_0 nicht, obwohl H_0 falsch ist, spricht man vom Typ 2-Fehler.
- *Alternativ:* Alpha- und Beta-Fehler.

Aufgabe 3: **[16 Punkte]**

Sie interessieren sich dafür, was den Wunsch eines Jobwechsels beeinflusst und untersuchen dies anhand von Querschnittsdaten aus dem Jahr 2019. Insgesamt 787 Erwebstätige wurden unter anderem dazu befragt, ob sie den Wunsch haben, ihren jetzigen Arbeitgeber zu wechseln.

Im Datensatz enthalten sind folgende Variablen:

job_change_i	=1, wenn Jobwechseln gewünscht ist, 0, wenn nicht
age_i	= Alter von Person i in Jahren
sat_emp_i	= Zufriedenheit mit jetzigem Arbeitgeber auf einer Skala von 1 (absolut unzufrieden) bis 10 (voll zufrieden)
$childr_i$	= Anzahl der Kinder von Person i
$female_i$	=1, wenn weiblich, 0, wenn männlich

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output:

$$job\_change_i = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot sat\_emp_i + \beta_3 \cdot childr_i + u_i$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.419	0.090	4.655	0.000
age	-0.013	0.003	-4.802	0.000
sat_emp	0.024	0.009	2.739	0.006
childr	0.057	0.034	???	???

Multiple R-squared: 0.062, Adjusted R-squared: 0.059

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

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

- Mit jedem zusätzlichem Lebensjahr sinkt die Wahrscheinlichkeit für den Wunsch eines Jobswechsels c.p. im Mittel um 1,3 Prozentpunkte.
- Der Koeffizient ist auf dem 1%-Niveau statistisch signifikant.

b) Wie hoch ist die Wahrscheinlichkeit, dass eine 45 Jahre alte Person ohne Kinder, die eine 10 auf der Skala der Zufriedenheit mit dem jetzigen Arbeitgeber angegeben hat, den Wunsch eines Jobwechsels äußert? (3 Punkte)

- $job\_change_i = 0,419 - 0,013 \cdot 45 + 0,024 \cdot 10 + 0,057 \cdot 0 = 0,074$
- Die vorhergesagte Wahrscheinlickeit der Äußerung des Wunsches eines Jobwechsels für einen Person mit diesen Eigenschaften liegt bei 7,4 Prozent.

c) Nennen Sie zwei Schwächen des linearen Wahrscheinlichkeitsmodells. (2 Punkte)

- Es ist möglich, dass vorhergesagte Werte außerhalb des Intervalls (0,1) liegen.
- 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.
- (Maximal zwei Punkte, andere Lösungen möglich.)

d) Sie vermuten, dass sich die Parameter des Modells für Frauen und Männer unterscheiden und führen ein Chow-Test auf Strukturbruch am 5%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung an. (6 Punkte)
Hinweise: $SSR_{pooled} = 749$, $SSR_1 = 340$ (für $female = 0$), $SSR_2 = 396$ (für $female = 1$)

- Hypothesen:
  H_0 : Es besteht kein Strukturbruch zwischen Männern und Frauen, oder $\beta_{j,g=1} = \beta_{j,g=2}$ mit $j = 0,\ldots,k$
  H_1 : Es besteht ein Strukturbruch zwischen Männern und Frauen, oder $\beta_{j,g=1} \neq \beta_{j,g=2}$ mit $j = 0,\ldots,k$
- Teststatistik: $F_{Chow} = \frac{\frac{SSR_p - (SSR_1 + SSR_2)}{k+1}}{\frac{(SSR_1 + SSR_2)}{(n-2(k+1))}} = \frac{\frac{749-(340+396)}{3+1}}{\frac{340+396}{787-2(3+1)}} = \frac{3,25}{0,945} = 3,440$
- kritischer Wert: $F_{0,05;4;779} = 2,37$
  (krit. Wert nicht tabelliert für df=779 $\rightarrow df = \infty$)
- Testentscheidung: Da $F_{Chow} = 3,44 > 2,37 = c$ kann die Nullhypothese auf dem 5%-Niveau verworfen werden. Das Modell unterscheidet sich für Frauen und Männer.

e) Erläutern Sie kurz eine alternative Vorgehensweise, um zu testen, ob sich die Parameter des Modells für Männer und Frauen signifikant unterscheiden. Geben Sie zusätzlich sowohl einen Vorteil als auch einen Nachteil gegenüber der Testung in Teilaufgabe d) an. (3 Punkte)

- Berechnung eines vollständig interagierten Modells und einem anschließendem F-Test.

- Ein Vorteil ist, dass an den p-Werten der Interaktionsterme direkt abgelesen werden kann, welche Regressionsparameter signifikant unterschiedlich für verschiedene Gruppen sind.
- Ein Nachteil ist die aufwändige Schätzung, wenn das Modell viele erklärende Variablen enthält.
- Andere Antworten möglich.

