Masterprüfung WS 2015/16 - 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: **[18 Punkte]**

Ein Unternehmensvorstand interessiert sich für die Determinanten der Überstunden seiner Mitarbeitenden. Im Jahr 2013 wird eine Befragung von 2150 Vollzeit-Mitarbeitenden durchgeführt, die folgenden Datensatz liefert:

$Ueberstunden_i$	Überstunden von Person i im letzten Kalenderjahr
$log\_Einkommen_i$	Logarithmiertes monatliches Arbeitseinkommen von Person i in €
$Kinderlos_i$	Dummy-Variable, =1, wenn Person i kein Kind hat, =0 sonst
$Frau_i$	Dummy-Variable, =1, wenn Person i weiblich, =0 sonst
$Erfahrung_i$	Potentielle Berufserfahrung in Jahren ($Erfahrung = Alter - 6 - Bildungsjahre$)

Es wird folgendes Regressionsmodell aufgestellt und anschließend mit Stata geschätzt:

$$Ueberstunden_i = \beta_1 + \beta_2 log\_Einkommen_i + \beta_3 Kinderlos_i + \beta_4 Frau_i + \beta_5 Erfahrung_i + \beta_6 Frau\_Kinderlos_i + \varepsilon_i$$

wobei $Frau\_Kinderlos_i = Frau_i \cdot Kinderlos_i$.

```
      Source |       SS       df       MS              Number of obs =    2150
-------------+------------------------------           F(  5,  2144) =  971.71
       Model |  ??????????   ???  465.99004            Prob > F      =  0.0000
    Residual |  ??????????   ???  0.2675297            R-squared     =  ??????
-------------+------------------------------           Adj R-squared =  ??????
       Total |  ??????????   ???  1.3511093            Root MSE      =  ??????

------------------------------------------------------------------------------
 Ueberstunden |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
log_Einkommen |   937.9434   204.3245     4.57   0.000     213.2925    534.4684
    Kinderlos |   -12.9863   2.342910    -5.43   0.000     -17.5785    -8.52105
         Frau |   -77.8549   23.80924    -3.27   0.001     -124.521    17.40015
    Erfahrung |   5.890824   12.84000     0.46   0.647     -19.2755    13.73922
Frau_Kinderlos |  47.01218   9.234201     5.09   0.000     19.21274    28.91315
        _cons |   0.518024   .0729190     7.10   0.000     0.375103    13.99696
------------------------------------------------------------------------------
```

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 Einkommens führt c.p. zu einem durchschnittlichen Anstieg der Überstunden um 9,37 Stunden.
- Der Koeffizient ist statistisch signifikant auf dem 1% Niveau. (da $p = 0{,}00 < 0{,}01$)

1.2 Leiten Sie den marginalen Effekt der Variable *Kinderlos* allgemein für die Schätzgleichungen her. Wie groß ist der marginale Effekt der Variable *Kinderlos* für Männer und Frauen? Stellen Sie Null- und Alternativhypothese für einen Test auf, mit welchem Sie die Unterschiede des Effekts zwischen den Geschlechtern testen können. [4 Punkte]

- Allgemein:

$$\frac{\Delta E(Ueberstunden_i)}{\Delta Kinderlos_i} = \beta_3 + \beta_6 Frau_i$$

- Kinderlose Frauen machen c.p. im Mittel 34,026 Überstunden mehr als Frauen mit Kindern.

- Kinderlose Männer machen c.p. im Mittel 12,986 Überstunden weniger als Männer mit Kindern.
- $H_0 : \beta_6 = 0$ und $H_1 : \beta_6 \neq 0$

1.3 Berechnen und interpretieren Sie den Wert für das Bestimmtheitsmaß R^2. [4 Punkte]

- $SSE = 465,99 \cdot (K-1) = 465,99 \cdot 5 = 2329,95$ (wobei K die Konstante enthält)
- $SST = 1,351 \cdot (N-1) = 1,351 \cdot 2149 = 2903,299$
- R^2=$SSE/SST = 2329,95/2903,299 = 0,803$
- Interpretation: Das Modell erklärt 80,3% der Variation in den Überstunden.

1.4 Testen Sie auf einem Signifikanzniveau von 1%, ob Fehlspezifikation vorliegt, indem Sie Polynome 2., 3. und 4. Grades verwenden. Benennen Sie den hierzu verwendeten Test und stellen Sie die benötigte Hilfsregression auf. Definieren Sie die verwendeten Polynome. Geben Sie außerdem Hypothesen, Teststatistik, Freiheitsgrade, kritischen Wert und Testentscheidung an. Das R^2 des unrestringierten Modells lautet: 0,810. [6 Punkte]
Hinweis: Wenn Sie in Aufgabe 1.3 kein R^2 berechnet haben, verwenden Sie $R^2 = 0,800$.

