Wirtschaftswissenschaftliches Prüfungsamt

Diplomprüfung / Bachelorprüfung

Ökonometrie

Sommersemester 2012, Erster Prüfungstermin, 01. August 2012

Prof. Dr. Ralph Friedmann

Name, Vorname: _____

Matrikelnummer: _____

B i t t e b e a c h t e n S i e :

(a) Kleben Sie bitte Ihr Namensschild auf die dafür vorgesehene **Markierung auf dem Deckblatt des Klausurhefts**!

(b) Legen Sie Ihren Lichtbild- und Ihren Studierendenausweis an Ihrem Platz aus.

(c) Die Klausur besteht aus vier Aufgaben. Die vollständige Lösung einer Aufgabe wird mit 40 Punkten bewertet. Die Klausur enthält keinen Tabellenanhang. Von den vier Aufgaben sind genau **drei** zu bearbeiten.

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

(e) Überprüfen Sie die Vollständigkeit Ihres Klausurexemplares. Spätere Reklamationen können nicht berücksichtigt werden.

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

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

Aufgabe 1 [12 + 4 + 4 + 6 + 4 + 10 = 40 Punkte]

Ein einfaches lineares Regressionsmodell sei wie folgt spezifiziert:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Geben Sie unter Verwendung dieser Notation kurz an:

(a) (i) die Modellannahmen
(ii) Regressand, Regressor und Störterm
(iii) abhängige und unabhängige Variable
(iv) (Stichproben- und Populations-) Regressionsgerade, Achsenabschnitt, Steigung
(v) Regressionskoeffizienten und geschätzte Regressionskoeffizienten
(vi) Methode der Kleinsten Quadrate

(b) Wie kann das Bestimmtheitsmaß R^2 allgemein interpretiert werden? Wie wird es berechnet und welche Werte kann diese Größe annehmen?
Welche Beziehung besteht zwischen dem Bestimmtheitsmaß und dem quadrierten Korrelationskoeffizienten von X und Y im einfachen Regressionsmodell?

(c) Entspricht das Bestimmtheitsmaß einer Regression von Y auf X dem der Regression von X auf Y? Begründen Sie Ihre Antwort.

(d) Betrachten Sie den Fall, dass $\beta_1 = 0$ gilt. Leiten Sie einen Schätzer für β_0 her. Zeigen Sie, dass dann $R^2 = 0$ gilt.

(e) Skizzieren Sie für ein einfaches lineares Regressionsmodell ein Streudiagramm für die Punkte (X, Y) und eine zugehörige Regressionsgerade für den Fall, dass $Korr(X, Y)$ den Wert 1 annimmt.

(f) Nehmen Sie nun an, dass $\beta_0 = 1$ bekannt ist. Geben Sie die Formel für den Kleinst-Quadrate-Schätzer $\widehat{\beta}_1$ an. Prüfen Sie die Erwartungstreue von $\widehat{\beta}_1$, also ob $\mathrm{E}(\widehat{\beta}_1|X) = \beta_1$ gilt.

Aufgabe 2 [10 + 3 + 4 + 3 + 4 + 6 + 4 + 6 = 40 Punkte]

Unter Verwendung der Daten des US-amerikanischen Current Population Survey für das Jahr 1988 wurde für Männer im Alter von 18 bis 70 Jahren ein multiples Regressionsmodell für den Wochenlohn (`wage`) mittels OLS geschätzt. Als erklärende Variablen wurden verwendet:

`experience`	Berufserfahrung in Jahren
`education`	Ausbildung in Jahren
`region`	qualitative Variable mit den vier Ausprägungen: *northeast,midwest,south,west*
`parttime`	War der Mann teilzeitbeschäftigt? Qualitative Variable mit den Ausprägungen: *no,yes*

