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

Bachelorprüfung SoSe 2023 - 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: **[21 Punkte]**

Der Datensatz enthält Informationen zu Umsätzen und drei Werbemedien (Youtube, Instagram und Zeitung) von 200 Unternehmen:

$umsatz_i$	= Umsatz von Unternehmen i in Tausend Euro
$youtube_i$	= Werbebudget für Youtube von Unternehmen i in Tausend Euro
$instagram_i$	= Werbebudget für Instagram von Unternehmen i in Tausend Euro
$newspaper_i$	= Werbebudget für Zeitungen von Unternehmen i in Tausend Euro

Hinweis: Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

Sie wollen die Verkäufe auf der Grundlage des ausgegebenen Werbebudgets vorhersagen und entscheiden sich zwischen den folgenden Regressionsmodellen (1-4):

$$umsatz_i = \beta_0 + \beta_1 \cdot youtube_i + u_i \tag{1}$$

$$umsatz_i = \beta_0 + \beta_1 \cdot instagram_i + \beta_2 \cdot instagram_i^2 + u_i \tag{2}$$

$$umsatz_i = \beta_0 + \beta_1 \cdot youtube_i + \beta_2 \cdot instagram_i + \beta_3 \cdot newspaper_i + u_i \tag{3}$$

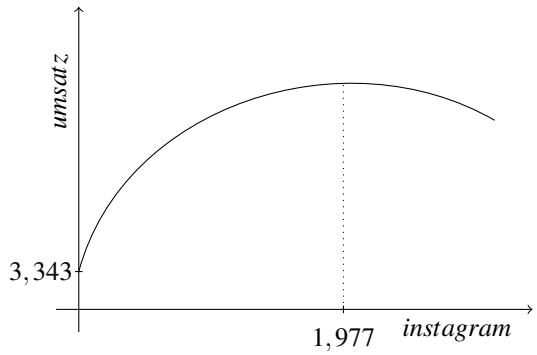
$$log(umsatz_i) = \beta_0 + \beta_1 \cdot youtube_i + \beta_2 \cdot instagram_i + \beta_3 \cdot newspaper_i + u_i \tag{4}$$

a) Sind Modelle 1 und 2 genestet? Begründen Sie kurz Ihre Antwort. (2 Punkte)

- Nein, Modelle 1 und 2 sind nicht genestet.
- Genestete Modelle entstehen durch Vereinfachung (Nullsetzen einiger Parameter) des unrestringierten Modells (hier Modell 2). Modell 1 ist nicht in Modell 2 enthalten.
- Andere Antworten möglich.

b) Geben Sie den marginalen Effekt von *instagram* im Modell 2 an. Bei welcher Höhe des Werbebudgets für Instagram wird der Umsatz maximiert, wenn $\hat{\beta}_0 = 3,343$, $\hat{\beta}_1 = 0,174$ und $\hat{\beta}_2 = -0,044$? Skizzieren Sie anschließend eine mögliche Beziehung zwischen *instagram* und *umsatz* basierend auf diesen Koeffizientenschätzwerten. Markieren Sie alle bekannten Werte auf den Achsen. (7 Punkte)

- Ableitung: $\frac{\Delta \widehat{umsatz_i}}{\Delta instagram_i} = \hat{\beta}_1 + 2 \cdot \hat{\beta}_2 \cdot instagram_i$
- =0 setzen: $\hat{\beta}_1 + 2 \cdot \hat{\beta}_2 \cdot instagram_i = 0$
- Nach *instagram* auflösen: $instagram^* = -\frac{\hat{\beta}_1}{2 \cdot \hat{\beta}_2}$
- Einsetzen und berechnen: $instagram^* = -\frac{0,174}{2 \cdot (-0,044)} = 1,977$
- Jeweils 1P für die Achsenbeschriftung und 1P für den umgekehrt u-förmigen Verlauf
- Jeweils 0.5P für die Eintragung der Werte auf den Achsen.



c) Sie schätzen das Modell 3 in R und erhalten folgenden Output:

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)     3.527      0.374   9.422     0.000
youtube         0.045      0.001  32.809     0.000
instagram       0.189      0.009  21.893     0.000
newspaper      -0.001      0.006  -0.177      0.86

