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

Prüfung im Fach Ökonometrie im WS 2012/13

Aufgabe 1 (17 Punkte)

Mit einer Regression wird die tägliche Schlafdauer in Stunden in Abhängigkeit von erklärenden Merkmalen geschätzt. Für die Beobachtungen $i = 1, ..., N$ stellen Sie folgendes Modell auf:

$$sleep_i = \beta_1 + \beta_2 \cdot workhours_i + \beta_3 \cdot m\_workhours_i + \beta_4 \cdot male_i + \beta_5 \cdot city_i + \beta_6 \cdot youngkid_i + \varepsilon_i$$

In den folgenden Tabellen finden Sie sowohl deskriptive Statistiken als auch die Regressionsergebnisse.

Variable	**Mittelwert**	**Std. Abw.**	**Min.**	**Max.**	**Beschreibung**
sleep	7.78	1.06	1.80	11.18	Tägliche Schlafdauer in Stunden
workhours	5.05	2.26	0	15.27	Tägliche Arbeitszeit in Stunden
m_workhours	3.28	3.25	0	15.27	Interaktionsterm: $male \cdot workhours$
male	0.56	0.50	0	1	=1, falls Mann, =0, falls Frau.
city	0.40	0.49	0	1	=1, falls in Großstadt wohnen, 0 sonst.
youngkid	0.13	0.34	0	1	=1, falls Person Kinder jünger als 3 Jahre hat, 0 sonst.

```
      Source |       SS       df       MS              Number of obs =     706
-------------+------------------------------           F(  5,   700) =   19.42
       Model |  96.1511501     5   19.23023            Prob > F      =  0.0000
    Residual |  693.190306   700  .990271865           R-squared     =  ??????
-------------+------------------------------           Adj R-squared =  ??????
       Total |  789.341456   705  1.11963327           Root MSE      =  .99512

------------------------------------------------------------------------------
       sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   workhours |  -.1420164   .0261594    -5.43   0.000    -.1933768   -.0906561
 m_workhours |  -.0539565   .0124899    -4.32   0.000    -.0784367   -.0294763
        male |   .4859738   .1939993     2.51   0.012     .1050837    .8668639
        city |  -.1839444   .0767387    -2.40   0.017    -.3346099   -.0332788
    youngkid |  -.0739644   .1127516    -0.66   0.512    -.2953362    .1474074
       _cons |   8.479748   .1257961    67.41   0.000     8.232765    8.726731
------------------------------------------------------------------------------
```

1.1 Interpretieren Sie den geschätzten Koeffizienten der Variable *city* statistisch und inhaltlich. (1.5 Punkte)

- Statistisch: der geschätzte Koeffizient von *city* ist auf dem 5%-Niveau statistisch signifikant von Null verschieden, nicht aber auf dem 1%-Niveau.
- Inhaltlich: Personen, die in einer Großstadt wohnen, schlafen c.p. im Durchschnitt 0.18 Stunden pro Tag weniger als Personen, die nicht in einer Großstadt wohnen.

1.2 Berechnen und interpretieren Sie das Bestimmtheitsmaß. Geben Sie den Wert des korrigierten Bestimmtheitsmaßes an. (3 Punkte)

- $R^2 = 96.15/789.34 = 0.1218$. Das Modell erklärt 12.18% der Variation der abhängigen Variable *sleep*.
- $\bar{R}^2 = 1 - [(1/700) \cdot 693.19]/[(1/705) \cdot 789.34] = 0.1155$.

1.3 Berechnen Sie die vorhergesagte Schlafdauer in Stunden für Männer, die in einer Großstadt leben, täglich 5 Stunden arbeiten und keine Kinder jünger als 3 Jahre haben. Runden Sie alle Werte auf die zweite Nachkommastelle. (2 Punkte)

Die vorhergesagte Schlafdauer mit den angegebenen Merkmalen beträgt $8.48 + (-0.14 - 0.05) \cdot 5 + 0.49 - 0.18 = 7.84$ Stunden.

1.4 Testen Sie mit einer Irrtumswahrscheinlichkeit von $\alpha = 0.01$, ob der Zusammenhang zwischen Arbeitszeit und Schlafdauer für Männer und Frauen unterschiedlich ist. Geben Sie hierfür Hypothesen, Teststatistik, kritischen Wert und Testentscheidung an. (3.5 Punkte)

- Hypothesen: $H_0 : \beta_3 = 0$, $H_1 : \beta_3 \neq 0$
- Teststatistik: $t^{emp} = \frac{-0.0539565}{0.0124899} = -4.32$
- Kritischer Wert: $t^{krit}_{1-\frac{\alpha}{2},N-K} = t^{krit}_{0.995,700} = 2.576$
- Testentscheidung: Da $4.32 = |t^{emp}| > t^{krit} = 2.576$ wird die Nullhypothese auf dem 1%-Niveau verworfen. Der Effekt der Arbeitszeit auf die Schlafdauer ist für Männern stärker als für Frauen ($b_3 < 0$).

