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

Masterprüfung SS 2016 - MUSTERLÖSUNG

Fach: Ökonometrie

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

Vorbemerkungen:

Anzahl der Aufgaben:	Die Klausur besteht aus 5 Aufgaben, die alle bearbeitet werden müssen. **Es wird nur der Lösungsbogen eingesammelt.**
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: **[13 Punkte]**

Sie interessieren sich dafür, wie Unterschiede in der durchschnittlichen Kalorienaufnahme zwischen Ländern erklärt werden können. Ihnen sind folgende Daten über 107 Länder gegeben:

$kcal_i$	Durchschnittliche tägliche Kalorienaufnahme in Land *i* in Kilokalorien (kcal)
ln_GDPpc_i	Logarithmiertes Bruttoinlandsprodukt von Land *i* in Dollar pro Kopf
$grain_i$	Anteil der Getreideanbaufläche an der gesamten Anbaufläche von Land *i* in Prozent (0 - 100)
$grain_i$ˆ2	$(grain_i)^2$
$trocken_i$	Dummy-Variable, =1, wenn in Land *i* trockenes Klima herrscht, =0 sonst
$tropisch_i$	Dummy-Variable, =1, wenn in Land *i* tropisches Klima herrscht, =0 sonst
$mediterran_i$	Dummy-Variable, =1, wenn in Land *i* mediterranes Klima herrscht, =0 sonst
$arktisch_i$	Dummy-Variable, =1, wenn in Land *i* arktisches Klima herrscht, =0 sonst

Es wird folgendes Regressionsmodell aufgestellt und anschließend mit Stata geschätzt:

$$kcal_i = \beta_1 + \beta_2 ln\_GDPpc_i + \beta_3 grain_i + \beta_4 grain_i\hat{}2 + \beta_5 tropisch_i + \beta_6 mediterran_i + \beta_7 arktisch_i + \varepsilon_i$$

```
      Source |       SS       df       MS              Number of obs =     107
-------------+------------------------------           F(  6,   100) =   27.49
       Model |  20372778.3     6  3395463.06           Prob > F      =  0.0000
    Residual |  12351372.6   100  123513.726           R-squared     =  0.6226
-------------+------------------------------           Adj R-squared =  ??????
       Total |  32724150.9   106  308718.405           Root MSE      =  351.45

------------------------------------------------------------------------------
        kcal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ln_GDPpc |   273.1558   25.85852    10.56   0.000     221.8532    324.4584
       grain |   13.38649   6.017523     2.22   0.028     1.447898    25.32509
     grain^2 |  -.1112228    .097891    -1.14   0.259    -.3054356    .0829901
    tropisch |   66.20466   110.7554     0.60   0.551     -153.531    285.9403
  mediterran |   117.6956    139.717     0.84   0.402    -159.4989    394.8902
    arktisch |    185.273   108.4746     1.71   0.091    -29.93748    400.4834
       _cons |   115.1619   238.1231     0.48   0.630    -357.2675    587.5913
------------------------------------------------------------------------------
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

1.1 Interpretieren Sie den geschätzten Koeffizienten b_2 inhaltlich und statistisch. [2 Punkte]

- Eine 1% Erhöhung des Pro Kopf Einkommens führt c.p. zu einem durchschnittlichen Anstieg der Kalorienaufnahme um 2,732 Kilokalorien.
- Der Koeffizient ist statistisch signifikant auf dem 1% Niveau.

1.2 Bestimmen Sie den marginalen Effekt des Anteils der Getreideanbaufläche. Bei welchem Anteil ist die Kalorienaufnahme maximal? [2,5 Punkte]

- Marginaler Effekt allgemein:

$$\frac{\Delta E(kcal_i)}{\Delta grain_i} = b_3 + 2 \cdot b_4 grain_i = 13,386 + 2 \cdot (-0,111) \cdot grain_i$$

- Maximierender Anteil:

$$13,386 + 2 \cdot (-0,111) \cdot grain_i \stackrel{!}{=} 0 \Leftrightarrow grain_i = \frac{-13,386}{2 \cdot (-0,111)} \Leftrightarrow grain_i = 60,297$$

Die Kalorienaufnahme wird maximiert bei einem Anteil von Getreide an der Gesamtanbaufläche von 60,297%.

1.3 Das Land Balkonien hat ein Bruttoinlandsprodukt pro Kopf von 20.000 Dollar, es herrscht mediterranes Klima und auf 20% der Anbaufläche wird Getreide angebaut. Wie hoch ist die erwartete tägliche Kalorienaufnahme eines Bewohners von Balkonien? [2,5 Punkte]

$$\begin{aligned}\widehat{kcal_i} &= b_1 + b_2 ln\_GDPpc_i + b_3 grain_i + b_4 (grain_i)^2 + b_5 tropisch_i + b_6 mediterran_i + b_7 arktisch_i \\ &= 115,162 + 273,156 \cdot ln(20.000) + 13,386 \cdot 20 - 0,111 \cdot (20)^2 + 66,205 \cdot 0 + 117,686 \cdot 1 + 185,273 \cdot 0 \\ &= 3161,365\end{aligned}$$

1.4 Ermitteln Sie das $\bar{R}^2$ (das angepasste R^2) der Regression. [2 Punkte]

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2/(N-K)}{\sum_{i=1}^{N} (y_i - \bar{y})^2/(N-1)} = 1 - \frac{SSR/(107-7)}{SST/(107-1)} = 1 - \frac{12351372,6/100}{32724150,9/106} = 0,600$$

1.5 Haben die Dummy-Variablen für das Klima gemeinsam einen signifikanten Einfluss auf die Kalorienaufnahme? Führen Sie einen geeigneten Test auf dem 5%-Signifikanzniveau durch. Geben Sie Null- und Alternativhypothese, Teststatistik, kritischen Wert und Testentscheidung an.
Hinweis: Das Bestimmtheitsmaß R^2 des restringierten Modells beträgt 0,609. [4 Punkte]

- **Hypothesen:** $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$; H_1 : mind. ein $\beta_j \neq 0$ mit $j = 5,6,7$
- **Teststatistik:**
$$F_{empirisch} = \frac{(R_1^2 - R_0^2)/J}{(1-R_1^2)/(N-K)} = \frac{(0,623 - 0,609)/3}{(1-0,623)/(107-7)} = 1,238$$
(Je nachdem an welcher Stelle überall auf die dritte Nachkommastelle gerundet wurde, kann das Ergebnis auch 1,184 oder 1,250 sein.)
- **Kritischer Wert:** $c = F_{kritisch} = F_{J;N-K;\alpha} = F_{3;100;0,05} = 2,70$
- **Testentscheidung:** Da $F_{empirisch} = 1,238 < 2,70 = F_{kritisch}$ kann die Nullhypothese nicht verworfen werden. Die Koeffizienten der Klima-Dummies sind gemeinsam nicht signifikant von Null verschieden auf dem 5%-Niveau.

Aufgabe 2: **[7 Punkte]**

Sie interessieren sich für den Einfluss der Fertilität auf die Erwerbstätigkeit von Frauen. Hierfür steht ihnen ein Datensatz über 753 Frauen mit folgenden Variablen zur Verfügung:

Job_i	Dummy-Variable, =1, wenn Frau i erwerbstätig ist, =0 sonst
$kidsl6_i$	Anzahl der Kinder unter 6 Jahren von Frau i
$kids618_i$	Anzahl der Kinder zwischen 6 und 18 Jahren von Frau i
$educ_i$	Bildung von Frau i in Jahren
age_i	Alter von Frau i in Jahren

Es wird folgendes Modell aufgestellt und mittels Logit geschätzt:

$$P(Job_i = 1|\mathbf{x}_i) = F(\beta_1 + \beta_2 kidsl6_i + \beta_3 kids618_i + \beta_4 educ_i + \beta_5 age_i)$$