- Test: RESET-Test
- Die Teststatistik basiert auf folgender Hilfsregression:

$$y_i = x_i'\beta + \alpha_2 \widehat{y}_i^2 + \alpha_3 \widehat{y}_i^3 + \alpha_4 \widehat{y}_i^4 + v_i$$

  wobei $\widehat{y}_i = x_i' b$.
- Hypothesen: H0: $\alpha_2 = \alpha_3 = \alpha_4 = 0$; H1: mindestens ein $\alpha_j \neq 0$ mit $j = 2,3,4$
- Teststatistik: $F^{emp} = \frac{(R_1^2 - R_0^2)/J}{(1-R_1^2)/(N-K)} \sim F_{J;N-K}$; $F^{emp} = \frac{(0,810-0,803)/3}{(1-0,810)/(2150-9)} = 26,293 \sim F_{3;2150-9}$

  (Alternativ für $R_0^2 = 0,800$: $F^{emp} = \frac{(0,810-0,800)/3}{(1-0,810)/(2150-9)} = 37,561 \sim F_{3;2150-9}$)
- Freiheitsgrade: J=3
- $F^{kritisch} = F_{3;2150-9;0,01} = 3,80$
- Da $F^{emp} > F^{krit}_{J;N-K;0,01}$ wird die H_0 auf dem 1% Signifikanzniveau verworfen. Der Test liefert Evidenz für Fehlspezifikation.

1.5 In einer Subgruppenanalyse begrenzen Sie Ihre Stichprobe auf 63-jährige Personen. Ein Vorstandsvorsitzender schlägt vor, die Variable *Bildungsjahre* in das Modell aufzunehmen. Benennen Sie das daraus entstehende Problem und begründen Sie Ihre Antwort. [2 Punkte]

- Die Variable *Bildungsjahre* sollte nicht aufgenommen werden, da dadurch perfekte Multikollinearität vorliegen würde.
- Begründung: Da das Alter für alle gleich ist, ergeben sich die Bildungsjahre aus

$$Bildungsjahre = 63 - Erfahrung - 6.$$

- Da *Erfahrung* kontrolliert wird, stellt *Bildungsjahre* eine Linearkombination der berücksichtigten Kontrollvariablen dar.

Aufgabe 2: **[15 Punkte]**

Als Besitzer eines Getränkestandes auf dem Nürnberger Christkindlesmarkt interessieren Sie sich für die Determinanten der abgesetzten Menge an Glühwein und befragen hierzu 180 Kunden im Alter von 20 bis 70 Jahren. Ihr Datensatz enthält folgende Variablen:

$Gluehwein_i$	Anzahl der von Person i getrunkenen Tassen Glühwein
$Alter_i$	Alter von Person i (in Jahren)
$Temperatur_i$	Außentemperatur zum Zeitpunkt des Glühweinkaufs von Person i (in Grad Celsius)

Die Regression der Variable $Gluehwein_i$ auf die Variablen $Alter_i$, $Temperatur_i$ und eine Konstante ergibt folgenden Output:

```
      Source |       SS       df       MS              Number of obs =     180
-------------+------------------------------           F( 2, 177)    =  676.02
       Model |  4083.24127   ???  81.6648253           Prob > F      =  0.0000
    Residual |  5570.51969   ???  .120801503           R-squared     =  0.4230
-------------+------------------------------           Adj R-squared =  0.4223
       Total |  9653.76095   ???  .209123344           Root MSE      =  .34757

------------------------------------------------------------------------------
   Gluehwein |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Alter |  -.0693153    .??????   ??????   ?????    -.0461786   -.092452
  Temperatur |  -.5919651    .091466    -6.47   0.000     -.412691   -.771239
       _cons |    7.39303    .481614    15.35   0.000     6.449074    8.33699
------------------------------------------------------------------------------
```

2.1 Interpretieren Sie statistisch den Zusammenhang zwischen Alter und Glühweinkonsum. [1 Punkt]

- Da das 95%-Konfidenzintervall den Wert 0 nicht enthält, ist der Koeffizient statistisch signifikant von Null verschieden auf dem 5%-Niveau.