1.5 Erläutern Sie formal, welches Problem auftritt, wenn Sie zusätzlich die Dummyvariablen *nokid* (=1, falls Person keine Kinder hat, =0 sonst) und *olderkid* (=1, falls Person Kind von mindesterns 3 Jahren hat, =0 sonst) in das Regressionsmodell einfügen. Was bedeutet das für den KQ-Schätzer? (2 Punkte)

- Problem der perfekten Multikollinearität: Linearkombination der Dummyvariablen *youngkid*, *nokid*, *oldkid* ergibt die Konstante.
- Da die Matrix $\mathbf{X'X}$ nicht mehr den vollen Rang besitzt, kann der KQ-Schätzer nicht berechnet werden.
- (STATA lässt folglich eine der Dummyvariablen aus.)

1.6 Führen Sie einen RESET-Test unter Verwendung von Polynomen 2., 3., 4. und 5. Grades mit einer Irrtumswahrscheinlichkeit von 5% durch. Beschreiben Sie Ihr Vorgehen, indem Sie Hilfsregression, Hypothesen, Teststatistik, Entscheidungsregel und Testentscheidung angeben. *Hinweis* : Verwenden Sie $S_1 = 691.842$ als Fehlerquadratsumme des unrestringierten Modells. (5 Punkte)

- Die Teststatistik basiert auf folgender Hilfsregression:

$$y_i = \mathbf{x_i'}\beta + \alpha_2 \hat{y}_i{}^2 + \alpha_3 \hat{y}_i{}^3 + \alpha_4 \hat{y}_i{}^4 + \alpha_5 \hat{y}_i{}^5 + \nu_i$$

- Hypothesen: $H_0 : \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$, H_1 : mindestens ein $\alpha_j \neq 0$ mit $j = 2,3,4,5$
- Teststatistik:

$$F^{emp} = \frac{(S_0 - S_1)/J}{S_1/(N-K)}$$

- Entscheidungsregel: H_0 verwerfen, falls $F^{emp} > F^{krit}_{J,N-K,0.05} = F^{krit}_{4,706-6-4,0.05} = 2.38$
- Berechnung:

$$F^{emp} = \frac{(693.190 - 691.842)/4}{691.842/696} = 0.339$$

- Testentscheidung: Da $0.339 = F^{emp} < F^{krit}_{4,696,0.05} = 2.38$ kann die Nullhypothese auf dem 5%-Niveau nicht verworfen werden. Das Modell ist nicht fehlspezifiziert.

Aufgabe 2 (15 Punkte)

Sie interessieren sich für die Determinanten der monatlichen Ausgaben für Luxusgüter. Ihnen steht ein Datensatz mit 284 Personen zur Verfügung, welcher folgende Informationen beinhaltet:

Variable:	Beschreibung:
luxus	Monatliche Ausgaben für Luxusgüter (in 1000 €)
lnEK	Monatliches Einkommen in € (logarithmiert)
alter	Alter in Jahren
mann	Mann (=1, falls Person männlich, =0 , falls Person weiblich.)

Sie führen eine KQ-Schätzung durch und erhalten folgenden STATA-Output:

```
      Source |       SS       df       MS              Number of obs =     284
-------------+------------------------------           F(  3,   280) =  165.99
       Model |  549.268697     3  183.089566           Prob > F      =  0.0000
    Residual |  73.9011373    67  1.10300205           R-squared     =  0.8814
-------------+------------------------------           Adj R-squared =  0.8761
       Total |  623.169834    70   8.9024262           Root MSE      =  1.0502

------------------------------------------------------------------------------
       luxus |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnEK |   1.038118   .0478108    21.71   0.000     .9426876    1.133549
       alter |   0.028573   .0094614     3.02   0.004     .0096998    .0474699
        mann |  -.0856857   .0321444    -2.67   0.010    -.1498462   -.0215251
       _cons |  -.1350778   .1272177    -1.06   0.292    -.3890054    .1188498
------------------------------------------------------------------------------
```

2.1 Interpretieren Sie den geschätzten Koeffizienten von $lnEK$ inhaltlich. (1 Punkt)

- Inhaltlich: Steigt das monatliche Einkommen um 1%, so steigen die monatlichen Ausgaben für Luxusgüter c.p. im Durchschnitt um $(1.038/100) \cdot 1000€ = 10.38€$.