```
Logistic regression                               Number of obs   =        753
                                                  LR chi2(4)      =      99.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -465.03978                       Pseudo R2       =     0.0968

------------------------------------------------------------------------------
         Job |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      kidsl6 |  -1.470007   .1949278    -7.54   0.000    -1.852058   -1.087956
     kids618 |  -.0940861   .0665954    -1.41   0.158    -.2246107    .0364385
        educ |   .1979224   .0373293     5.30   0.000     .1247583    .2710864
         age |   -.063381   .0124618    -5.09   0.000    -.0878057   -.0389562
       _cons |   1.036848   .7701192     1.35   0.178    -.4725582    2.546254
------------------------------------------------------------------------------
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

2.1 Berechnen Sie die Differenz in der Wahrscheinlichkeit erwerbstätig zu sein zwischen einer Frau mit einem 5-jährigen Kind und einer Frau mit einem 8-jährigen Kind. Beide haben einen Hauptschulabschluss (10 Jahre Bildung), sind 30 Jahre alt und haben keine weiteren Kinder. [5 Punkte]

- Wahrscheinlichkeit, dass die erste Frau erwerbstätig ist:

$$P(Job_i = 1|kidsl6 = 1, kids618 = 0, educ = 10, age = 30) = F(1{,}037 - 1{,}47 \cdot 1 - 0{,}094 \cdot 0 + 0{,}198 \cdot 10 - 0{,}063 \cdot 30)$$

$$= F(-0{,}343) = \frac{1}{1 + exp(-(-0{,}343))} = 0{,}415$$

(Das Ergebnis ohne in den Zwischenschritten zu runden ist 0,412.)

- Wahrscheinlichkeit, dass die zweite Frau erwerbstätig ist:

$$P(Job_i = 1|kidsl6 = 0, kids618 = 1, educ = 10, age = 30) = F(1{,}037 - 1{,}47 \cdot 0 - 0{,}094 \cdot 1 + 0{,}198 \cdot 10 - 0{,}063 \cdot 30)$$

$$= F(1{,}033) = \frac{1}{1 + exp(-(1{,}033))} = 0{,}737$$

(Das Ergebnis ohne in den Zwischenschritten zu runden ist 0,735.)

- Differenz: Die Wahrscheinlichkeit, dass die Frau mit dem 5-jährigen Kind erwerbstätig ist, ist um $0{,}735 - 0{,}412 = 0{,}323 \rightarrow 32{,}3$ Prozentpunkte geringer als die Wahrscheinlichkeit, dass die Frau mit dem 8-jährigen Kind erwerbstätig ist.