2.2 Christian ist 22 Jahre alt und kauft seinen Glühwein bei einer Außentemperatur von -1 Grad Celsius, Klara ist 35 Jahre alt und kauft ihren Glühwein bei einer Außentemperatur von 5 Grad Celsius. Berechnen Sie unter Angabe des Rechenweges die Differenz der mittels obigem Modell vorhergesagten Anzahl an getrunkenen Tassen Glühwein zwischen Christian und Klara. [2 Punkte]
Hinweis: Runden Sie alle Angaben und Ihr Ergebnis auf die zweite Nachkommastelle.

- Christian trinkt $\widehat{Gluehwein} = 7{,}39 - 0{,}07 \cdot 22 - 0{,}59 \cdot (-1) = 6{,}44$ Tassen Glühwein.
- Klara trinkt $\widehat{Gluehwein} = 7{,}39 - 0{,}07 \cdot 35 - 0{,}59 \cdot 5 = 1{,}99$ Tassen Glühwein.
- Die Differenz beträgt 6,44 - 1,99 = 4,45 Tassen Glühwein.

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

- Unter Heteroskedastie versteht man eine Situation, in der die Varianz der Störterme im (linearen) Regressionsmodell nicht konstant ist, sondern über die Beobachtungen variiert.

- Liegt Heteroskedastie im Modell vor, so sind die Koeffizienten weiterhin unverzerrt geschätzt, ssolange A1 und A2 gelten.
- Die Standardfehler sind allerdings falsch berechnet. (Die Schätzung ist ineffizient.)

2.4 Erläutern Sie verbal, warum im vorliegenden Fall Heteroskedastie vorliegen könnte. [2 Punkte]

- Es könnte sein, dass die Varianz des Störterms mit steigendem Alter abnimmt. Personen in jüngerem Alter zeigen möglicherweise eine sehr hohe Variation in ihrem Glühweinkonsum, wohingegen bei älteren Personen eine niedrige Variation in ihrem Glühweinkonsum möglich ist. (andere Antwort möglich)

2.5 Der folgende Stata-Output enthält die Ergebnisse eines White-Tests auf Heteroskedastie. Führen Sie den Test auf einem Signifikanzniveau von $\alpha = 0,01$ durch. Geben Sie Hypothesen und Hilfsregression an. Definieren Sie die abhängige Variable der Hilfsregression. Bestimmen Sie zudem Teststatistik, Freiheitsgrade, kritischen Wert und Testentscheidung. [5,5 Punkte]