Aufgabe 4: **[14 Punkte]**

Sie interessieren sich weiterhin dafür, was die Zufriedenheit mit dem jetzigen Arbeigeber beeinflusst und untersuchen dies anhand des in Aufgabe 3 beschriebenen Datensatzes. Ihnen stehen zwei zusätzliche Variablen zur Verfügung.

Im Datensatz enthalten sind nun folgende Variablen:

sat_emp_i	= Zufriedenheit mit jetzigem Arbeitgeber auf einer Skala von 1 (absolut unzufrieden) bis 10 (voll zufrieden)
age_i	= Alter von Person i in Jahren
$female_i$	=1, wenn weiblich, 0, wenn männlich
$seniority_i$	Anzahl an Jahren beim jetzigen Arbeitgeber
ln_inc_i	Logarithmiertes monatliches Bruttoeinkommen in Euro

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output für Modell I:

$$sat\_emp_i = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot female_i + \beta_3 \cdot seniority_i + u_i$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.334	0.710	13.144	0.000
age	0.073	0.016	4.713	0.000
female	-0.118	0.304	-0.387	0.699
seniority	-0.023	0.019	-1.206	0.228

Multiple R-squared: 0.045, Adjusted R-squared: 0.040

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Sie möchten testen, ob der Koeffizient der Variable *age* auf dem 10%-Niveau statistisch signifikant größer als 0,02 ist. Führen Sie einen entsprechenden Test durch. Geben Sie Testverfahren, Null- und Alternativhypothese, Teststatistik, Freiheitsgrade, kritischen Wert und Ihre Testentscheidung an. (6,5 Punkte)

- Testverfahren: einseitiger t-Test
- Hypothesen: $H_0 : \beta_1 <= 0{,}02,\ H_1 : \beta_1 > 0{,}02$
- Teststatistik: $t = \dfrac{\hat{\beta}_1 - 0{,}2}{se(\hat{\beta}_1)} = \dfrac{0{,}073 - 0{,}02}{0{,}016} = 3{,}313$
- Freiheitsgrade: $n-k-1 = 787-3-1 = 783$
- Kritischer Wert c 10%-Signifikanzniveau: $c = t_{\alpha;n-k-1} = t_{0{,}10;783} = 1{,}282$
- Testentscheidung: Da $t_{empirisch} = 3{,}31 > 1{,}282 = c$ kann die Nullhypothese auf dem 10%-Niveau verworfen werden, der Koeffizient *age* ist statistisch signifikant größer als 0,02.

Sie nehmen die Höhe des monatlichen Bruttolohnes als logarithmierte Variable zusätzlich in Ihr Modell mit auf und erhalten folgenden Output für Modell II:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.562	1.592	-2.238	0.026
age	0.057	0.015	3.838	0.000
female	0.918	0.308	2.980	0.003
seniority	-0.052	0.018	-2.896	0.004
ln_inc	1.631	0.183	8.923	0.000

Multiple R-squared: 0.169, Adjusted R-squared: 0.163

b) Interpretieren Sie den geschätzten Koeffizienten von *ln_inc* inhaltlich und statistisch. (2 Punkte)

- Ein Anstieg des monatlichen Bruttoeinkommes um 1% steigert die Zufriedenheit mit dem derzeitigen Arbeitgeber c.p. im Durchschnitt um 0,016 Skalenpunkte.
- Der Koeffizient ist auf dem 1%-Niveau statistisch signifikant von Null verschieden.

c) Ihre Kommilitonin vermutet, dass die Annahme $E(u|age_i, female_i, seniority_i) = 0$ in Model I ohne die zusätzliche Aufnahme von *ln_inc* verletzt sein könnte. Erläutern Sie diese Annahme und beschreiben Sie, wie sich die Verletzung dieser Annahme auf die Interpretation der Koeffizienten allgemein auswirkt, gehen Sie dabei zusätzlich genauer auf die Auswirkungen auf den Koeffizienten von $female$ ein. (3,5 Punkte)

- Die Annahme $E(u|age_i, female_i, seniority_i) = 0$ beschreibt die mittlere bedingte Unabhängigkeit des Störterms von den erklärenden Variablen.
- Bei Verletzung dieser Annahme werden die Koeffizienten verzerrt geschätzt.
- Der Koeffizient von $female$ wurde in Modell I im Vergleich zu Modell II unterschätzt.

d) Welche Auswirkung hätte die zusätzliche Aufnahme der Variable des Geburtsjahres in Ihr Modell für die Schätzung? Begründen Sie Ihre Aussage mit der relevanten Gauss-Markov Annahme. (2 Punkte)

- Bei Aufnahme der Variable des Geburtsjahres wäre der KQ-Schätzer nicht mehr berechenbar.
- Alternativ: Die Konstante könnte nicht mitgeschätzt werden.
- Es würde perfekte Multikollinearität vorliegen (MLR.3).