2.2 Es gibt zwei gebräuchliche Arten, den marginalen Effekt einer unabhängigen Variable bei einer Logit Schätzung zu berechnen. Benennen sie kurz in Worten die beiden Möglichkeiten. [2 Punkte]

i. Berechnung des marginalen Effekts am Mittelwert aller unabhängigen Variablen.

ii. Berechnung des Mittelwerts des marginalen Effekts über alle Beobachtungen.

Aufgabe 3: **[20 Punkte]**

Ein Verkehrsverbund interessiert sich für die Determinanten der Anzahl verkaufter U-Bahn-Tickets in seinem Gebiet. Es liegt ein Datensatz mit Informationen für 365 Tage des Jahres 2015 vor:

$ln\_Tickets_t$	Logarithmierte Anzahl verkaufter Einzelfahrscheine am Tag t
ln_Price_t	Logarithmierter Preis eines Einzelfahrscheines in € am Tag t
$Rain_t$	Niederschlagsmenge in Millimeter pro Quadratmeter am Tag t
$Speed_t$	Durchschnittliche Straßenverkehrsgeschwindigkeit in km/h am Tag t

Es wird folgendes Regressionsmodell aufgestellt und anschließend mit Stata geschätzt:

$$ln\_Tickets_t = \beta_1 + \beta_2 ln\_Price_t + \beta_3 Rain_t + \beta_4 Speed_t + \varepsilon_t$$

```
      Source |       SS       df       MS              Number of obs =     365
-------------+------------------------------           F(  3,   361) =   54.71
       Model |   31.240971     3  10.413657             Prob > F      =  0.0000
    Residual |   45.881003   361  0.1270942            R-squared     =  0.4051
-------------+------------------------------           Adj R-squared =  0.3968
       Total |   77.121974   364  0.2118736            Root MSE      =  .28725
```

```
------------------------------------------------------------------------------
   ln_Tickets |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    ln_Price |  -.6156393   .0897841    -6.86   0.000    -.7917348   -.4395438
        Rain |   .1298635   .0234291     5.43   0.000     .1757853    .0852105
       Speed |  -.0049900   .0008750    -5.70   0.000    -.0067070   -.0032740
       _cons |   0.518024   .0729190     7.10   0.000     0.375103    13.99696
------------------------------------------------------------------------------
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

3.1 Interpretieren Sie den geschätzten Koeffizienten b_2 inhaltlich und statistisch. [2 Punkte]

- Eine Erhöhung im Preis eines Einzelfahrscheines um 1% führt c.p. im Mittel zu einem Rückgang der verkauften Einzelfahrscheine um 0,615%.
- Der Koeffizient ist statistisch signifikant auf dem 1% Niveau. (da $p = 0{,}00 < 0{,}01$)

3.2 Der folgende Stata-Output enthält die Ergebnisse eines White-Tests auf Heteroskedastie. Führen Sie den Test auf einem Signifikanzniveau von 5% durch. Geben Sie Null- und Alternativhypothese sowie Hilfsregression an. Definieren Sie die abhängige Variable der Hilfsregression. Bestimmen Sie zudem Teststatistik, Freiheitsgrade, kritischen Wert und Testentscheidung. [7 Punkte]

```
White's test for Ho: ???
        against Ha: ???