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

         chi2(?)      =    67,90
         Prob > chi2  =    ??????
```

- Hypothesen: $H_0 : V(\varepsilon_i) = \sigma^2$ für alle i (Homoskedastie); H_1 : nicht H_0
- Hilfsregression: $e_i^2 = \alpha_1 + \alpha_2 Alter_i + \alpha_3 Alter_i^2 + \alpha_4 Temperatur_i + \alpha_5 Temperatur_i^2 + \alpha_6(Alter_i \cdot Temperatur_i) + \nu_i$ wobei $e_i^2 = (y_i - x_i'b)^2$.
- 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=5
- Kritischer Wert: $\chi^2_J = \chi^2_5 = 15,09$
- Da $\chi^2_{empirisch} = 67,90 > 15,09 = \chi^2_{kritisch}$ wird die Nullhypothese auf dem 1%-Niveau verworfen. Der Test weist auf Heteroskedastie in dem vorliegenden Modell hin.

2.6 Es sei $Var(\varepsilon_i) = \sigma^2 \cdot Alter_i^9$. Zeigen Sie formal die aus einer 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: $\frac{Gluehwein_i}{Alter_i^{4,5}} = \frac{\beta_1}{Alter_i^{4,5}} + \frac{\beta_2 Alter_i}{Alter_i^{4,5}} + \frac{\beta_3 Temperatur_i}{Alter_i^{4,5}} + \frac{\varepsilon_i}{Alter_i^{4,5}}$
- $Var(\frac{\varepsilon_i}{Alter_i^{4,5}}) = \frac{1}{Alter_i^9} \cdot Var(\varepsilon_i) = \frac{1}{Alter_i^9} \cdot \sigma^2 \cdot Alter_i^9 = \sigma^2$

Aufgabe 3: **[9 Punkte]**

Mit Querschnittsdaten wird folgende Arbeitsangebotsfunktion geschätzt:

$$Arbeitszeit_i = \beta_1 + \beta_2 Lohn_i + \beta_3 Alter_i + \varepsilon_i.$$

Die Variablen sind wie folgt kodiert:

$Arbeitszeit_i$	Wochenarbeitszeit von Person i (in Stunden)
$Lohn_i$	Stundenlohn von Person i (in €)
$Alter_i$	Alter von Person i (in Jahren)

3.1 Erläutern Sie zwei mögliche Ursachen, warum die Variable $Lohn_i$ mit dem Störterm korreliert sein kann. [3 Punkte]

- Mögliche Ursachen (2 Nennungen werden bepunktet):
  - (1): Ausgelassene erklärende Variablen (omitted variable bias)
    * Unbeobachtete relevante Größen wie z.B. Motivation sind nicht im Modell enthalten.
    * Motivation korreliert mit $Arbeitszeit_i$ und mit $Lohn_i$ $\Rightarrow Cov(Lohn_i, \varepsilon_i) \neq 0$.
  - (2): Umgekehrte Kausalität (reverse causality bias)
    * Umgekehrte Kausalität liegt vor, wenn der Stundenlohn $Lohn_i$ von der Wochenarbeitszeit $Arbeitszeit_i$ abhängt.
  - (3): Messfehler in der erklärenden Variable
    * Die Variable $Lohn_i$ kann Messfehler aufweisen, wenn der Stundenlohn nur approximativ aus der Wochenarbeitszeit ($Arbeitszeit_i$) und einem (Monats-)Einkommen ausgerechnet wird.

3.2 Sie wählen die regionale Arbeitslosenquote (ALQ_i), um die Variable $Lohn_i$ zu instrumentieren. Nennen Sie die Voraussetzungen, unter denen ALQ_i als gültiges Instrument verwendet werden kann. [2 Punkte]

- Relevanz: die Instrumentvariable ALQ_i muss mit der endogenen Variable $Lohn_i$ partiell korrelieren: $Cov(Lohn_i, ALQ_i) \neq 0$.
- Exogenität: die Instrumentvariable ALQ_i darf nicht mit dem Störterm korrelieren: $Cov(ALQ_i, \varepsilon_i) = 0$.

3.3 Stellen Sie die beiden Modellgleichungen auf, welche für die 2SLS-Schätzung benötigt werden und erläutern Sie die einzelnen Schritte der Schätzung verbal. [4 Punkte]

- Der 2SLS-Schätzer ist ein zweistufiger KQ-Schätzer.

- 1. Stufe:

$$Lohn_i = \theta_1 + \theta_2 Alter_i + \theta_3 ALQ_i + \nu_i$$

  Die endogene Variable Stundenlohn $Lohn_i$ wird auf das verwendete Instrument ALQ_i und den exogenen Regressor $Alter_i$ regressiert. Hieraus ergibt sich $\widehat{Lohn_i}$.

- 2. Stufe:

$$Arbeitszeit_i = \beta_1 + \beta_2 \widehat{Lohn_i} + \beta_3 Alter_i + u_i$$

Das ursprüngliche Modell wird geschätzt, wobei die beobachtete Stundenlohn $Lohn_i$ durch die in der ersten Stufe ermittelten Werte $\widehat{Lohn}_i$ ersetzt wird.

Aufgabe 4: **[18 Punkte]**

Die Wahrscheinlichkeit einen Diebstahl zu begehen hängt von den sozioökonomischen Charakteristika einer Person ab. Dieser Zusammenhang wird mit folgendem binären Modell analysiert:

$$P(theft_i = 1|\mathbf{x}_i) = F(\beta_1 + \beta_2 unemp_i + \beta_3 male_i + \beta_4 age_i + \beta_5 famdisr_i + \beta_6 city_i)$$

Dabei wird für F die logistischen Verteilungsfunktion angenommen. Es liegt Ihnen ein Querschnittsdatensatz aus dem Jahr 2014 mit 450 befragten Personen vor. Die Variablen aus dem Datensatz sind wie folgt definiert:

Variable	Mittelwert	Std.Abw.	Min.	Max.	Beschreibung
theft	0,036	0,185	0	1	Person i hat im Jahr 2014 einen Diebstahl begangen (= 1, sonst= 0)
unemp	0,073	0,261	0	1	Person i ist arbeitslos (= 1, sonst= 0)
male	0,471	0,499	0	1	Person i ist männlich(= 1, sonst= 0)
age	36,81	5,388	18	65	Alter der Person i (in Jahren)
famdisr	0,162	0,369	0	1	Person i wurde von alleinerziehender Person erzogen (= 1, sonst= 0)
city	0,519	0,500	0	1	Person i wohnt in einer Stadt (= 1, sonst= 0)

4.1 Im Folgenden finden Sie die Ergebnisse einer Logit-Schätzung des Modells. Berechnen Sie jeweils für 25- und 50-jährige Männer die Wahrscheinlichkeiten, einen Diebstahl zu begehen. Berechnen und interpretieren Sie die Differenz in den Wahrscheinlichkeiten. Nehmen Sie für die übrigen Variablen Stichprobenmittelwerte an. [7 Punkte]