2.2 Erläutern Sie, weshalb bezüglich $lnEK$ ein Heteroskedastie-Problem vorliegen könnte. (2 Punkte)

Mögliche Lösung: Es wäre vorstellbar, dass Personen, die vergleichsweise wenig Einkommen erzielen, ähnliche (nämliche niedrigere) Ausgaben für Luxusgüter aufweisen. Zugleich könnten Personen, die viel Einkommen erzielen, im Hinblick auf *luxus* sehr heterogen sein. In diesem Fall würde $Var(luxus|lnEK,\mathbf{x}) = Var(\varepsilon|lnEK,\mathbf{x})$ mit steigendem $lnEK$ zunehmen.

2.3 Erläutern Sie das Vorgehen beim Breusch-Pagan Test. Gehen Sie dabei auf Hypothesen, Teststatistik, Hilfsregression und Schlußlogik ein. (4 Punkte)

- Hypothesen: $H_0 : Var(\varepsilon_i) = \sigma^2$ (Homoskedastie), $H_1 : Var(\varepsilon_i) = \sigma_i^2$ (Heteroskedastie)
- Teststatistik:
$$N \cdot R^2 \sim \chi^2_K$$
wobei N (Anzahl der Beobachtungen) und R^2 aus einer Hilfsregression stammen.
- Hilfsregression: $e_i^2 = \alpha_1 + \alpha_2 lnEK_i + \alpha_3 alter_i + \alpha_4 mann_i + \nu_i$
- Schlußlogik: ist $\chi^2_{emp} > \chi^2_{krit}$, so kann die Nullhypothese verworfen werden.

2.4 Nach der Schätzung führen Sie einen Breusch-Pagan Test auf Heteroskedastie durch.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
        Ho: Constant variance
        Variables: ln_EK alter mann

        chi2(3)      =    15.55
        Prob > chi2  =   0.0014
```

Interpretieren Sie das Ergebnis kurz. (1 Punkt)

Es liegt Heteroskedastie vor, da der p-Wert kleiner als 0.01 ist. Die Nullhypothese kann somit auf dem 1%-Niveau verworfen werden.

2.5 Erläutern Sie die Schritte des FGLS-Verfahrens, um das Heteroskedastie-Problem zu lösen, wenn Sie davon ausgehen, dass die Varianz der Störterme eine Funktion von $lnEK$ ist. (4 Punkte)

- Problem: die funktionale Form der Heteroskedastie ist nicht bekannt und muss geschätzt werden ($Var(\varepsilon|\mathbf{x}) = \sigma^2 \cdot h(\mathbf{x})$).
- 1. Man schätzt die Regression, z.B. $y_i = x_i'\beta + \varepsilon_i$, und erhält die Residuen e_i für jede Beobachtung.
- 2. Man schätzt ein parametrisches Modell der quadrierten Residuen als Funktion von $lnEK$.
- 3. Die vorhergesagten Werte werden für die Transformation des Modell genutzt.
- 4. Man schätzt das Hauptmodell auf die transformierten Variablen.

2.6 Erläutern Sie ausgehend von der Formel $Var(\mathbf{b}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Diag\{\sigma_i^2\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ kurz die Idee von White (1980) zur Berechnung robuster Standardfehler und legen Sie dar, wie diese durchgeführt wird. (3 Punkte)

- Die Matrix $Diag\{\sigma_i^2\}$ kann nicht ohne zusätzliche Annahmen geschätzt werden, da sie N unbekannte Parameter σ_i^2 enthält.
- Die Idee von White (1980) ist, stattdessen die $[KxK]$-Matrix $\mathbf{X}'Diag\{\sigma_i^2\}\mathbf{X}$ konsistent mit Hilfe der Residuen (d.h. mit $\sum_{i=1}^N e_i^2 \mathbf{x_i x_i'}$) zu schätzen.
- Die Wurzeln der Diagonalelemente der auf diese Weise geschätzten Kovarianzmatrix der Schätzer werden als robuste Standardfehler bezeichnet.

Aufgabe 3 (17 Punkte)

3.1 Definieren Sie verbal, was unter Autokorrelation zu verstehen ist und welche Gauß-Markow Annahme verletzt wird. (1 Punkt)

Unter Autokorrelation versteht man eine Situation, in der die Störterme miteinander korrelieren. Die Gauß-Markow Annahme 4, $cov(\varepsilon_i, \varepsilon_j) = 0$ für alle $i \neq j$, ist verletzt.

3.2 Zeigen Sie formal, ob sich Autokorrelation auf die Unverzerrtheit des KQ-Schätzers auswirkt. Gehen Sie von deterministischen X aus. (3 Punkte)

Für deterministische X gilt:

$$E[\mathbf{b}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] = E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})\mathbf{X}'\boldsymbol{\epsilon}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \cdot E[\boldsymbol{\epsilon}]$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot E[\boldsymbol{\epsilon}] = \boldsymbol{\beta}$$

Solange $E[\boldsymbol{\epsilon}] = 0$ gilt, beeinflusst Autokorrelation die Unverzerrtheit des KQ-Schätzers nicht.

Sie interessieren sich für die Determinanten von Immobilienpreisen. Ihnen liegt ein US-amerikanischer Datensatz für die Jahre 1947-1988 mit folgenden Variablen vor:

Variable:	Beschreibung:
lprice	Realer Hauspreis (in US-Dollar, logarithmiert)
t	Zeittrend in Jahren: t=1,...,42
inv	Reales Investitionsvolumen in Immobilien (in Millionen US-Dollar)
pop	Bevölkerung (in 1000)

Mittels KQ-Verfahren schätzen Sie folgendes Modell:

$$\ln price_t = \beta_1 + \beta_2 t_t + \beta_3 inv_t + \beta_4 pop_t + \varepsilon_t$$

