Wirtschaftswissenschaftliches Prüfungsamt

Master of Economics, Finance and Philosophy
Diplomprüfung
Econometric Methods and Applications
Wintersemester 2011/12
22. Februar 2012
Prof. Dr. Ralph Friedmann

B i t t e b e a c h t e n S i e F o l g e n d e s:

1. Kleben Sie bitte Ihr Namensschild auf die dafür vorgesehene Markierung **auf dem Deckblatt des Klausurhefts**!

2. Legen Sie einen Lichtbildausweis an Ihrem Platz aus.

3. Die Klausur besteht aus 5 Aufgaben, von denen die Aufgaben 1, 2 und 3 sowie **eine** der Aufgaben 4-5 zu bearbeiten sind (entspricht 120 Punkten).
   Prüfen Sie die Vollständigkeit Ihres Exemplares nach; spätere Reklamationen können nicht berücksichtigt werden.

4. Die Reihenfolge der Bearbeitung der Aufgaben kann beliebig gewählt werden, beginnen Sie aber für jede Aufgabe eine neue Seite.

5. Bei der Korrektur werden nur die Lösungen auf den dafür vorgesehenen Blättern im Klausurheft berücksichtigt.

6. Die Benutzung von zwei beidseitig beschriebenen bzw. vier einseitig beschriebenen DIN A4-Blättern sowie (auch programmierbaren) Taschenrechnern ist erlaubt.

7. Bei allen statistischen Tests sind die Hypothesen, die Teststatistik sowie deren Verteilung unter H_0, der kritische Bereich, die Realisation der Teststatistik sowie die Entscheidung anzugeben! Ist das Signifikanzniveau nicht explizit angegeben, so ist $\alpha = 0.05$ zu verwenden.

1. Aufgabe (10 Punkte)

Es sei Y ein dreidimensionaler normalverteilter Zufallsvektor mit:

$$\mathbf{E}(Y) = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \quad \text{und} \quad \mathbf{V}(Y) = \begin{pmatrix} 3 & 1 & 0.5 \\ 1 & 4 & 0.75 \\ 0.5 & 0.75 & 2 \end{pmatrix}$$

Bestimmen Sie die Verteilung von

$$Z = \begin{pmatrix} 2 \\ -2 \\ -3 \end{pmatrix} + \begin{pmatrix} 3 & -1 & 2 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{pmatrix} Y.$$

2.Aufgabe (5 + 6 + 5 + 4 = 20 Punkte)

Gegeben seien die identisch unabhängig verteilten Zufallsvariablen $X_1, X_2, \ldots, X_n$ mit $X_1 \sim Exp(\lambda)$ und

$$E(X_1) = \frac{1}{\lambda}, \quad \mathrm{Var}(X_1) = \frac{1}{\lambda^2}.$$

Es gelte weiter $\lambda > 0$ und $n > 10$.

(a) Es sei S eine Schätzfunktion für den Erwartungswert der Zufallsvariablen $X_1, \ldots, X_n$ der Form $S = \sum_{i=1}^{n} \gamma_i X_i$.

Zeigen Sie, dass S genau dann erwartungstreu ist wenn gilt: $\sum_{i=1}^{n} \gamma_i = 1$.

(b) Entscheiden Sie für die folgenden Schätzfunktionen $S_1(X_1, \ldots, X_n), \ldots, S_6(X_1, \ldots, X_n)$, ob diese erwartungstreue Schätzer für den Erwartungswert der Zufallsvariablen $X_1, \ldots, X_n$ sind.

$$S_1(X_1, \ldots, X_n) = 4X_1 - 2X_n, \qquad S_2(X_1, \ldots, X_n) = X_1 + X_2,$$

$$S_3(X_1, \ldots, X_n) = \frac{1}{4}\left(X_1 + X_4 + X_8 + X_n\right), \qquad S_4(X_1, \ldots, X_n) = \frac{1}{9}\left(X_1 + X_9 + X_n\right),$$

$$S_5(X_1, \ldots, X_n) = \sum_{i=1}^{n} \frac{X_i}{n}, \qquad S_6(X_1, \ldots, X_n) = \sum_{i=1}^{n} \frac{1}{i} X_i$$

(c) Überlegen Sie wie Sie die nicht-erwartungstreuen Schätzfunktionen aus Aufgabenteil (b) transformieren können, so dass diese die Eigenschaft der Erwartungstreue besitzen.