```
Logistic regression                              Number of obs   =        450
                                                 LR chi2(5)      =     281.33
                                                 Prob > chi2     =     0.0000
Log likelihood =  -1245.887                      Pseudo R2       =     0.1614

------------------------------------------------------------------------------
       theft |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       unemp |   .3664153   .0922960     3.97   0.000      .185515    .5473155
        male |   .2005059   .0196574    10.20   0.000      .161977    .2390345
         age |  -.1262412   .0368050    -3.43   0.000     -.198379   -.0541034
     famdisr |   .2125789   .0523593     4.06   0.000      .109955    .3152032
        city |   .1725274   .0481920     3.58   0.000      .078071    .2669837
       _cons |   .6978023   .1226366     5.69   0.000      .457435     0.93817
------------------------------------------------------------------------------
```

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

- Berechnung der Wahrscheinlichkeit für 25-jährige Männer:

$$P(theft = 1|male = 1, age = 25, \bar{\mathbf{x}}) = F(0,366 \cdot 0,073 + 0,201 \cdot 1 - 0,126 \cdot 25 + 0,213 \cdot 0,162 + 0,173 \cdot 0,519 + 0,698) =$$

$$F(-2,100) = \frac{1}{1 + \exp(-(-2,100))} = 0,109$$

- Berechnung der Wahrscheinlichkeit für 50-jährige Männer:

$$P(theft = 1|male = 1, age = 50, \bar{\mathbf{x}}) = F(0,366 \cdot 0,073 + 0,201 \cdot 1 - 0,126 \cdot 50 + 0,213 \cdot 0,162 + 0,173 \cdot 0,519 + 0,698) =$$

$$F(-5,250) = \frac{1}{1 + \exp(-(-5,250))} = 0,005$$

- Differenz: Die Wahrscheinlichkeit, einen Diebstahl zu begehen, unterscheidet sich zwischen 25-jährigen und 50-jährigen Männern am Stichprobenmittelwert um 10,9-0,5=10,4 Prozentpunkte.

4.2 Berechnen und interpretieren Sie den marginalen Effekt eines zusätzlichen Lebensjahres für einen arbeitslosen 25-jährigen Mann, der von einer alleinerziehenden Person erzogen wurde und in einer Stadt wohnt.
[4 Punkte]

Runden Sie alle Zahlenangaben auf die dritte Nachkommastelle.

- Bestimmung des marginalen Effekts im Logit-Modell für einen arbeitslosen 25-jährigen Mann, der von einer alleinerziehenden Person erzogen wurde und in einer Stadt wohnt:

$$\mathbf{x}'\widehat{\beta} = 0,366 \cdot 1 + 0,201 \cdot 1 - 0,126 \cdot 25 + 0,213 \cdot 0 + 0,173 \cdot 1 + 0,698 = -1,712$$

$$\frac{\partial F(\mathbf{x}_i'\widehat{\beta})}{\partial age} = \frac{\exp(\mathbf{x}_i'\widehat{\beta})}{(1 + \exp(\mathbf{x}_i'\widehat{\beta}))^2}\widehat{\beta}_{age} = \frac{\exp(-1,712)}{(1 + \exp(-1,712))^2}(-0,126) = -0,016.$$

- Ein zusätzliches Lebensjahr reduziert für einen arbeitslosen 25-jährigen Mann, der von einer alleinerziehenden Person erzogen wurde und in einer Stadt wohnt, c. p. im Mittel die Wahrscheinlichkeit, einen Diebstahl zu begehen um 1,6 Prozentpunkte.

4.3 In Ihrem Datensatz wird das Angebot an öffentlichen Freizeitaktivitäten in der Gemeinde mit der Variable *leisure* gemessen. Prüfen Sie, ob das öffentliche Freizeitaktivitätenangebot einen signifikanten Effekt auf die Wahrscheinlichkeit, einen Diebstahl zu begehen, hat. Führen Sie dazu einen Likelihood-Ratio-Test auf dem 5% Signifikanzniveau durch. Stellen Sie die Schätzgleichungen für das restringierte und das unrestringierte Modell auf. Benennen Sie dann die Hypothesen und berechnen den Wert der Teststatistik, sowie den kritischen Wert. Benennen Sie die Testentscheidung und interpretieren Sie diese.