```
      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  3,    38) =   60.85
       Model |  .136717016     3  .045572339           Prob > F      =  0.0000
    Residual |   .02846113    38  .000748977           R-squared     =  0.8277
-------------+------------------------------           Adj R-squared =  0.8141
       Total |  .165178146    41  .004028735           Root MSE      =  .02737

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   .0329951   .0061595     5.36   0.000     .0205257    .0454644
         inv |  -4.16e-07   2.68e-07    -1.55   0.128    -9.58e-07    1.26e-07
         pop |  0.0000111   2.43e-06     4.59   0.000     .0000161    6.22e-06
       _cons |   1.439702   .3525777     4.08   0.000     .7259455    2.153458
```

3.3 Interpretieren Sie den geschätzten Koeffizienten der Variable *t* inhaltlich. (1 Punkt)

- Inhaltlich: Mit jedem zusätzlichen Jahr steigt der Hauspreis c.p. im Mittel um ca. 3.29%.

3.4 Nennen Sie zwei Gründe dafür, dass Autokorrelation in diesem Fall vorliegen könnte (d.h. der Autokorrelationskoeffizient $\rho \neq 0$). (1.5 Punkte)

- Ausgelassene Variable, z.B. Hauspreise unterliegen den Schwankungen des Wirtschaftswachstums, welches nicht in der Regression vorkommt.
- Fehlspezifikation, z.B. der Zeittrend muss nicht linear im Modell berücksichtigt werden.
- (auch andere Nennungen möglich).

3.5 Beschreiben Sie die Grundidee des Breusch-Godfrey Tests auf Autokorrelation 1. Ordnung. Führen Sie den Test am 1% Signifikanzniveau für das Modell durch. Geben Sie Hypothesen, Teststatistik, Hilfsregression und Entscheidungsregel an. Wie entscheiden Sie sich, basierend auf folgendem STATA-Output? (5.5 Punkte)