Residual standard error: 2.023 on 196 degrees of freedom
Multiple R-squared:  0.897,        Adjusted R-squared:  0.896
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 0.000
```

Sie möchten nun testen, ob die Variablen *instagram* und *newspaper* gemeinsam signifikant zu der Erklärung vom Umsatz beitragen. Benennen Sie das Testverfahren und formulieren Sie die Null- und Alternativhypothese, berechnen Sie die Teststatistik und bestimmen Sie den kritischen Wert. Kann die Nullhypothese auf dem 5%-Signifikanzniveau abgelehnt werden? (6 Punkte)
Hinweis: Das R^2 aus der Schätzung von Modell 1 ist $0,885$.

- Testverfahren: F-Test auf gemeinsame Signifikanz
- H_0: $\beta_2 = \beta_3 = 0$ und H_1: mindestens ein Parameter $\neq 0$
- Teststatistik: $F = \frac{[(R_U^2 - R_R^2)/q]}{(1-R_U^2)/(n-k-1)} = \frac{(0{,}897-0{,}885)/2}{(1-0{,}897)/(200-3-1)} \approx 11{,}417$
- Kritischer Wert c: $c = F_{(0{,}05;2;200-3-1)} = F_{(0{,}05;2;200-3-1)} = 3{,}00$
- Testentscheidung: $F = 11{,}417 > 3{,}00 = c$. Die Nullhypothese kann auf dem 5%-Signifikanzniveau verworfen werden. Die beiden Parameter sind gemeinsam statistisch signifikant von 0 verschieden.

d) Sie schätzen nun das Modell 4 und erhalten folgenden Output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.921      0.035  55.574    0.000
youtube         0.003      0.001  23.735    0.000
instagram       0.009      0.001  12.365    0.000
newspaper       0.000      0.001   0.545    0.587

Residual standard error: 0.1868 on 196 degrees of freedom
Multiple R-squared:  0.7998,        Adjusted R-squared:  0.7967
F-statistic: 260.9 on 3 and 196 DF,  p-value: <0.000
```

Interpretieren Sie den Koeffizientenschätzer von β_2 inhaltlich und statistisch. (2 Punkte)

- Steigt das Werbebudget für instagram von Unternehmen i um 1000 Euro, so steigt der erwartete Umsatz c.p. i.M. um $0,009 \cdot 100 = 0,9\%$.
- Der Koeffizient ist signifikant am 1% Signifikanzniveau.

e) Nehmen Sie an, dass der Umsatz in dem Modell 4 nicht in Tausend Euro, sondern in Euro geschätzt wird. Wie verändert sich der Schätzer der Koeffizienten β_0 und β_1? (4 Punkte)

- $log(\underbrace{umsatz_i \cdot 1000}_{=\widetilde{umsatz_i}} \cdot \frac{1}{1000}) = \beta_0 + \ldots$
- $log(\widetilde{umsatz_i}) - log(1000) = \beta_0 + \ldots$
- $log(\widetilde{umsatz_i}) = \underbrace{\beta_0 + log(1000)}_{=\tilde{\beta}_0} + \ldots$
- β_1 verändert sich nicht.

Aufgabe 2: **[15,5 Punkte]**

Sie interessieren sich für die Determinanten von Wohnungseinbrüchen. Es steht Ihnen dafür ein Datensatz mit 2756 Wohnbezirken aus bayerischen Großstädten zur Verfügung. Sie beobachten die folgenden Variablen:

$einbruch_i$	=1, bei Einbruch in den letzten 6 Wochen; =0, sonst
$einw_i$	= Anzahl der EinwohnerInnen im Wohnbezirk i in Hundert
alo_i	= Arbeitslosenquote im Wohnbezirk i (0,00 - 1,00)
$hheink_i$	= Durchschnittliches Haushaltsnettoeinkommen im Wohnbezirk i in Euro

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output:

$$einbruch_i = \beta_0 + \beta_1 \cdot einw_i + \beta_2 \cdot alo_1 + \beta_3 \cdot hheink_i + u_i$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    35.028      2.045  17.129    0.000
einw            0.016      0.006   4.404    0.000
alo             0.121       ???    3.043    0.002
hheink         -0.014      0.008  -1.620    0.105
---

Residual standard error: 0.8548 on 2721 degrees of freedom
Multiple R-squared:  0.01093, Adjusted R-squared:  0.009839
F-statistic: 10.02 on 3 and 2721 DF,  p-value: 0.0000
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

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

- Pro Hundert zusätzliche Einwohner im Wohnbezirk steigt die Wahrscheinlichkeit eines Wohnungseinbruches in den letzten 6 Wochen c.p. im Mittel um 1,6 Prozentpunkte.
- Der Koeffizient ist auf dem 1%-Niveau statistisch signifikant von Null verschieden.

b) Berechnen Sie den Standardfehler von $\hat{\beta}_2$. (1 Punkt)

- $se(\hat{\beta}_2) = \hat{\beta}_2 / t(\hat{\beta}_2) = 0,121/3,043 = 0,040$.