Hinweis: Der Log-Likelihood-Wert des Modells inkl. dem Freizeitaktivitätenangebot beträgt -1243,774
[7 Punkte]

- Spezifikation des restringierten Modells:
  $P(theft_i = 1|\mathbf{x}_i) = F(\beta_1 + \beta_2 unemp_i + \beta_3 male_i + \beta_4 age_i + \beta_5 famdisr_i + \beta_6 city_i)$
- Spezifikation des unrestringierten Modells:
  $P(theft_i = 1|\mathbf{x}_i) = F(\beta_1 + \beta_2 unemp_i + \beta_3 male_i + \beta_4 age_i + \beta_5 famdisr_i + \beta_6 city_i + \beta_7 leisure_i)$
- Nullhypothese: $H_0 : \beta_7 = 0$, Alternativhypothese $H_0 : \beta_7 \neq 0$
- kritischer Wert: $LR_{kritisch} = \chi^2_{1,0.05} = 3,84$
- Teststatistik: $LR_{empirisch} = -2(\ln L_R - \ln L_U) = -2(-1245,887 - (-1243,774)) = -2(-2,113) = 4,226$
- Entscheidung: Die H0 kann auf dem 5% Signifikanzniveau verworfen werden, da $LR_{empirisch} = 4,226 > 3,84 = LR_{kritisch}$. Das öffentliche Freizeitaktivitätenangebot hat einen signifikanten Effekt auf die Wahrscheinlichkeit, einen Diebstahl zu begehen.

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.	Sind zwei Zufallsvariablen X und Y unkorreliert, so
a	ist die Summe von X und Y gleich 0.
b	sind X und Y unabhängig.
c	ist die Varianz von X und die Varianz von Y jeweils gleich 0.
d	**X** ist die Kovarianz zwischen X und Y gleich 0.

2.	Welche Aussage ist richtig?
a	**X** Die Teststatistik des Lagrange-Multiplier-Tests folgt asymptotisch der χ^2-Verteilung.
b	Der Likelihood-Ratio-Test hat N Freiheitsgrade.
c	Zur Durchführung eines Wald-Tests ist die zweimalige Schätzung des Modells (d.h. mit bzw. ohne Restriktion) notwendig.
d	In kleinen Stichproben sind Wald-, Likelihood-Ratio- und Lagrange-Multiplier-Tests äquivalent.

3.	Welcher Wert der Teststatistik liefert in einem Test auf schwache Instrumente ein Indiz für schwache Instrumente?
a	28,73.
b	12,35.
c	**X** 6,94.
d	16,94.

4.	Für die Schätzgleichung $\hat{y}_i = 2,5 - 4 \cdot x_i$ und die Beobachtung $(y_i, x_i) = (-1, 1)$ beträgt das Residuum
a	-2,5
b	**X** 0,5.
c	5,5.
d	-1,5.

5.	Der Prais-Winston Schätzer
a	ist nicht BLUE.
b	benutzt die erste Beobachtung nicht.
c	wird bei Vorhandensein von Heteroskedastie verwendet.
d	**X** ist effizienter als der Cochrane-Orcutt-Schätzer.

6.	Bei Vorliegen von perfekter Multikollinearität ist die Kreuzproduktmatrix $\mathbf{X}'\mathbf{X}$
a	invertierbar.
b	**X** singulär.
c	ein Skalar.
d	idempotent.

7.	Die FGLS-Schätzung bei heteroskedastischen Störtermen beruht darauf, dass
a	nur die abhängige Variable so transformiert wird, dass Homoskedastie vorliegt.
b	die Standardfehler des OLS-Schätzers neu berechnet werden.
c	die Varianz-Kovarianz-Matrix mit einem Vorfaktor multipliziert wird.
d	**X** Beobachtungen mit großer Störtermvarianz ein kleineres Gewicht erhalten als Beobachtungen mit kleiner Störtermvarianz.

8.	Die Summe quadrierter, standardnormalverteilter, unabhängiger Zufallsvariablen ist
a	normalverteilt.
b	**X** χ^2-verteilt.
c	F-verteilt.
d	t-verteilt.

9.	Der Durbin-Watson Test
a	ist auch bei einer Schätzung ohne Konstante gültig.
b	**X** ist auch in kleinen Stichproben gültig.
c	eignet sich zum Testen auf Heteroskedastie.
d	ist bei positiver Autokorrelation nicht durchführbar.

10.	Die Alternativhypothese im Breusch-Pagan Test besagt,
a	**X** dass Heteroskedastie vorliegt.
b	dass Autokorerelation vorliegt.
c	dass sowohl Heteroskedastie als auch Autokorrelation vorliegen.