```
estat bgodfrey, lags(1)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         23.100               1                      ?
---------------------------------------------------------------------------
                        H0: no serial correlation
```

Der Breusch-Godfrey Test untersucht, ob die verzögerten KQ-Residuen einen signifikanten Einfluss auf die KQ-Residuen haben.

- H_0: Es liegt keine Autokorrelation 1. Ordnung vor: $\rho = 0$. H_1: nicht H_0.
- $\chi^2_{empirisch} = (T-1) \cdot R^2$ ist χ^2_1 verteilt, mit R^2 und $T-1$ (Anzahl der Beobachtungen) aus der Hilfsregression.
- Hilfsregression (je nach Annahme bzgl. Exogenität der erklärenden Variablen: $e_t = \alpha_1 + \rho e_{t-1} + \alpha_2 t_t + \alpha_3 inv_t + \alpha_4 pop_t + \nu_t$ oder $\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t$.
- Wenn $\chi^2_{empirisch} > \chi^2_{kritisch}$ wird die Nullhypothese verworfen.
- Kritischer Wert: $\chi^2_{kritisch} = \chi^2_1 = 6.635$
- Da $23.1 > 6.635$ wird die Nullhypothese verworfen. Es liegt wohl Autokorrelation vor.

3.6 In dem Modell $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$ folgt der Störterm einem autoregressiven Prozess zweiter Ordnung, d.h. $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \nu_t$ mit $E(\nu_t) = 0$ und $Var(\nu_t) = \sigma^2_\nu$. Zeigen Sie anhand dieses Beispiels formal die Vorgehensweise einer GLS-Transformation nach Cochrane-Orcutt. Wie ist die Konstante im transformierten Schätzmodell definiert? Wieviele Beobachtungen stehen zur Schätzung des transformierten Modells zur Verfügung? (5 Punkte)

- AR(2) Prozess: $\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \nu_t$
- Für ε_t in Hauptgleichung einsetzen:
  $y_t = \beta_1 + \beta_2 x_t + \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \nu_t$
- Mit $\varepsilon_{t-1} = y_{t-1} - \beta_1 - \beta_2 x_{t-1}$, und $\varepsilon_{t-2} = y_{t-2} - \beta_1 - \beta_2 x_{t-2}$
- ε_{t-1} und ε_{t-2} einsetzen:
  $y_t = \beta_1 + \beta_2 x_t + \rho_1 (y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + \rho_2 (y_{t-2} - \beta_1 - \beta_2 x_{t-2}) + \nu_t$
- Auflösen und Umstellen zu: $y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} = \beta_1 (1 - \rho_1 - \rho_2) + \beta_2 (x_t - \rho_1 x_{t-1} - \rho_2 x_{t-2}) + \nu_t$.
- Die Konstante ist wie folgt definiert: $\beta_1 (1 \cdot 1 - \rho_1 \cdot 1 - \rho_2 \cdot 1) = \beta_1 (1 - \rho_1 - \rho_2)$
- Das transformierte Modell kann für die Beobachtungen $3, \ldots, T$ mit dem KQ-Verfahren geschätzt werden. Die ersten beiden Beobachtungen können nicht genutzt werden.

Aufgabe 4 (11 Punkte)

Sie schätzen den Zusammenhang zwischen Kindergesundheit und ausgewählten erklärenden Variablen mit einem binären Logit Modell. Die folgenden Tabellen präsentieren die deskriptiven Statistiken der verwendeten Variablen und die Schätzergebnisse.

Variable	Mittelwert	Std. Abw.	Min.	Max.	Beschreibung
ghealth	0.85	0.36	0	1	=1, falls Kind in guter Gesundheit, =0 sonst.
lincome	9.93	0.66	6.33	11.37	Reales Haushaltseinkommen (in €, logarithmiert)
mage	34.97	6.10	18	69	Alter der Mutter in Jahren
male	0.51	0.49	0	1	=1, falls Kind ein Junge ist, =0, falls Mädchen.
lhhsize	1.13	0.41	0	2.94	Haushaltsgröße (logarithmiert)

```
Logistic regression                               Number of obs   =      28743
                                                  LR chi2(4)      =     788.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -11949.049                       Pseudo R2       =     0.0320