(d) Gehen Sie kurz auf die Konsistenz einer Schätzfunktion ein. Welche der Schätzfunktionen aus (b) besitzt diese Eigenschaft?

3. Aufgabe (5 + 12 + 3 + 10 + 10 + 10 + 10 = 60 Punkte)

Mit Hilfe der Statistiksoftware R soll der Datensatz *CPSSW9204* aus dem Paket *AER* untersucht werden, welcher Informationen bezüglich der Lohnverteilung von Vollzeitangestellten im Alter zwischen 25 und 34 in den Jahren 1992 und 2004 zur Verfügung stellt. Als abhängige Variable soll *earnings* verwendet werden, welche den durchschnittlichen Stundenlohn der Arbeiter repräsentiert, erklärende Variablen sollen aus dem Geschlecht (*female,male* ; binär), dem Alter (*age*) und den zwei Binärvariablen für den höchsten erreichten Abschluss (*bachelor,highschool*) gewählt werden.
Folgender Output entstand bei der Analyse der Daten:

```
Call:
lm(formula = earnings ~ age + female + bachelor, data = CPSSW9204)

Residuals:
    Min      1Q  Median      3Q     Max
-17.411  -4.529  -1.171   3.070  46.787

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.72215        ???   1.205    0.228
age          0.40732    0.01987  20.500   <2e-16 ***
female           ???    0.11514 -22.940   <2e-16 ***
bachelor     5.99748    0.11505     ???   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.068 on 15584 degrees of freedom
Multiple R-squared: 0.1809,     Adjusted R-squared: ???
F-statistic:  ??? on ??? and ??? DF,  p-value: < 2.2e-16
```

(a) Formulieren Sie die Modellgleichung, durch deren Schätzung obiger Output entstanden ist. Warum wurden die Variablen *male* und *highschool* nicht auch in das Modell aufgenommen, beziehungsweise was hätte dies zur Folge gehabt?

(b) Berechnen Sie die fehlenden Werte (???) im Output und treffen Sie eine Entscheidung, welche der Koeffizienten signifikant von Null verschieden sind bei einem Signifikanzniveau von $\alpha = 0.01$. Geben Sie weiterhin die Hypothesen H_0 und H_1 an, auf die sich die jeweilige t-Statistik und der p-Wert beziehen.

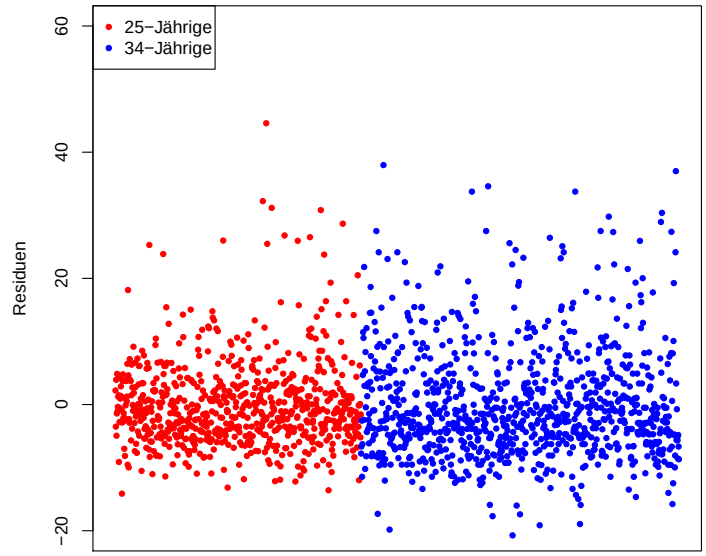
(c) Können Sie eine Aussage zur Signifikanz des Erklärungsansatzes machen? Geben Sie auch hier die entsprechenden Hypothesen H_0 und H_1 an.