c) Wie hoch ist die Wahrscheinlichkeit eines Wohnungseinbruches in den letzten sechs Wochen für einen Wohnbezirk mit 1000 Einwohnern, einer Arbeitslosenquote von 0,03 und einem durchschnittlichen Nettohaushaltseinkommen von 2500 Euro? (3 Punkte)

- $\widehat{einbruch}_i = 35,028 + 0,016 \cdot 10 + 0,121 \cdot 0,03 - 0,014 \cdot 2500 = 0,192$
- Die vorhergesagte Wahrscheinlickeit für einen Wohnungseinbruch in den letzten sechs Wochen liegt bei 19,2 Prozent.

d) Sie vermuten, dass die Anzahl von in den Wohnbezirken lebenden Frauen ($einw\_fem_i$) und Männern ($einw\_male_i$) einen unterschiedlichen Einfluss haben könnten. Sie haben beide Maße vorliegen und nehmen diese zusätzlich in ihr Modell mit auf. Welches Problem tritt bei der Schätzung auf? Wie lässt es sich lösen? Begründen Sie Ihre Antwort knapp. (4 Punkte)

- Es besteht das Problem der perfekten Multikollinearität.
- Die Variablen $einw_i$, $einw\_fem_i$ und $einw\_male_i$ sind linear abhängig; es ist nicht möglich, alle drei Variablen gleichzeitig in die Regression aufzunehmen.
- Lösung: Entweder $einw_i$, oder $einw\_fem_i$ bzw. $einw\_male_i$ aus dem Modell nehmen. (Eine der Möglichkeiten reicht aus)

e) Beschreiben Sie allgemein, welche Bedingungen zutreffen müssen, damit ein Problem ausgelassener Variablen vorliegt. Welche Folgen hätte dies für den geschätzten Koeffizienten? Diskutieren Sie, ob es sich bei $\hat{\beta}_1$ um den kausalen Effekt der Einwohnerzahl auf die Einbruchswahrscheinlichkeit handelt. (3,5 Punkte)

- Damit ein Problem ausgelassener Variablen vorliegt, muss die ausgelassene Variable (i) sowohl mit der abhängigen Variable, als auch (ii) mit einer erklärenden Variable korrelieren.
- Folge: Verzerrung des geschätzten Koeffizienten, bildet nicht mehr den kausalen Effekt ab.
- Im vorliegenden Fall ist es plausibel, dass z.B. der Anteil an Einfamilienhäusern im Wohnbezirk sowohl mit der Einwohnerzahl, als auch mit der Einbruchswahrscheinlichkeit korreliert. Die hier geschätzte Korrelation zwischen der Einwohnerzahl und der Einbruchswahrscheinlichkeit bildet vermutlich keinen kausalen Effekt ab.
- Andere Antworten möglich.

f) 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.

Aufgabe 3: **[23,5 Punkte]**

Sie interessieren sich weiterhin für die Determinanten von Wohnungseinbrüchen. Es steht Ihnen dafür der gleiche Datensatz wie aus Aufgabe 2 mit 2756 Wohnbezirken zur Verfügung, jedoch verwenden Sie die Häufigkeit der Wohnungseinbrüche im letzten Jahr. Sie beobachten den folgenden Satz von Variablen:

$anzahleinbr_i$	= Anzahl der Einbrüche im Jahr in Wohnbezirk i
$einw_i$	= Anzahl der EinwohnerInnen im Wohnbezirk i in Hundert
alo_i	= Arbeitslosenquote im Wohnbezirk i (0,00-1,00)
$hheink_i$	= Durchschnittliches Haushaltsnettoeinkommen im Wohnbezirk i in Euro

Sie schätzen das folgende lineare Regressionsmodell und erhalten untenstehenden Output:

$$anzahleinbr_i = \beta_0 + \beta_1 \cdot einw_i + \beta_2 \cdot alo_i + \beta_3 \cdot hheink_i + u_i$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.211      0.010  22.078    0.000
einw_i          0.004      0.002   2.397    0.017
alo_i           0.011      0.004   2.718    0.007
hheink_i        0.009      0.003     ???      ???
---

Residual standard error: 0.4113 on 2721 degrees of freedom
Multiple R-squared:  ???,        Adjusted R-squared:  0.007
F-statistic: 7.039 on 3 and 2721 DF,  p-value: 0.0001
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

a) Berechnen Sie das R^2 der Schätzung und interpretieren Sie dieses. *Hinweis:* Das adjustierte Bestimmtheitsmaß beträgt 0,055. Nennen Sie einen Nachteil des R^2. (4 Punkte)

- $\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-k-1} \leftrightarrow R^2 = 1-(1-\bar{R}^2)\frac{n-k-1}{n-1} = 1-(1-0{,}055)\frac{2756-3-1}{2756-1} = 0{,}056$
- Das Modell erklärt 5,6% der Variation in der abhängigen Variable (= der jährlichen Anzahl der Einbrüche)
- 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.