Eine Schätzung unter der Annahme von homoskedastisch-verteilten Störgrößen mit der Software `R` liefert folgenden Output:

```
Call:
lm(formula = log(wage) ~ experience + I(experience^2) + education +
    region + parttime, data = CPS1988)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6699 -0.3080  0.0347  0.3394  4.8344

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         ???     1.891e-02 243.420  < 2e-16
experience       5.511e-02  8.620e-04   ???    < 2e-16
I(experience^2) -8.551e-04  1.852e-05 -46.182  < 2e-16
education        8.770e-02  1.166e-03  75.223  < 2e-16
regionmidwest   -7.106e-02     ???     -7.649 2.09e-14
regionsouth     -1.385e-01  8.804e-03 -15.729  < 2e-16
regionwest      -5.521e-02  9.588e-03  -5.759 8.57e-09
parttimeyes     -8.860e-01  1.194e-02 -74.231  < 2e-16
---

Residual standard error: 0.535 on 28147 degrees of freedom
Multiple R-squared: 0.4416,Adjusted R-squared:  ???
F-statistic:   ??? on ? and ??? DF,  p-value: < 2.2e-16
```

(a) Geben Sie die Modellgleichung an und berechnen Sie die fehlenden Werte (mit `???` gekennzeichnet).

(b) Welche Koeffizienten sind signifikant zu einem Niveau von $\alpha = 0.05$. Was würde sich ändern wenn Sie $\alpha = 0.01$ verwenden? Besteht ein nicht-linearer Zusammenhang des Wochenlohnes und der Berufserfahrung bei $\alpha = 0.01$? Begründen Sie Ihre Antwort.

(c) Ist der Erklärungsansatz signifikant? Geben Sie Nullhypothese und Gegenhypothese sowie die Verteilung der Teststatistik unter H_0 an, wobei Sie davon ausgehen können, dass homoskedastisch normalverteilte Störgrößen vorliegen.

(d) Berechnen Sie ein 95%-Konfidenzintervall für $\beta_{experience}$. Wie können Sie dieses Intervall interpretieren?
Verwenden Sie zur Berechnung des Konfidenzintervalles eines der folgenden Quantile der Standardnormalverteilung.

$$z_{0.99} = 2.326, \quad z_{0.95} = 1.645, \quad z_{0.975} = 1.960$$

(e) Bestimmen Sie, wie sich der für den Süden (`region = regionsouth`) prognostizierte Wochenlohn eines Vollzeitarbeitnehmers mit gegebener Ausbildung und Erfahrung prozentual unterscheidet von den prognostizierten Wochenlöhnen eines identischen Arbeitnehmers in den übrigen drei Regionen. Beachten Sie, dass als abhängige Variable das logarithmierte wöchentliche Einkommen verwendet wurde.