(d) Als ein Kritikpunkt der Schätzung wird angeführt, dass man die Daten der Jahre 1992 und 2004 nicht zusammen für eine Modellschätzung verwenden sollte, da eine zu große Zeitspanne besteht und man nicht die gleichen Koeffizienten für beide Jahre annehmen kann. Aus diesem Grund wurden zwei separate Schätzungen durchgeführt, zum einem mit den Daten aus dem Jahr 1992 und zum anderen mit denen aus dem Jahr 2004, was folgende zwei Outputs lieferte.

```
Call:
lm(formula = earnings ~ age + female + bachelor, data = CPSSW9204,
    subset = 1:7602)

Residuals:
    Min      1Q  Median      3Q     Max
-13.547  -3.317  -0.602   2.597  32.674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.60568    0.61556   0.984    0.325
age          0.34223    0.02041  16.767   <2e-16 ***
female      -2.00384    0.11589 -17.290   <2e-16 ***
bachelor     4.38391    0.11770  37.247   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.988 on 7598 degrees of freedom
Multiple R-squared: 0.1949,     Adjusted R-squared: 0.1946
F-statistic: 613.2 on 3 and 7598 DF,  p-value: < 2.2e-16

Call:
lm(formula = earnings ~ age + female + bachelor, data = CPSSW9204,
    subset = 7603:15588)

Residuals:
    Min      1Q  Median      3Q     Max
-20.477  -5.098  -1.266   3.431  44.828

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.88380    0.92029   2.047   0.0407 *
age          0.43920    0.03053  14.387   <2e-16 ***
female      -3.15786    0.18036 -17.508   <2e-16 ***
bachelor     6.86515    0.17837  38.489   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.884 on 7982 degrees of freedom
Multiple R-squared:  0.19,      Adjusted R-squared: 0.1897
F-statistic: 624.1 on 3 and 7982 DF,  p-value: < 2.2e-16
```

Verwenden Sie einen geeigneten Test um die vorgebrachte Kritik entweder zu verifizieren oder zu entkräften.

(e) Im Weiteren werden nur noch die Daten für das Jahr 2004 verwendet. Ein weiterer Kritikpunkt des Modells bestehe nun in der Annahme homoskedastischer Störgrößen, wobei sich die Kritiker auf folgende Grafik stützen.



Verwenden Sie einen geeigneten Test zur Überprüfung der Hypothese, dass die Störtermvarianz für die Gruppe der 34-Jährigen größer ist als für die Gruppe der 25-Jährigen. Gehen Sie hierbei auf die einzelnen Schritte ein, die notwendig sind, um die Teststatistik berechnen zu können.
Verwenden Sie folgende Ergebnisse für die Schätzung des Modells

$$y_i = \beta_0 + \beta_1 female_i + \beta_2 bachelor_i + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma^2)$$

für die Gruppe der 25-Jährigen mit einer Gruppenstärke von 725 Personen,

$$\widehat{u}_{25}'\widehat{u}_{25} = 33542.5720,$$

und für die Gruppe der 34-Jährigen mit einer Gruppenstärke von 939 Personen,

$$\widehat{u}_{34}'\widehat{u}_{34} = 70325.6337.$$

(f) Aufgrund von (e) wird in Betracht gezogen die Modellannahmen bezüglich der Störterme zu verallgemeinern und

$$u_i \sim N(0, \sigma^2_{age_i}), \quad \text{wobei} \quad age_i \in \{25, 26, \ldots, 34\}$$

zu verwenden.
Welche Gestalt hat bei diesen Annahmen die Varianz-Kovarianz-Matrix der Störterme,wenn die Beobachtungen nach dem Alter der Merkmalsträger geordnet werden? Geben Sie formal den erwartungstreuen und varianzminimalen Schätzer $\check{\beta}$ und dessen Varianz-Kovarianz-Matrix an. Welche Schritte sind jetzt noch notwendig, um eine berechenbare Version des Schätzers $\check{\beta}$ zu erhalten?