        chi2(?)      =    21,14
        Prob > chi2  =    ??????
```

- Hypothesen: $H_0 : V(\varepsilon_t) = \sigma^2$ für alle t (Homoskedastie); $H_1 : V(\varepsilon_t) \neq V(\varepsilon_s)$ für mind. ein $t \neq s$ (Heteroskedastie)
- Hilfsregression: $e_t^2 = \alpha_1 + \alpha_2 ln\_Price_t + \alpha_3 ln\_Price_t^2 + \alpha_4 Rain_t + \alpha_5 Rain_t^2 + \alpha_6 Speed_t + \alpha_7 Speed_t^2 + \alpha_8 (ln\_Price_t \cdot Rain_t) + \alpha_9 (ln\_Price_t \cdot Speed_t) + \alpha_{10} (Rain_t \cdot Speed_t) + \nu_t$ [2P]
  wobei $e_t^2 = (y_t - x_t' b)^2$.
- Teststatistik: $\chi^2_{emp} = N \cdot R^2$, wobei N die Anzahl der Beobachtungen bezeichnet; R^2 ist das Bestimmtheitsmaß aus der Hilfsregression.
- Freiheitsgrade: J=9
- Kritischer Wert: $\chi^2_J = \chi^2_9 = 16,92$
- Da $\chi^2_{empirisch} = 21,14 > 16,92 = \chi^2_{kritisch}$ wird die Nullhypothese auf dem 5%-Niveau verworfen. Der Test weist auf Heteroskedastie in dem vorliegenden Modell hin.

3.3 Es sei $V(\varepsilon_t) = \sigma^2 \cdot Rain_t^3$. Zeigen Sie formal die aus einer generalized least squares (GLS)-Transformation resultierende Schätzgleichung, welche zu konstanter Störtermvarianz führt. Leiten Sie zudem die Varianz des Störterms im transformierten Modell her. [2,5 Punkte]

- GLS-Transformation: $\frac{ln\_Tickets_t}{Rain_t^{\frac{3}{2}}} = \frac{\beta_1}{Rain_t^{\frac{3}{2}}} + \frac{\beta_2 ln\_Price_t}{Rain_t^{\frac{3}{2}}} + \frac{\beta_3 Rain_t}{Rain_t^{\frac{3}{2}}} + \frac{\beta_4 Speed_t}{Rain_t^{\frac{3}{2}}} + \frac{\varepsilon_t}{Rain_t^{\frac{3}{2}}}$
- $V(\frac{\varepsilon_t}{Rain_t^{\frac{3}{2}}}) = \frac{1}{Rain_t^3} \cdot Var(\varepsilon_t) = \frac{1}{Rain_t^3} \cdot \sigma^2 \cdot Rain_t^3 = \sigma^2$

3.4 Erläutern Sie verbal den Begriff Autokorrelation. Welche Folgen hat unkorrigierte Autokorrelation für die geschätzten Parameter und deren Standardfehler? [2 Punkte]

- Unter Autokorrelation versteht man eine Situation, in der die Störterme im (linearen) Regressionsmodell korrelieren.
- Liegt Autokorrelation im Modell vor, so sind die Koeffizienten unverzerrt geschätzt (solange A1 und A2 gelten.)
- Die Standardfehler sind allerdings falsch berechnet. (Die Schätzung ist ineffizient.)

3.5 Führen Sie einen Breusch-Godfrey-Test auf Autokorrelation 1. Ordnung auf einem Signifikanzniveau von 1% durch. Geben Sie Null- und Alternativhypothese, Hilfsregression, Teststatistik, Freiheitsgrade, kritischen Wert und Testentscheidung an. Definieren Sie dabei formal Autokorrelation 1. Ordnung.
Hinweis: Das Bestimmtheitsmaß aus der Hilfsregression lautet: $R^2 = 0,02$. [6,5 Punkte]

- Autokorrelation 1.Ordnung: $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$
- Hypothesen: $H_0 : \rho = 0$; $H_1 : \rho \neq 0$
- Hilfsregression: $e_t = \alpha_1 + \alpha_2 e_{t-1} + \alpha_3 ln\_Price_t + \alpha_4 Rain_t + \alpha_5 Speed_t + \nu_t$
- Teststatistik: $\chi^2_{emp} = (T-1) \cdot R^2 = (365-1) \cdot 0,02 = 7,28$, wobei T die Anzahl der beobachteten Perioden bezeichnet; R^2 ist das Bestimmtheitsmaß aus der Hilfsregression.
- Freiheitsgrade: J=1
- Kritischer Wert: $\chi^2_J = \chi^2_1 = 6,635$
- Da $\chi^2_{empirisch} = 7,28 > 6,635 = \chi^2_{kritisch}$, wird die Nullhypothese auf dem 1%-Niveau verworfen. Der Test weist auf Autokorrelation 1. Ordnung in dem vorliegenden Modell hin.

Aufgabe 4: **[20 Punkte]**

Die Universitätsleitung beauftragt Sie, die Determinanten der Klausurleistungen von Erstsemesterstudierenden zu analysieren. Sie stellen dazu folgendes Regressionsmodell auf:

$$Punkte_i = \tilde{\beta}_1 + \tilde{\beta}_2 Anwesenheit_i + \gamma Motivation_i + \varepsilon_i \quad \text{(Modell I)}$$

wobei Sie die Variablen wie folgt kodieren:

$Punkte_i$	Durchschnittlich erreichte Punktzahl des Studierenden i in den Klausuren des 1. Semesters (0-100)
$Anwesenheit_i$	Anteil der besuchten Vorlesungen des Studierenden i in % (0-100)