d	dass weder Heteroskedastie noch Autokorrelation vorliegen.

11.	Das angepasste R^2
a	kann nicht negativ sein.
b	trifft eine Aussage über den Teil der Variation in x, der durch y erklärt werden kann.
c	**X** ist zum Vergleich von genesteten Modellen geeignet.
d	kann bei der Aufnahme zusätzlicher Variablen in das Modell nicht sinken, selbst wenn der Erklärungsgehalt steigt.

12.	Sie schätzen die Bildungsrendite mit $log\_Stundenlohn_i = \beta_1 + \beta_2 Bildungsjahre + \varepsilon_i$. b_2 beträgt 0,0479 und der dazugehörige p-Wert beträgt 0,08.
a	Bei einem Anstieg der Bildung um 1% steigt der Lohn c.p. im Mittel um 4,79 Euro.
b	**X** Mit jedem zusätzlichen Jahr Bildung steigt der Lohn c.p. im Mittel um 4,79%.
c	Der Koeffizient ist signifikant am 1% Niveau.
d	Der Koeffizient kann als Elasizität interpretiert werden.

13.	In welchem Fall wird b_2 im Modell $y_i = \beta_1 + \beta_2 x_{2i} + \varepsilon_i$, in dem die Variable x_{3i} ausgelassen wurde, nach unten verzerrt geschätzt?
a	**X** $\beta_3 < 0$ und Cov$(x_{i2}; x_{i3}) > 0$.
b	$\beta_2 > 0$ und Cov$(x_{i2}; x_{i3}) < 0$.
c	$\beta_3 < 0$ und Cov$(x_{i2}; x_{i3}) < 0$.
d	$\beta_2 > 0$ und Cov$(x_{i2}; x_{i3}) > 0$.

14.	Wenn $\Psi = \mathbf{I}$ gilt, dann impliziert $Var(\varepsilon) = \sigma^2 \Psi$, dass
a	**X** $cov(\varepsilon_i, \varepsilon_j) = 0$ für alle $i \neq j$.
b	Heteroskedastie vorliegt.
c	$cov(\varepsilon_i, \varepsilon_j) \neq 0$ für alle $i \neq j$.
d	$Var(\varepsilon) = 1$.

15.	Bei unabhängigen Beobachtungen entspricht die Likelihoodfunktion
a	der Summe der individuellen Dichten der beobachteten abhängigen Variablen.
b	der Summe der quadrierten individuellen Dichten der beobachteten abhängigen Variablen.
c	**X** dem Produkt der individuellen Dichten der beobachteten abhängigen Variablen.
d	dem Logarithmus der individuellen Dichten der beobachteten abhängigen Variablen.

16.	Sei $y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \varepsilon_i$ mit x_{1i} exogen und x_{2i} endogen, wobei x_{2i} durch z_{1i} und z_{2i} instrumentiert werden kann. Welche der folgenden Momentenbedingungen wird dann in einem Test auf überidentifizierende Bedingungen unter der H0 <u>nicht</u> getestet?
a	$E[y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}] = 0$.
b	$E[(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}) x_{1i}] = 0$.
c	**X** $E[(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}) x_{2i}] = 0$.
d	$E[(y_i - \beta_1 - \beta_2 x_{2i} - \beta_3 x_{3i}) z_{1i}] = 0$.

17.	Sie messen den Effekt der Anzahl der Polizisten in einem Stadtviertel (x) auf die Kriminalität (y). Was kann Ursache für das Problem $E(\varepsilon_i x_i) \neq 0$ in Ihrem Modell sein?
a	**X** Die Höhe der Kriminalität beeinflusst auch gleichzeitig die Zahl der Polizisten.
b	Die Kriminalität wird ungenau und dadurch fehlerhaft gemessen.
c	Die Residuenvariation steigt mit der Zahl der Polizisten.
d	Die Anzahl der Polizisten ist irrelevant für die Höhe der Kriminalität in einem Stadtviertel.

18.	Die Determinante der Matrix $\mathbf{A} = \begin{pmatrix} 3 & -1 \\ 4 & 2 \end{pmatrix}$ lautet
a	-10.
b	-2.
c	2.
d	**X** 10.

19.	Mit welchem Test kann unter Annahme der Gültigkeit der Instrumente die Endogenität einer Variable überprüft werden?
a	Wald-Test.
b	Zweiseitiger t-Test.
c	Strukturbruchtest.
d	**X** Hausman-Test.

20.	Wird ein Modell mit einer binären abhängigen Variable mit KQ geschätzt
a	ist die Varianz der Störterme konstant (Homoskedastie).