(g) Es wird eingewandt, dass die Heteroskedastizität nicht nur vom Alter der Merkmalsträger abhängig ist. Wird, wenn dieser Einwand zutrifft, dadurch die Gültigkeit des LS-Outputs und/oder des in (d) durchgeführten Strukturbruchtests beeinträchtigt? (kurze Erläuterung)

Zur Lösung der Aufgabenteile (d) und (e) stehen Ihnen die folgenden Quantile zur Verfügung:

$$F_{936,722;0.95} = 1.1228, \quad F_{3,15588;0.95} = 2.6055, \quad F_{4,15580;0.95} = 2.3725, \quad F_{4,7982;0.95} = 2.373$$

Wahlteil:

Bearbeiten Sie genau EINE der verbleibenden ZWEI Aufgaben (4. + 5.)!

4. Aufgabe (8 + 7 + 5 + 8 + 2 = 30 Punkte)

(a) Formulieren Sie (allgemein) ein Modell für Paneldaten indem Sie *fixed effects* sowohl für die Untersuchungseinheiten als auch die verschiedenen Perioden berücksichtigen. Auf was müssen Sie besonders achten?

Bei einer großen Anzahl an Untersuchungseinheiten und Perioden kann die Schätzung des Modells eventuell sehr zeitintensiv werden; gehen Sie kurz darauf ein, wie man dieses Problem umgehen kann.

(b) Ein Gastronom denkt über eine Umstrukturierung seines Geschäftskonzepts nach und plant unter Umständen verstärkt sogenannte *Themen-Abende* anzubieten. Da er sich über die Preisgestaltung an solchen Abenden noch nicht im Klaren ist, will er auch das Verhältnis von Rechnungsbetrag und Höhe des Trinkgeldes mit in seine Überlegungen einfliessen lassen, um eventuell seine Mehrkosten durch das erhöhte Trinkgeld abzufangen.
Hierzu notiert er sich die Datenpaare (x_i = Rechnungsbetrag, y_i = Trinkgeld, beides in EUR) von vier seiner Stammgäste an insgesamt vier Abenden, wobei die ersten drei gewöhnliche Abende waren und es sich beim vierten Abend um ein Themen-Abend handelt.

	Gast 1	Gast 2	Gast 3	Gast 4
1. Abend	$(x_1, y_1) = (20, 2)$	$(x_2, y_2) = (40, 2)$	$(x_3, y_3) = (50, 4)$	$(x_4, y_4) = (30, 3)$
2. Abend	$(x_5, y_5) = (30, 3)$	$(x_6, y_6) = (35, 1)$	$(x_7, y_7) = (41, 3)$	$(x_8, y_8) = (20, 2)$
3. Abend	$(x_9, y_9) = (25, 2)$	$(x_{10}, y_{10}) = (14, 1)$	$(x_{11}, y_{11}) = (20, 1)$	$(x_{12}, y_{12}) = (30, 2)$
4. Abend	$(x_{13}, y_{13}) = (50, 7)$	$(x_{14}, y_{14}) = (40, 4)$	$(x_{15}, y_{15}) = (40, 5)$	$(x_{16}, y_{16}) = (40, 6)$

i) Gehen Sie mit Hilfe der obigen Daten auf die Begriffe *Zeitreihe*,*Querschnittsdaten* und *Paneldaten* ein. Geben Sie je ein Beispiel mit konkreten Datenpaaren an. Handelt es sich hier um ein balanciertes Panel?

ii) Formulieren Sie nun ein Modell für das konkrete Beispiel des Gastronoms, in dem nur zeitliche Effekte berücksichtigt werden, in zwei Varianten,