$Motivation_i$	Motivation des Studierenden i für das Studium

Der Ihnen durch die Universitätsleistung zur Verfügung gestellte Datensatz mit Daten für 680 Studierende enthält jedoch keine Information über die Motivation der Studierenden, so dass Sie lediglich die Gleichung

$$Punkte_i = \beta_1 + \beta_2 Anwesenheit_i + \nu_i \quad \text{(Modell II)}$$

mit $\nu_i = \gamma Motivation_i + \varepsilon_i$ schätzen können.

4.1 Leiten Sie ausgehend von der Formel $b = (X'X)^{-1}X'(X\beta + U\gamma + \varepsilon)$ mit

$$X = \begin{pmatrix} 1 & Anwesenheit_1 \\ 1 & Anwesenheit_2 \\ \vdots & \vdots \\ 1 & Anwesenheit_{680} \end{pmatrix} \text{ und } U = \begin{pmatrix} Motivation_1 \\ Motivation_2 \\ \vdots \\ Motivation_{680} \end{pmatrix}$$

in Matrizenschreibweise formal her, unter welchen Bedingungen die Kleinstquadrateschätzung (KQ) in Modell II zu einem unverzerrten Schätzer für β_2 führt. Erläutern Sie die beiden Bedingungen knapp verbal. Machen Sie kenntlich, an welcher Stelle der Herleitung Sie welche Annahme benötigen. Gehen Sie davon aus, dass *Anwesenheit* und *Motivation* deterministisch sind. [5,5 Punkte]

i. $b = (X'X)^{-1}X'(X\beta + U\gamma + \varepsilon)$ [Aufgabenstellung]

ii. $b = \beta + (X'X)^{-1}X'U\gamma + (X'X)^{-1}X'\varepsilon$

iii. $E[b] = E[\beta] + E[(X'X)^{-1}X'U\gamma] + E[(X'X)^{-1}X'\varepsilon]$

iv. $E[b] = \beta + (X'X)^{-1}X'U\gamma + (X'X)^{-1}X'E[\varepsilon]$

v. $E[b] = \beta + (X'X)^{-1}X'U\gamma$

- Für Schritt v) wird A1 ($E[\varepsilon] = 0$) benötigt.
- Der KQ-Schätzer b ist dann unverzerrt, wenn
  - $X'U = 0$, d.h. wenn Motivation mit Anwesenheit unkorreliert ist und/oder
  - $\gamma = 0$, d.h. wenn Motivation keinen Einfluss auf Punkte hat, also irrelevant ist.

4.2 Wie unterscheidet sich der geschätzte Koeffizient für *Anwesenheit* zwischen Modell I und II, wenn Sie vermuten, dass $Cov(Anwesenheit_i, Motivation_i) > 0$ und $\gamma > 0$ gilt? Erläutern Sie Ihre Antwort. Wie können Sie den KQ-Schätzer für β_2 in Modell II interpretieren? [3 Punkte]

- Der Koeffizient für *Anwesenheit* wird im Modell II durch die pos. Korrelation und den positiven Koeffizienten für Motivation überschätzt. [1P] Der Koeffizient wäre daher in Modell I kleiner als in Modell II.
- Der KQ-Schätzer für β_2 misst nicht den kausalen Effekt der Anwesenheit, sondern den mittleren Punkteunterschied für Studierende mit unterschiedlicher Anwesenheitsquote.

4.3 Ihr Kommilitone rät Ihnen, ihr Endogenitätsproblem durch Instrumentierung von *Anwesenheit* zu lösen. Er schlägt vor, als Instrumente für *Anwesenheit* die geographische Distanz zwischen dem Wohnort der Studierenden und des Hörsaals (*Distanz*) sowie die Gesamtanzahl an verpflichtenden Semesterwochenstunden (*SWS*) zu nutzen. Erläutern Sie verbal, unter welchen Bedingungen die beiden Variablen geeignete Instrumente darstellen. [2 Punkte]

- Relevanz: *Distanz* und *SWS* müssen jeweils mit der endogenen Variable *Anwesenheit* korreliert sein.
- Exogenität: *Distanz* und *SWS* dürfen jeweils nicht mit *Motivation* und nicht mit ε, also nicht mit ν, korreliert sein.

4.4 Sie entscheiden sich, beide Instrumente in einer two-stage-least-squares (2SLS)-Schätzung von Modell II zu verwenden. Erläutern Sie kurz verbal die Vorgehensweise des 2SLS-Schätzers und stellen Sie die für die Schätzung benötigten Modellgleichungen auf. [4 Punkte]

- Der 2SLS-Schätzer ist ein zweistufiger KQ-Schätzer.
- 1. Stufe: Die endogene Variable *Anwesenheit* wird auf die verwendeten Instrumente *Distanz* und *SWS* regressiert. Hieraus ergibt sich eine Vorhersage für *Anwesenheit*.
- 2.Stufe: Modell II wird geschätzt, wobei für *Anwesenheit* die vorhergesagten Werte aus Stufe 1 eingesetzt werden.

1.Stufe:

$$Anwesenheit_i = \theta_1 + \theta_2 Distanz_i + \theta_3 SWS_i + \upsilon_i$$

2.Stufe:

$$Punkte_i = \beta_1 + \beta_2 \widehat{Anwesenheit}_i + u_i$$

Für Ihre 2SLS-Schätzung mit Stata erhalten Sie folgenden Output:

```
First-stage regressions

                                                  Number of obs   =        680
                                                  F(   2,    677) =     266.03
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.5626
                                                  Adj R-squared   =     0.5613
                                                  Root MSE        =    11.2908

------------------------------------------------------------------------------
             |               Robust
 Anwesenheit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Distanz |  -.9327058   .1060123    -8.80   0.000    -1.140858   -.7245534
         SWS |  -1.374835   .1755636    -7.83   0.000    -1.719549    -1.03012
       _cons |   111.8296   2.256388    49.56   0.000     107.3993      116.26
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression          Number of obs =     680
                                                  Wald chi2(1)  =    8.41
                                                  Prob > chi2   =  0.0037
                                                  R-squared     =  0.0195
                                                  Root MSE      =  4.6602

------------------------------------------------------------------------------
             |               Robust
      Punkte |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Anwesenheit |   .0364422     .01257     2.90   0.004     .0118055    .0610789
       _cons |    22.9135   1.029011    22.27   0.000     20.89668    24.93032
------------------------------------------------------------------------------
Instrumented:  Anwesenheit
Instruments:   Distanz SWS
```

4.5 Interpretieren Sie den Koeffizienten für *Anwesenheit* inhaltlich und statistisch. [2 Punkte]

- $b_2 = .0364422$: Ein Anstieg des Anteils besuchter Vorlesungen um einen Prozentpunkt erhöht die durchschnittliche Punktzahl im Durchschnitt c.p. um 0,0364422.
- Der Koeffizient ist statistisch signifikant auf dem 1%-Niveau.

4.6 Erläutern Sie den Begriff schwacher Instrumente und diskutieren Sie, wie sich die Stärke von Instrumenten messen lässt. Erklären Sie, ob es sich bei Ihren Instrumenten *Distanz* und *SWS* um schwache Instrumente handelt. [3,5 Punkte]

- Schwache Instrumente sind Instrumente, die nur schwach mit der endogenen Variablen korrelieren.
- Die Schwäche bzw. Stärke kann anhand des Wertes der F-Statistik der Schätzung der ersten Stufe beurteilt werden.
- Bei einer F-Statistik kleiner als 10 liegen schwache Instrumente vor.
- Hier: $F(2,677) = 266,03 > 10$, bei *Distanz* und *SWS* handelt es sich somit nicht um schwache Instrumente.

Aufgabe 5 - MC Fragen **[30 Punkte]**

Bitte geben Sie die zutreffende Antwort **auf Ihrem Multiple-Choice-Lösungsblatt** an. Zu jeder Frage gibt es genau eine richtige Antwort. Für jede korrekt beantwortete Frage erhalten Sie einen Punkt. Falsche Antworten führen nicht zu Punktabzug. Bei mehr oder weniger als einer markierten Antwort auf eine Frage gilt diese als nicht beantwortet. **Angaben auf dem Aufgabenblatt werden nicht gewertet.**

1.	Ein Messfehler in x führt bei einer KQ-Schätzung des Modells $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
a	zum Problem des ability bias.
b	zu einer Überschätzung von β_2.
c	zu endogenen Störtermen.
d	zu einer Verzerrung von β_1. **X**

2.	Bei GLS (generalized least squares) Schätzern
a	muss die Varianz-Kovarianz Matrix geschätzt werden.
b	sind nach der GLS-Transformation die Gauß-Markov Annahmen verletzt.
c	ist die Varianz-Kovarianz Matrix des Störterms bekannt. **X**
d	sind t- und F-Tests nicht gültig.

3.	Was ist eine Eigenschaft des KQ-Schätzers bei binärer abhängiger Variable?
a	Der Fehlerterm ist immer heteroskedastisch. **X**
b	Es werden immer erwartete Wahrscheinlichkeiten außerhalb des Intervalls $[0,1]$ errechnet.
c	Er ist immer verzerrt.
d	Das R^2 hat keine sinnvolle Interpretation.

4.	Der feasible GLS (FGLS)-Schätzer ist bei Gültigkeit der unterliegenden Annahmen
a	konsistent und ineffizient.
b	erwartungstreu und asymptotisch effizient.
c	erwartungstreu und effizient.
d	konsistent und asymptotisch effizient. **X**

5.	Unkorrigierte Heteroskedastie im linearen Regressionsmodel führt zu
a	Effizienz des KQ-Schätzers.
b	Verzerrung des KQ-Schätzers.
c	falschen Werten der t-Statistik.**X**
d	korrekten Standardfehlern des KQ-Schätzers.

6.	Sie schätzen das Modell $Stundenlohn_i = \beta_1 male_i + \beta_2 female_i + \beta_3 educ_i + \beta_4 (male_i \cdot educ_i) + \varepsilon_i$ mittels KQ ($educ_i \hat{=}$ Bildung in Jahren). Welche Aussage über die Schätzung trifft zu?