b	**X** können die prognostizierten Wahrscheinlichkeiten außerhalb des Intervalls [0,1] liegen.
c	sind Vorzeichen und statistische Signifikanz der Ergebnisse nicht mit einem Probit/Logit Modell vergleichbar.
d	entspricht der Mittelwert der vorhergesagten Wahrscheinlichkeiten in einem Modell ohne Konstante dem Mittelwert der abhängigen Variable.

21.	Eine Proxy Variable, welche zur Lösung des Problems ausgelassener Variablen verwendet wird, sollte möglichst hoch mit
a	**X** der ausgelassenen Variable korreliert sein.
b	den übrigen erklärenden Variablen im ursprünglichen Modell korreliert sein.
c	dem Störterm des neuen Modells korreliert sein.
d	der abhängigen Variable korreliert sein.

22.	Welcher Ausdruck beschreibt den marginalen Effekt von exp für Frauen gegeben das Modell $lohn_i = \beta_1 + \beta_2 exp_i + \beta_3 mann_i + \beta_4 age_i + \beta_5 exp_i^2 + \beta_6 exp_i \cdot mann_i + u_i$?
a	**X** $\beta_2 + 2\beta_5 exp_i$.
b	$\beta_1 + \beta_2 + 2\beta_5 exp_i$.
c	$\beta_2 + 2\beta_5 exp_i + \beta_6 mann_i$.
d	$\beta_1 + \beta_2 + 2\beta_5 exp_i + \beta_6 mann_i$.

23.	Für das Modell $daten_i = \beta_1 + \beta_2 log(freizeit_i) + \beta_3 alter_i + \beta_4 alter_i^2 + u_i$ ergibt eine KQ-Schätzung $b_1 = 10$, $b_2 = 50$, $b_3 = -0,40$ und $b_4 = 0,01$. In welchem Alter wird der durchschnittliche Datenverbrauch c.p. maximiert?
a	40.
b	**X** 20.
c	18.
d	30.

24.	Gegeben ist folgendes Modell: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$. Die aufgestellten Hypothesen lauten $H_0 : \beta_3 \geq -4$ vs. $H_1 : \beta_3 < -4$. Auf einem Signifikanzniveau von 1% führt eine Teststatistik von $-2,51$
a	bei $n = 20$ zur Ablehung von H_0.
b	**X** bei $n = 25$ zur Ablehung von H_0.
c	bei $n = 30$ nicht zur Ablehung von H_0.
d	bei $n = 35$ nicht zur Ablehung von H_0.

25.	Ein hoher AIC-Wert weist auf
a	eine Fehlspezifikation des Modells hin.
b	einen hohen Erklärungsgehalt des Modells hin.
c	**X** einen niedrigen Erklärungsgehalt des Modells hin.
d	ausgelassene Variablen hin.

26.	Sei $lny_i = \beta x_i + \varepsilon_i$. Dann gilt
a	$\beta = \frac{\Delta E(y)}{\Delta x}$.
b	$\beta = \frac{\Delta E(y)}{E(y)} \frac{x}{\Delta x}$.
c	$\beta = \Delta E(y) \frac{x}{\Delta x}$.
d	**X** $\beta = \frac{\Delta E(y)}{E(y)} \frac{1}{\Delta x}$.

27.	Das Maximum-Likelihood Schätzverfahren
a	minimiert die Summe der quadrierten Residuen.
b	liefert für ein lineares Modell dieselben Ergebnisse wie das KQ-Verfahren.
c	**X** bestimmt die geschätzten Parameter so, dass die Wahrscheinlichkeit für das Auftreten der Stichprobe maximiert wird.
d	bestimmt die Stichprobe so, dass die Wahrscheinlichkeit für die Populationsparameter maximiert wird.

28.	Das Informationskriterium BIC
a	kann nur zum Vergleich genesteter Modelle verwendet werden.
b	nimmt ausschließlich negative Werte an.
c	**X** bestraft die Hinzunahme zusätzlicher Regressoren stärker als das AIC.
d	berechnet sich aus einer Funktion der Fehlerquadratsumme und Steigungsparameter.

29.	Logit- und Probit-Schätzer
a	ergeben identische Koeffizienten.
b	ergeben identische marginale Effekte.
c	unterliegen identischen Annahmen bezüglich der Fehlertermverteilung.
d	**X** werden nicht bei stetigen abhängigen Variablen angewendet.

30.	Welche Folge hat die Aufnahme einer irrelevanten Variable in Ihr Modell?
a	Die Schätzer sind inkonsistent.
b	**X** Die Varianz der Schätzer steigt.
c	Die Schätzer sind verzerrt.
d	Der Betrag des Schätzers steigt.