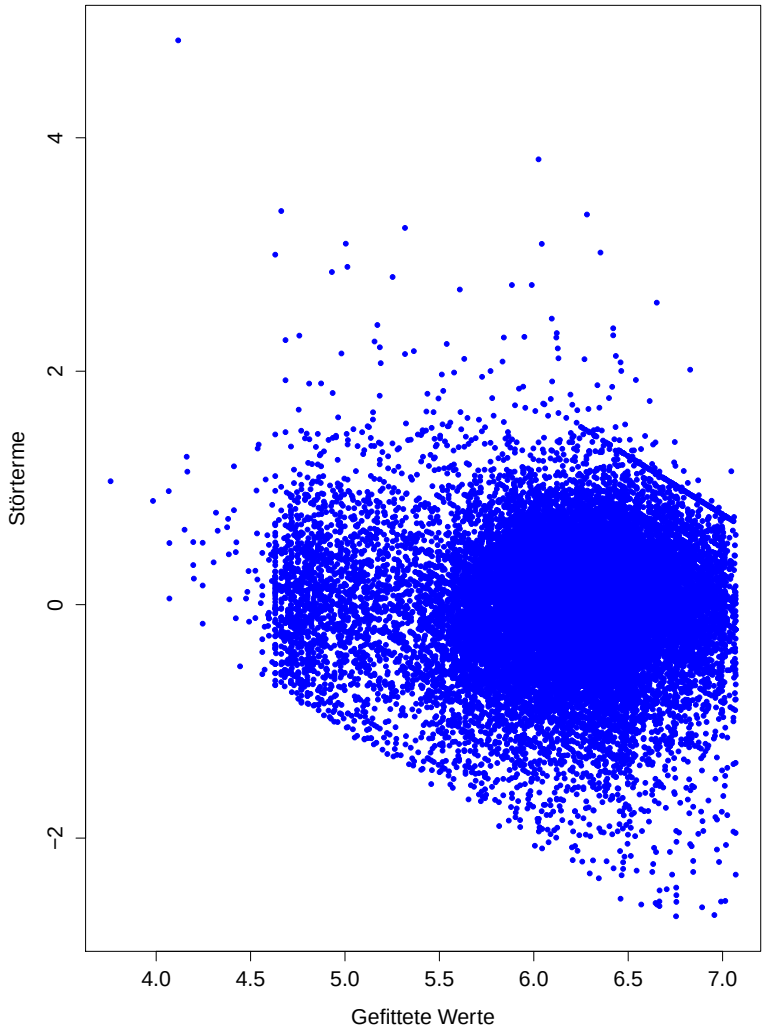
(f) Ändern sich die Koeffizienten für

- `experience`,`experience`^{2},`education`,`parttimeyes` und
- `regionmidwest`,`regionsouth`,`regionwest`,

wenn Sie die binäre Variable `regionnortheast` anstatt des Absolutgliedes in die Modellgleichung aufnehmen? Welchen Wert wird der Koeffizient von `regionnortheast` annehmen?

(g) Begründen Sie, warum nicht für jede Ausprägung von `region` eine binäre Variable und zusätzlich das Absolutglied in die Modellgleichung aufgenommen werden können. Welche Konsequenzen würden sich hierduch für die OLS-Schätzung ergeben?

(h) Bei der Analyse des geschätzten Modells wurden die gefitteten Werte $x_i'\widehat{\beta}$ gegen die Residuen $\widehat{u}_i$ geplottet, wobei folgende Grafik entstand:



Welche Verteilungseigenschaft der Störterme wird durch den Plot deutlich? Wurde diese Eigenschaft bei den bisherigen Berechnungen berücksichtigt? Wieweit behält der Output der Kleinst-Quadrate-Schätzung seine Gültigkeit (Schätzer, Standardfehler, t-Teststatistiken, F-Teststatistik, p-Werte)? Machen Sie einen Verbesserungsvorschlag!

Aufgabe 3 [4 + 4 + 6 + 6 + 10 + 4 + 6 = 40 Punkte]

Mit einem linearen Wahrscheinlichkeitsmodell soll die Wahrscheinlichkeit untersucht werden, mit der eine Person in den USA krankenversichert ist. Als abhängige Variable wurde `insurance` verwendet, welche den Wert 1 annimmt, wenn die Person krankenversichert ist und 0 sonst. Folgende erklärende Variablen wurden genutzt:

`gender`	Geschlecht: Factor mit den Levels *female* und *male*
`age`	Alter der Person
`family`	Familengröße: Ehepartner/in + Anzahl der Kinder
`married`	Ist die Person verheiratet? Factor mit den Levels *no* und *yes*
`selfemp`	Ist die Person selbständig? Factor mit den Levels *no* und *yes*

Eine Schätzung mit `R` produzierte den folgenden Output:

```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  0.70346506  0.02022132  34.7883 < 2.2e-16 ***
gendermale  -0.05001710  0.00820907  -6.0929 1.155e-09 ***
age          0.00348548  0.00041715   8.3554 < 2.2e-16 ***
marriedyes   0.16743034  0.01001336  16.7207 < 2.2e-16 ***
selfempyes  -0.16593087  0.01452594 -11.4231 < 2.2e-16 ***
family      -0.03065321  0.00319156  -9.6045 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Geben Sie explizit die Modellgleichung an. Warum spricht man hier von einem linearen Wahrscheinlichkeitsmodell? Würden Sie eine Modellschätzung mit Annahme homoskedastischer Störgrößen als sinnvoll erachten?

(b) Wie können Sie die geschätzten Koeffizienten in diesem Modell interpretieren? Wie groß ist die Differenz bei der Wahrscheinlichkeit krankenversichert zu sein zwischen ledigen und verheirateten Personen, wenn die übrigen Regressoren identisch sind? Ist diese Differenz signifikant?

(c) Berechnen Sie die prognostizierte Wahrscheinlichkeit, dass eine 55-jährige, verheiratete aber kinderlose Angestellte krankenversichert ist. Nehmen Sie Stellung zu Ihrem Ergebnis.

(d) Beschreiben Sie den Modellansatz des Probit- und Logit-Modell für binäre abhängige Variablen. Kann es mit diesen Modellen auch passieren, dass Wahrscheinlichkeiten größer 1 prognostiziert werden? Wenn nein, warum nicht?

(e) Durch Schätzung des Probit- beziehungsweise Logit-Modells entstanden die folgenden beiden Outputs:

```
  Call:
glm(formula = insurance ~ gender + age + married + selfemp +
    family, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5276   0.3674   0.5279   0.6920   1.5861
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.506306   0.070735   7.158 8.20e-13 ***
gendermale  -0.186536   0.032085  -5.814 6.11e-09 ***
age          0.013614   0.001577   8.633  < 2e-16 ***
marriedyes   0.597743   0.035981  16.613  < 2e-16 ***
selfempyes  -0.593481   0.045800 -12.958  < 2e-16 ***
family      -0.104420   0.010332 -10.106  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8780.2  on 8801  degrees of freedom
Residual deviance: 8097.8  on 8796  degrees of freedom
AIC: 8109.8

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = insurance ~ gender + age + married + selfemp +
    family, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4665   0.3811   0.5259   0.6848   1.6452

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.822997   0.123516   6.663 2.68e-11 ***
gendermale  -0.342071   0.056742  -6.029 1.65e-09 ***
age          0.023949   0.002807   8.532  < 2e-16 ***
marriedyes   1.046932   0.063629  16.454  < 2e-16 ***
selfempyes  -1.030240   0.078078 -13.195  < 2e-16 ***
family      -0.180939   0.017661 -10.245  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8780.2  on 8801  degrees of freedom
Residual deviance: 8095.2  on 8796  degrees of freedom
AIC: 8107.2

Number of Fisher Scoring iterations: 4
```

Bestimmen Sie im Logit-Modell den Indexwert $x'\widehat{\beta}$ für ledige, kinderlose, selbständige Männer im Alter von 30 Jahren. Wie groß ist die geschätzte Wahrscheinlichkeit, dass ein solcher Mann krankenversichert ist und wie groß ist der marginale Effekt des Alters?

(f) Testen Sie im Probit- und Logit-Modell auf die Signifikanz des Erklärungsansatzes mittels eines Likelihood-Ratio-Tests. Welche (approximative) Verteilung besitzt die Teststatistik unter der Nullhypothese?
Als kritischen Wert nehmen Sie eines der folgenden Quantile:

$$\chi^2_{0.95;1} = 3.8415, \quad \chi^2_{0.95;5} = 11.0705, \quad \chi^2_{0.95;6} = 12.5916$$

(g) Gehen Sie im Rahmen des Logit-Modells auf die Begriffe *Odds* und *Log-Odds* ein und wie die Koeffizienten bezogen auf diese beiden Größen interpretiert werden können.

Aufgabe 4 [4 + 4 + 6 + 6 + 6 + 6 + 3 + 5 = 40 Punkte]

Betrachtet wird ein balanciertes Panel für 48 US-Bundesstaaten (ohne Hawaii und Alaska) für die Jahre 1982 bis 1988 zu den folgenden Variablen:

`state`	Indikator des Bundesstaates
`year`	Indikator des Jahres
`unemp`	Arbeitslosenrate
`beertax`	Steuer für einen Kasten Bier (in %)
`drinkage`	Mindestalter für legalen Alkoholkonsum
`miles`	Durchschnittlich gefahrene Kilometer pro Jahr
`jail`	Existiert eine Mindestgefängnisstrafe für Alkohol am Steuer?
`service`	Muss gemeinnützige Arbeit geleistet werden?
`income`	Jährliches Pro-Kopf-Einkommen
`fatal`	tödliche Verkehrsunfälle pro 10000 Einwohner

Mit Hilfe dieser Daten soll untersucht werden, ob die Anzahl an tödlichen Verkehrsunfällen mit einer Erhöhung der Biersteuer gesenkt werden kann.

(a) Erklären Sie, was man unter einem Panel versteht und worin der Unterschied zu Zeitreihen und Querschnittsdaten besteht. Erläutern Sie zudem den Unterschied zwischen einem balancierten und unbalancierten Panel.

(b) Was versteht man unter *interner Validität* und *externer Validität*?

(c) Eine einfache Regression von `fatal` auf `beertax` produziert folgenden (gekürzten) Output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.85331    0.04357  42.539  < 2e-16
beertax          ???    0.06217   5.865 1.08e-08
```