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=2721, 1-α /2 = 0,995).
- Obere Grenze: $0{,}009 + 2{,}576 * 0{,}003 = 0{,}017$.
- Untere Grenze: $0{,}009 - 2{,}576 * 0{,}003 = 0{,}001$.
- (Das 99%-Konfidenzintervall des Koeffizienten von *hheink* lautet: [0,001; 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) Sie vermuten, dass sich das Modell zwischen Wohnbezirken mit direkter Autobahnanbindung und jenen ohne direkte Autobahnanbindung unterscheidet.

i. Erläutern Sie kurz das Vorgehen und die Entscheidungslogik des Chow-Tests auf Strukturbruch, den Sie mittels der Variable $autobahn_i$ durchführen können. (3 Punkte)

- Alle erklärenden Variablen aus dem Modell werden mit der Variable $autobahn_i$ interagiert. Diese Interaktionsterme und die Variable $autobahn_i$ werden als zusätzliche erklärende Variablen in Modell I aufgenommen (vollständig interagiertes Modell).
- Anschließend wird ein F-Test auf gemeinsame Signifikanz der zusätzlich aufgenommenen Variablen durchgeführt.
- Ergibt der F-Test Hinweise auf gemeinsame Signifikanz, so wird davon ausgegangen, dass ein Strukturbruch vorliegt und sich das Modell für Wohnbezirken mit bzw. ohne direkte Autobahnanbindung unterscheidet.

(Andere Antworten möglich.)

ii. Führen Sie einen Chow-Test auf Strukturbruch am 10%-Niveau durch. Geben Sie Hypothesen, Teststatistik, kritischen Wert und Ihre Testentscheidung für das 10%-Signifikanzniveau an. (6 Punkte)
Hinweise: $SSR_{pooled} = 3159,781$, $SSR_1 = 1603,674$ (für $autobahn = 0$), $SSR_2 = 1550,100$ (für $autobahn = 1$)

- Hypothesen: $H_0 : \beta_{1,0} = \beta_{2,0},\ \beta_{1,1} = \beta_{2,1},\ \beta_{1,2} = \beta_{2,2},\ \beta_{1,3} = \beta_{2,3}$. H_1 : mindestens ein β_j mit $j = 0,1,2,3$ unterscheidet sich zwischen den Gruppen $g = 1$ und $g = 2$.
- Teststatistik: $F_{Chow} = \frac{\frac{SSR_p-(SSR_1+SSR_2)}{k+1}}{\frac{(SSR_1+SSR_2)}{(n-2(k+1))}} = \frac{\frac{3159,781-(1603,674+1550,100)}{3+1}}{\frac{1603,674+1550,100}{2756-2(3+1)}} = \frac{1,502}{1,148} = 1,308$
- kritischer Wert: $F_{4,2748,10\%} = 1,94$
  (krit. Wert nicht tabelliert für df=2748, nächstgelegener niedrigerer Wert: df=120)
- Testentscheidung: Da $F_{Chow} = 1,307 < 1,94 = c$ kann die Nullhypothese auf dem 10%-Niveau nicht verworfen werden. Das Modell unterscheidet sich für Wohnbezirke ohne ($g = 1$) und mit direktert Autobahnanbidung ($g = 2$) nicht signifikant.

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

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

e) Benennen Sie den Fehler in folgendem R-Code: (1,5 Punkte)

```
filter(autobahn=1) %>%

ggplot(data = einbruchstat, aes(x=hheink, y=anzahleinbr)) +

geom_point()
```

- Die Filterfunktion ist nicht korrekt, korrekt wäre autobahn==1

f) Was wird mit nachfolgender Code-Zeile ausgeführt? (1 Punkt)

```
fit <-  lm(anzahleinbr ~ einw, data=einbruchstat)
```

- In der Code-Zeile wird die Variable Anzahl der EinwohnerInnen auf die Variable Anzahl der Einbrüche im Wohnbezirk regressiert.

g) Sie erweitern Ihren Code. Welche Wirkung hat das auf die Ausgabe? (1 Punkt)

```
fit1 <- lm(anzahleinbr ~ einw , data=einbruchstat, subset=female!=1)
```

- Die Berechnung schließt die Gruppe der Frauen aus. Die lineare Regression wird nur für die Gruppe der Männer ausgeführt.