------------------------------------------------------------------------------
     ghealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lincome |   .6105914   .0276457    22.09   0.000     .5564069    .6647759
        mage |   .0143377   .0029246     4.90   0.000     .0086056    .0200699
        male |  -.2050511   .0332941    -6.16   0.000    -.2703064   -.1397958
     lhhsize |  -.4392864   .0404419   -10.86   0.000    -.5185511   -.3600217
       _cons |  -4.184366   .2467145   -16.96   0.000    -4.667918   -3.700814
------------------------------------------------------------------------------
```

4.1 Berechnen und interpretieren Sie den marginalen Effekt des Alters der Mutter auf die Kindergesundheit am Mittelwert der Kovariaten. Runden Sie alle Werte auf die dritte Nachkommastelle.
Hinweis: $\overline{\mathbf{x}'}\hat{\beta} = 1.62$. (3 Punkte)

- ME im Logit Modell: $\frac{exp(\overline{\mathbf{x}'}\hat{\beta})}{(1+exp(\overline{\mathbf{x}'}\hat{\beta}))^2} \cdot \hat{\beta}_{mage} = \frac{exp(1.62)}{(1+exp(1.62))^2} \cdot 0.014 = 0.0019$
- Interpretation: Steigt das Alter der Mutter um ein Jahr, so steigt die Wahrscheinlichkeit, dass ein Kind in guter Gesundheit ist, c.p. im Mittel um 0.19 Prozentpunkte (bei durchschnittlichen Merkmalsausprägungen).

4.2 Berechnen Sie die Wahrscheinlichkeit, dass die Tochter einer Familie mit einem Haushaltseinkommen von 15000€ in guter Gesundheit ist, bei sonst durchschnittlichen Merkmalen. Interpretieren Sie Ihr Ergebnis.
Hinweis: $P(y_i = 1|\mathbf{x_i}) = F(\mathbf{x_i'}\beta) = \frac{exp(\mathbf{x_i'}\beta)}{1+exp(\mathbf{x_i'}\beta)}$. (3.5 Punkte)

- ln(15000)=9.616
- $P(y_i = 1|lincome = ln(15000), male = 0, \overline{\mathbf{x}}) = F(-4.184 + 0.611 \cdot 9.616 + 0.014 \cdot 34.97 - 0.205 \cdot 0 - 0.439 \cdot 1.13) = F(1.685) = \frac{exp(1.685)}{1+exp(1.685)} = 0.844$
- Interpretation: Die Wahrscheinlichkeit, dass ein Mädchen einer Familie mit einem Haushaltseinkommen von 15000€ und sonst durchschnittlichen Merkmalsausprägungen in guter Gesundheit ist, liegt bei 84.4 Prozent.

4.3 Sie vermuten, dass die Gesundheit eines Kindes auch von der Gesundheit der Mutter abhängt. Sie nehmen die Variable zusätzlich in das Modell auf und erhalten nach erneuter Schätzung einen Log-Likelihood Wert von -11233.995. Testen Sie am 5% Signifikanzniveau, ob sich der Erkärungsgehalt des Modells verbessert hat. Nennen Sie Testverfahren, Hypothesen, Teststatistik, Entscheidungsregel und Testentscheidung. (4.5 Punkte)

- Testverfahren: LR-Test
- Hypothesen: H_0: Steigungsparameter für Muttergesundheit = 0. H_1: nicht H_0
- Teststatistik: $\xi_{LR} = -2[\ln(\tilde{\theta}) - \ln(\hat{\theta})] \sim \chi^2_J$
- Entscheidungsregel: H0 verwerfen, falls $\xi_{LR} > \chi^2_{J=1;\alpha=0.05} = 3.84$.
- Berechnung: $\xi_{LR} = -2[-11949.049 - (-11233.995)] = 1430.108$
- Entscheidung: H_0 wird verworfen. Die neu aufgenommene Variable (Gesundheit der Mutter) enthält Informationen zur Erklärung der Kindergesundheit und verbessert den Erklärungsgehalt des Modells.

Aufgabe 5 - MC Fragen (30 Punkte)

Welche Antwort ist richtig? Kreuzen Sie nur <u>eine Antwort</u> pro Aufgabe an. Falls mehrere Aussagen korrekt sind, kreuzen Sie <u>nur</u> die entsprechende <u>Antwortkombination</u> an. Für jede richtige Antwort gibt es 1 Punkt. Für falsche Antworten werden keine Punkte abgezogen. **Verwenden Sie für Ihre Antworten ausschließlich den separaten MC-Lösungsbogen. Einträge in dieser Aufgabenstellung werden nicht gewertet!**

1.		Das Bestimmtheitsmaß R^2 im einfachen linearen Regressionsmodell
a	☑	ist größer als das korrigierte Bestimmtheitsmaß.
b	☐	berücksichtigt die zur Schätzung benötigten Freiheitsgrade.
c	☐	gibt das Verhältnis von unerklärter Variation zu Gesamtvariation an.
d	☐	kann negative Werte annehmen.

2.		Wenn es einen nicht-linearen Zusammenhang zwischen zwei Zufallsvariablen gibt, dann kann
a	☐	die Kovarianz gleich 0 sein.
b	☐	die Kovarianz negativ sein.
c	☐	dieser Zusammenhang nicht korrekt mit der Kovarianz abgebildet werden.
d	☑	alle der Antworten.