a	Das angepasste R^2 gibt an, wie viel Prozent der Variation im Stundenlohn durch das Modell erklärt werden.
b	Das Modell lässt sich aufgrund perfekter Multikollinearität nicht schätzen.
c	b_3 gibt die geschätzte Bildungsrendite für Frauen an. **X**
d	b_4 gibt die geschätzte Bildungsrendite für Männer an.

7.	Autokorrelation im Störterm kann behoben werden durch
a	eine Vergrößerung der Stichprobe.
b	eine OLS-Transformation.
c	die Aufnahme von irrelevanten erklärenden Variablen.
d	eine FGLS Schätzung. **X**

8.	Sie schätzen das Modell $ln\_Stundenlohn_i = \beta_1 + \beta_2 male_i + \beta_3 educ_i + \varepsilon_i$ mittels KQ-Schätzer. Der geschätzte Koeffizient b_2 ist $0,286$. Wie hoch ist der erwartete Lohnunterschied zwischen Männern und Frauen mit gleicher Ausbildung genau?
a	$28,60\%$
b	$33,11\%$ **X**
c	$48,97\%$
d	$71,40\%$

9.	Sie führen einen Chow-Test auf Geschlechterunterschiede für die KQ-Schätzung des Modells $Stundenlohn_i = \beta_1 + \beta_2 male_i + \beta_3 educ_i + \varepsilon_i$ durch. Die Teststatistik des Chow-Tests hat den p-Wert $0,4432$. Was können Sie daraus schließen?
a	Männer haben im Mittel einen höheren Stundenlohn als Frauen.
b	Der mittlere Stundenlohn von Männern und Frauen unterscheidet sich nicht signifikant.
c	Die Bildungsrendite von Männern und Frauen unterscheidet sich signifikant.
d	Die Bildungsrendite von Männern und Frauen unterscheidet sich nicht signifikant. **X**

10.	Die obere und untere Grenze für den kritischen Wert eines Durbin-Watson Tests auf Autokorrelation für ein lineares Modell mit 90 beobachteten Perioden, einer Konstante, 3 unabhängigen Variablen beträgt auf einem Signifikanzniveau von 5%
a	$d_L = 1,61$ und $d_W = 1,70$.
b	$d_L = 1,62$ und $d_W = 1,71$.
c	$d_L = 1,59$ und $d_W = 1,73$. **X**
d	$d_L = 1,57$ und $d_W = 1,75$.

11.	Welche Annahme des Gauß-Markov-Theorems ist verletzt, wenn die Matrix $\mathbf{X}$ die $(N \times K)$-Dimension (11×10) und einen Rang von 10 hat?
a	A2: $\{x_1, ..., x_N\}$ und $\{\varepsilon_1, ..., \varepsilon_N\}$ sind unabhängig.
b	A3: $V\{\varepsilon_i\} = \sigma^2$ für $i = 1, 2, ..., N$.
c	A6: $\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ konvergiert gegen eine positiv definite nichtsinguläre Matrix $\sum_{xx}$.
d	Keine. **X**

12.	Für die Matrizen $A = \begin{pmatrix} 3 & 4 \\ 6 & 9 \end{pmatrix}$ und $B = \begin{pmatrix} 3 & 7 \\ 8 & 2 \\ 4 & 1 \end{pmatrix}$ gilt
a	$AB' = \begin{pmatrix} 51 & 36 & 18 \\ 75 & 50 & 25 \end{pmatrix}$
b	$AB' = \begin{pmatrix} 18 & 36 & 51 \\ 25 & 50 & 75 \end{pmatrix}$
c	$AB' = \begin{pmatrix} 37 & 32 & 16 \\ 81 & 66 & 33 \end{pmatrix}$ **X**
d	$AB' = \begin{pmatrix} 16 & 32 & 37 \\ 33 & 66 & 81 \end{pmatrix}$

13.	Bei unabhängigen Beobachtungen entspricht die Log-Likelihoodfunktion
a	dem Produkt der individuellen logarithmierten Wahrscheinlichkeitsdichten der beobachteten abhängigen Variable.
b	der Summe der individuellen logarithmierten Wahrscheinlichkeitsdichten der beobachteten abhängigen Variable. **X**
c	dem Produkt der individuellen logarithmierten Wahrscheinlichkeitsdichten der beobachteten unabhängigen Variablen.
d	der Summe der individuellen logarithmierten Wahrscheinlichkeitsdichten der beobachteten unabhängigen Variablen.

14.	Im linearen Modell $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i$ mit $\beta_2 > 0$ und $\beta_3 < 0$ ist $f(y)$ eine
a	lineare Funktion von x_i.
b	konvexe Funktion von x_i.
c	konkave Funktion von x_i. **X**
d	konstante Funktion von x_i.

15.	Für die Hilfsregression eines RESET-Test bildet man das Quadrat und höhere Polynome der
a	abhängigen Variable.
b	Residuen.
c	vorhergesagten Werte der abhängigen Variable. **X**
d	der unabhängigen Variable, die man testen will.

16.	Schwarzs Bayesianisches Informationskriterium
a	wird mit Hilfe der Residuenvarianz ausgerechnet. **X**
b	ist nicht geeignet, genestete Modelle miteinander zu vergleichen.
c	lässt sich für eine Probitschätzung aus der Informationsmatrix ermitteln.
d	bewertet die Effizienz einer Schätzung.

17.	Die irrelevante Variable z in der KQ-Schätzung des Modells $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
a	führt zu inkonsistenten Koeffizientenschätzern, wenn die Stichprobe klein ist.
b	erhöht die Störtermvarianz.
c	führt schon bei leichter Multikollinearität zu verzerrten Koeffizientenschätzern.
d	erhöht die Varianz des Koeffizientenschätzers b_2, wenn $cov(x,z) \neq 0$. **X**

18.	Wenn für zwei Zufallsvariablen X und Y gilt, dass $E(Y\|X) = E(Y)$, dann
a	ist der Korrelationskoeffizent zwischen X und Y gleich 0. **X**
b	ist die Kovarianz zwischen X und Y gleich 1.