- derart, dass das Basisniveau der Höhe des Trinkgeldes am $i-$ten Abend an den Koeffizienten abgelesen werden kann und alternativ dazu,
- derart, dass die Differenz der Höhe des Trinkgeldes zum Referenzabend an den Koeffizienten abgelesen werden kann.

iii) Es soll nun eine LS-Schätzung des Modells $y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \delta_1 B_1 + \delta_2 B_2 + \delta_3 B_3 + u_{i,t}$

$$\text{mit} \qquad B_i = \begin{cases} 1 & , \text{ Beobachtung am i-ten Abend} \\ 0 & , \text{ sonst} \end{cases}$$

durchgeführt werden. Berechnen Sie den zur Durchführung der Schätzung benötigten Vektor $X'y$.

iv) Nachdem der fehlende Vektor berechnet wurde, konnte die Schätzung durchgeführt werden und lieferte den Koeffizientenschätzer

$$\widehat{\beta} = (3.0468, 0.0577, -2.3170, -2.6150, -2.8310)'$$

Wie hoch ist das Basisniveau der Höhe des Trinkgeldes am zweiten Abend?

5. Aufgabe (7 + 10 + 6 + 7 = 30 Punkte)

Ein lineares Modell der Form

$$y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + u_i, \quad u_i \overset{iid}{\sim} N(0, \sigma^2)$$

soll mit der Kleinst-Quadrate-Methode geschätzt werden. Die ursprünglichen Daten sind allerdings nicht mehr vorhanden, sondern nur noch die von den Daten erzeugten Summen

$$\sum_{i=1}^{100} x_{2,i} = 123, \sum_{i=1}^{100} x_{3,i} = 96, \sum_{i=1}^{100} x_{2,i}^2 = 252, \sum_{i=1}^{100} x_{3,i}^2 = 167, \sum_{i=1}^{100} x_{2,i} \cdot x_{3,i} = 125, \sum_{i=1}^{100} y_i = 460,$$

$$\sum_{i=1}^{100} y_i \cdot x_{2,i} = 810, \quad \sum_{i=1}^{100} y_i \cdot x_{3,i} = 615, \quad \sum_{i=1}^{100} y_i^2 = 3200$$

und die Matrix

$$(X'X)^{-1} = \begin{pmatrix} 0.0353 & -0.0114 & -0.0118 \\ -0.0114 & 0.0100 & -0.0009 \\ -0.0118 & -0.0009 & 0.0134 \end{pmatrix}$$

(a) Berechnen Sie den KQ-Schätzer $\widehat{\beta}$ und dessen Varianz-Kovarianz-Matrix.

(b) Testen Sie mit einem geeigneten Test auf die Signifikanz des Erklärungsansatzes.

(c) Überprüfen Sie mit einem geeigneten Test jeweils die Hypothese

i) $H_1 : \beta_1 \neq 0$.

ii) $H_1 : \beta_2 > 0$.

iii) $H_1 : \beta_3 < 0$.

(d) Bestimmen Sie den Varianz-Inflations-Faktor für X_1 und X_2. Kann man bei diesen Daten sagen, dass Multikollinearität keine große Rolle spielt?

Hinweis: Sie dürfen verwenden, dass im linearen Modell mit $k = 3$ Parametern folgende Beziehung besteht:

$$Korr\left(\widehat{\beta}_2, \widehat{\beta}_3\right) = -\frac{s_{x_2,x_3}}{\sqrt{s_{x_2,x_2} \cdot s_{x_3,x_3}}}$$

Zur Lösung der Aufgabenteile (b) und (c) stehen Ihnen die folgenden Quantile zur Verfügung:

$t_{97;0.975} = 1.9847, \quad t_{103;0.95} = 1.6598, \quad t_{97;0.95} = 1.6607, \quad F_{2,97;0.95} = 3.0902, \quad F_{3,97;0.95} = 2.6984$

Viel Erfolg!