3.		Die Nullhypothese im Breusch-Pagan Test
a	☐	wird abgelehnt, wenn $R^2 \cdot N \leq \chi^2_{J,kritisch}$.
b	☐	besagt, dass keine Autokorrelation vorliegt.
c	☐	besagt, dass Heteroskedastie vorliegt.
d	☑	keine der Antworten.

4.		Wird die Nullhypothese eines Strukturbruchtests abgelehnt, so
a	☐	liegt ein Problem ausgelassener Variablen vor.
b	☑	liegen signifikante Unterschiede in den Steigungsparametern der verschiedenen Gruppen vor.
c	☐	liegt Heteroskedastie vor.
d	☐	ist das restringierte Modell korrekt spezifiziert.

5.		Der geschätzte Koeffizient b_2 des Modells $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ ist nach unten verzerrt, falls eine ausgelassene Variable z_i
a	☐	positiv mit x_i korreliert ist und y_i positiv beeinflusst.
b	☐	negativ mit x_i korreliert ist und y_i negativ beeinflusst.
c	☐	nicht mit x_i korreliert ist und y_i negativ beeinflusst.
d	☑	negativ mit x_i korreliert ist und y_i positiv beeinflusst.

6.		Bei GLS-Schätzern
a	☐	ist die funktionale Form der Varianz-Kovarianz Matrix des Störterms unbekannt und muss geschätzt werden.
b	☐	sind die Gauß-Markow Annahmen verletzt.
c	☑	gilt für das transformierte Modell die Annahme A3: $Var(\varepsilon_i) = \sigma^2$, für alle $i = 1, ..., N$.
d	☐	b und c.

7.		Welche der angegebenen Verteilungsfunktionen ist symmetrisch?
a	☑	Die t-Verteilung.
b	☐	Die χ^2-Verteilung.
c	☐	Die F-Verteilung.
d	☐	alle der genannten Verteilungen.

8.		Der Prais-Winsten Schätzer
a	☐	ist effizienter als der Cochrane-Orcutt Schätzer.
b	☐	ist nicht BLUE.
c	☐	benutzt die erste Beobachtung.
d	☑	a und c.

9.		Das Informationskriterium AIC
a	☐	nimmt ausschließlich Werte zwischen 0 und 1 an.
b	☐	sinkt, wenn zusätzliche Regressoren dem Modell hinzugefügt werden (bei gegebener Fehlerquadratsumme und Beobachtungszahl).
c	☑	steigt an, wenn die Fehlerquadratsumme der Regression ansteigt (bei gegebener Parameter- und Beobachtungszahl).
d	☐	wird ausschließlich zum Vergleich genesteter Modelle verwendet.

10.		Unabhängigkeit der zwei Zufallsvariablen X und Y impliziert,
a	☐	dass die Korrelation zwischen X und Y gleich 0 ist.
b	☐	dass die Kovarianz von X und Y gleich 0 ist.
c	☐	dass die Summe von X und Y gleich 0 ist.
d	☑	a und b.

11.		Im log-linearen Modell $ln(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i$
a	☑	ist die abhängige Variable eine lineare Funktion von x_i.
b	☐	werden prozentuale Änderungen von x_i betrachtet.
c	☐	wird b_2 als Elastizität interpretiert.
d	☐	ist y_i eine konkave Funktion von x_i.

12.		Mit welcher Annahme lässt sich die Erwartungstreue von **b** bei stochastischem **X** herleiten, wenn die Annahme, dass $\{\varepsilon_1,...,\varepsilon_N\}$ und $\{x_1,...,x_N\}$ unabhängig sind, verletzt ist?
a	☐	A4: $cov(\varepsilon_i, \varepsilon_j) = 0$ für $i, j = 1, ..., N$ und $i \neq j$.
b	☐	A5: $\varepsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I_N})$.
c	☐	A7: $E\{\mathbf{x_i} \varepsilon_\mathbf{i}\} = 0$.
d	☑	keine der Antworten.

13.		Gegeben sei $y_i = \beta_1 + \beta_2 D_i + \varepsilon_i$, wobei D_i eine binäre Variable darstellt. Welche Schätzergebnisse ändern sich, wenn Sie in dem Modell D_i durch $(1 - D_i)$ ersetzen?
a	☐	b_2.
b	☑	b_1 und b_2.
c	☐	b_1 und R^2.
d	☐	b_2 und R^2.

14.		Wenn $\Psi \neq \mathbf{I}$ gilt und Ψ eine diagonale Matrix ist, dann impliziert $Var(\varepsilon) = \sigma^2 \Psi$,
a	☐	dass $cov(\varepsilon_i, \varepsilon_j) \neq 0$ für alle $i \neq j$.
b	☑	dass Heteroskedastie vorliegt.
c	☐	dass der KQ-Schätzer verzerrt ist.
d	☐	a und b.

15.		Ein Interaktionsterm
a	☑	ist das Produkt von mindestens zwei Variablen.
b	☐	ist die Summe von mindestens zwei Variablen.
c	☐	ist eine Linearkombination mehrerer Variablen.
d	☐	muss eine Dummyvariable beinhalten.

16.		Weist der Störterm einen autoregressiven Prozess zweiter Ordnung auf, so
a	☐	sind die mit den KQ-Schätzern ausgewiesenen p-Werte gültig.
b	☐	gilt die Annahme A11 ($\varepsilon_t \sim IID(0, \sigma^2)$) nicht mehr.
c	☐	sind alle Elemente der Varianz-Kovarianz Matrix des Störterms von Null verschieden.
d	☑	b und c.