c	ist der auf X bedingte Erwartungswert von Y gleich 0.
d	sind X und Y statistisch unabhängig.

19.	Sie führen einen rechtsseitigen und einen beidseitigen t-Test durch. Wie unterscheiden sich die kritischen Werte, wenn sie beide Tests für das gleiche Modell, die gleiche Stichprobe und das gleiche Signifikanzniveau durchführen?
a	Der kritische Wert des einseitigen Tests ist größer.
b	Der kritische Wert des beidseitigen Tests ist größer. **X**
c	Der kritische Wert ist in beiden Tests gleich groß.
d	Die Antwort hängt von dem Vorzeichen des Koeffizienten ab.

20.	Die Annahme $\varepsilon_t \sim i.i.d.(0, \sigma^2)$ impliziert, dass die Störterme
a	nicht heteroskedastisch und nicht autokorreliert sein können. **X**
b	heteroskedastisch und autokorreliert sein können.
c	autokorreliert sein können.
d	heteroskedastisch sein können.

21.	Die Alternativhypothese im Durbin-Watson Test besagt, dass
a	Heteroskedastie vorliegt.
b	Autokorrelation vorliegt. **X**
c	sowohl Heteroskedastie als auch Autokorrelation vorliegen.
d	weder Heteroskedastie noch Autokorrelation vorliegen.

22.	Bei Vorliegen von Homoskedastie und Autokorrelation gilt für die Varianz-Kovarianz Matrix des Störterms, dass die Einträge
a	auf der Hauptdiagonalen konstant und die Einträge abseits der Diagonalen ungleich 0 sind. **X**
b	auf der Hauptdiagonalen konstant und die Einträge abseits der Diagonalen 0 sind.
c	auf der Hauptdiagonalen unterschiedlich und die Einträge abseits der Diagonalen ungleich 0 sind.
d	auf der Hauptdiagonalen unterschiedlich und die Einträge abseits der Diagonalen 0 sind.

23.	Eine t-verteilte Zufallsvariable
a	hat für große Stichproben mit hoher Wahrscheinlichkeit einen Wert bei 0. **X**
b	entspricht der Wurzel einer F-verteilten Zufallsvariable.
c	entsteht durch die Summierung mehrerer standardnormalverteilter Variablen.
d	nähert sich mit steigender Stichprobengröße der χ^2-Verteilung an.

24.	$cov(\varepsilon_i, \varepsilon_j) = 0$ für alle $i \neq j$ impliziert, dass
a	$V(\varepsilon) = \sigma^2 I$.
b	$\varepsilon_i \sim NID(0, \sigma^2)$.
c	$cov(x_i, \varepsilon_i) = 0$.
d	$E(\varepsilon_1 \varepsilon_2) = 0$. **X**

25.	Die Inverse der Matrix $A = \begin{pmatrix} 3 & 5 \\ 2 & 1 \end{pmatrix}$ lautet
a	$A^{-1} = \begin{pmatrix} -1/7 & 5/7 \\ 2/7 & -3/7 \end{pmatrix}$. **X**
b	$A^{-1} = \begin{pmatrix} -3/7 & 5/7 \\ 2/7 & -1/7 \end{pmatrix}$.
c	$A^{-1} = \begin{pmatrix} 1/7 & -5/7 \\ -2/7 & 3/7 \end{pmatrix}$.
d	$A^{-1} = \begin{pmatrix} 3/7 & -2/7 \\ -5/7 & 1/7 \end{pmatrix}$.

26.	Für die Beobachtung $(y_i, x_i) = (2, -1)$ beträgt das Residuum für die Schätzgleichung $\widehat{y_i} = 3,5 + 0,5x_i$
a	-2.
b	1.
c	3.
d	-1. **X**

27.	Die FGLS-Schätzung bei heteroskedastischen Störtermen beruht darauf, dass
a	Beobachtungen mit kleiner Störtermvarianz ein kleineres Gewicht erhalten als Beobachtungen mit großer Störtermvarianz.
b	Beobachtungen mit kleiner Störtermvarianz ein größerers Gewicht erhalten als Beobachtungen mit großer Störtermvarianz. **X**
c	nur die abhängige Variable so transformiert wird, dass Homoskedastie vorliegt.
d	die Standardfehler des KQ-Schätzers neu berechnet werden.

28.	Im linearen Modell $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i1}^2 + \beta_4 x_{i2} + \varepsilon_i$ seien x_{i1} und x_{i1}^2 exogen und x_{i2} endogen, wobei x_{i2} durch z_{i1} und z_{i2} instrumentiert werden kann. Welche der folgenden Bedingungen benötigt man zur Herleitung des IV-Schätzers nicht:
a	$\frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 - b_2 x_{i1} - b_3 x_{i1}^2 - b_4 x_{i2})x_{i1}^2 = 0.$
b	$\frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 - b_2 x_{i1} - b_3 x_{i1}^2 - b_4 x_{i2}) = 0.$
c	$\frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 - b_2 x_{i1} - b_3 x_{i1}^2 - b_4 x_{i1})x_{i2} = 0.$ **X**
d	$\frac{1}{N}\sum_{i=1}^{N}(y_i - b_1 - b_2 x_{i1} - b_3 x_{i1}^2 - b_4 x_{i2})z_{i2} = 0.$

29.	Welche der Aussagen für den KQ-Schätzer ist richtig?
a	Unverzerrtheit erfordert, dass der Störterm unabhängig von allen erklärenden Variablen ist.
b	Für $\sum_{i=1}^{N} x_i x_i'$ muss Singularität gegeben sein.
c	Konsistenz ist eine asymptotische Eigenschaft. **X**
d	Asymptotische Normalverteilung des Schätzers setzt $\varepsilon \sim N(0, \sigma^2 I)$ voraus.

30.	Der Breusch-Pagan Test
a	kann eine H_0, die nicht zutrifft, mit höherer Wahrscheinlichkeit verwerfen als der White Test. **X**
b	ist allgemeiner als der White Test.
c	prüft, ob e^2 durch die ersten und zweiten Momente und Interaktionsterme der ursprünglichen Regressoren erklärt werden kann.
d	hat eine Teststatistik, die auch in kleinen Stichproben χ^2-verteilt ist.