17.		Der marginale Effekt in einem binären Modell
a	☐	entspricht dem logarithmierten Koeffizienten.
b	☐	wird immer als Elastizität interpretiert.
c	☑	hat dasselbe Vorzeichen wie der geschätzte Koeffizient.
d	☐	a und c.

18.		Das Kleinstequadrateverfahren basiert auf der Minimierung
a	☐	der Summe der Störterme.
b	☐	der Summe der Residuen.
c	☑	der Summe der quadrierten Residuen.
d	☐	der quadrierten Summe der Residuen.

19.		Für das Modell $Lnlohn_i = \beta_1 + \beta_2 Alter_i + \beta_3 Alter_i^2 + \varepsilon_i$ berechnen Sie folgende statistisch signifikante Koeffizienten: $b_1 = 8.7$, $b_2 = 0.4$, $b_3 = -0.005$. Der marginale Effekt des Alters auf den logarithmierten Stundenlohn
a	☐	beträgt für jedes Alter 0.4.
b	☐	ist positiv für $Alter_i = 65$.
c	☐	ist negativ für $Alter_i = 35$.
d	☑	ist negativ für $Alter_i = 45$.

20.		Der Durbin-Watson-Test
a	☑	ist auch bei kleinen Stichproben gültig.
b	☐	darf angewendet werden, wenn verzögerte endogene Variablen im Modell vorkommen.
c	☐	unterstellt positive Autokorrelation in der Nullhypothese.
d	☐	keine der Antworten.

21.		Der kritische Wert eines White-Tests auf Heteroskedastie für ein lineares Modell mit 3 stetigen unabhängigen Variablen und einer Konstanten gegeben eine Irrtumswahrscheinlichkeit von $\alpha = 0.1$ beträgt
a	☐	6.25
b	☐	10.64
c	☐	13.36
d	☑	14.68

22.		Welche Aussage ist richtig?
a	☑	Zur Durchführung des Likelihood-Ratio- oder Wald-Tests wird das unrestringierte Modell geschätzt.
b	☐	Die Teststatistik des Wald-Tests folgt asymptotisch der F-Verteilung.
c	☐	In kleinen Stichproben sind Wald-, Likelihood-Ratio- und Lagrange-Multiplier-Tests äquivalent.
d	☐	a und c.

23.		Welche Größe ändert sich, wenn in einem einfachen linearen Regressionsmodell die erklärende Variable umskaliert wird?
a	☐	das R^2.
b	☐	die geschätzte Konstante.
c	☑	das für den geschätzten Steigungskoeffizienten ausgegebene Konfidenzintervall.
d	☐	b und c.

24.		Um einen RESET-Test durchzuführen, regressiert man in einer Hilfsregression
a	☑	die abhängige Variable auf Polynome der vorhergesagten abhängigen Variable.
b	☐	quadrierte Residuen auf alle unabhängigen Variablen, deren Quadrate und deren Interaktionen.
c	☐	Residuen auf alle unabhängigen Variablen und zeitlich verzögerte Residuen.
d	☐	keine der Antworten ist richtig.

25.		Das Maximum-Likelihood Schätzverfahren
a	☐	maximiert die Wahrscheinlichkeit, die Summe der quadrierten Residuen zu minimieren.
b	☑	bestimmt die geschätzten Parameter so, dass die Wahrscheinlichkeit für das Auftreten der Stichprobe maximiert wird.
c	☐	maximiert die Wahrscheinlichkeit der quadrierten Residuen.
d	☐	berechnet die Bevölkerungsparameter so, dass die Wahrscheinlichkeit für das Auftreten der Stichprobe maximiert wird.

26.		Welcher Test kann verwendet werden, um in dem linearen Modell $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$ auf gemeinsame Signifikanz der Parameter β_3 und β_4 zu testen?
a	☑	Wald-Test.
b	☐	RESET-Test.
c	☐	Durbin-Watson-Test.
d	☐	White-Test.

27.		Bei korrekter Spezifizierung der Likelihoodfunktion ist der ML-Schätzer:
a	☑	asymptotisch effizient.
b	☐	unverzerrt.
c	☐	t-verteilt.
d	☐	alle der Antworten.

28.		Der KQ-Schätzer $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ lässt sich nur berechnen, wenn
a	☐	die Matrix $\mathbf{Xy}$ eine quadratische Matrix ist.
b	☐	die Matrix $\mathbf{Xy}$ invertierbar ist.
c	☑	die Matrix $\mathbf{X}'\mathbf{X}$ vollen Rang besitzt.
d	☐	die Matrix $\mathbf{X}'\mathbf{X}$ nicht symmetrisch ist.

29.		Im Log-log-Modell $ln(y_i) = \beta_1 + \beta_2 ln(x_i) + \varepsilon_i$ mit $\beta_2 > 0$
a	☐	variiert die Elastizität von y_i bezüglich x_i mit dem Ausgangsniveau von x_i.
b	☑	besteht ein nicht-linearer Zusammenhang zwischen $ln(y_i)$ und x_i.
c	☐	gibt der Koeffizient b_2 eine Semi-Elastizität an.
d	☐	ist der Logarithmus von y_i eine nicht-lineare Funktion des Logarithmus von x_i.

30.		Autokorrelation im Störterm kann behoben werden durch:
a	☐	die Aufnahme von irrelevanten erklärenden Variablen.
b	☑	eine GLS Transformation.
c	☐	die Aufnahme zusätzlicher Beobachtungen.
d	☐	b und c.