Berechnen und interpretieren Sie den Koeffizienten von `beertax`.
Bei diesem Modellansatz ist `beertax` der einzige Regressor. Welche Probleme können hierdurch entstehen?

(d) Das Modell soll nun um *fixed effects* bezüglich der Bundesstaaten erweitert werden, um für nicht-beobachtbare Variablen zu kontrollieren. Zusätzlich werden alle restlichen zur Verfügung stehenden Variablen (bis auf `year`) in den Modellansatz aufgenommen. Skizzieren Sie die Modellgleichung (Ansatz mit Absolutglied).

(e) Wie viele Koeffizienten wurden im Modell aus (d) insgesamt geschätzt? Beschreiben Sie eine Möglichkeit wie Sie diese Anzahl deutlich reduzieren und trotzdem bei Bedarf die *fixed effects* für die Bundesstaaten berechnen können.

(f) Die Schätzung des Modells aus (d) liefert folgendes Ergebnis (ohne geschätzte Koeffizienten für die Effekte der Bundesstaaten):

```
                Estimate   Std. Error    t value   Pr(>|t|)
(Intercept) -2.339832e+00 3.573909e+00 -0.6546982 0.51319949
beertax     -3.647409e-01 1.951978e-01 -1.8685703 0.06272676
unemp       -1.970982e-02 1.162998e-02 -1.6947417 0.09123623
drinkage    -3.785145e-02 1.989061e-02 -1.9029807 0.05806864
miles       -3.523850e-06 1.014584e-05 -0.3473197 0.72861207
jailyes     -1.924556e-02 1.402653e-01 -0.1372083 0.89096475
serviceyes  -1.985855e-02 1.620295e-01 -0.1225613 0.90254239
log(income)  6.784052e-01 3.759687e-01  1.8044194 0.07224004
```

Vergleichen Sie den Koeffizienten von `beertax` mit Ihrem Ergebnis aus (b). Durch welche zusätzlichen Variablen können Sie dem Einwand begegnen, dass der (nicht berücksichtigte) Fortschritt in der Automobilsicherheit mit der Variablen `beertax` korreliert sein könnte und dadurch der entsprechende Schätzer verzerrt ist? Skizzieren Sie eine Modellgleichung (Ansatz mit Absolutglied).

(g) Es wird diskutiert, ob die Aufnahme von *fixed effects* für die Zeit in das Modell aus (d) (Ansatz mit Absolutglied) sinnvoll ist. Aus diesem Grund wurden beide Modelle, mit und ohne *fixed effects* für die Jahre, miteinander verglichen und man erhielt folgenden Output:

```
Model 1: fatal ~ beertax + unemp + drinkage + miles + jail + service +
    log(income) + state
Model 2: fatal ~ beertax + unemp + drinkage + miles + jail + service +
    log(income) + state + year
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    280 9.4106
2    274 6.6169  6    2.7936 19.28 < 2.2e-16
```

Welcher Test wurde hier verwendet und auf wie viele Restriktionen wurde getestet? Treffen Sie eine Entscheidung und begründen Sie diese.

(h) In dem Regressionsmodell mit Absolutglied und *fixed effects* sowohl für die Bundesstaaten als auch über die Zeit wurden der Staat Alabama und das Jahr 1982 als Referenz verwendet. Die Effekte der übrigen Jahre und der Bundesstaaten Kalifornien, New Mexico und Mississippi wurden geschätzt mit:

```
 year1983    year1984    year1985    year1986    year1987    year1988
-0.09845142 -0.28642193 -0.37725942 -0.34245597 -0.44125119 -0.52921997


 stateca    statenm    statems
-2.0892305  0.5156790  0.3547913
```

weiter gilt $\widehat{\beta}_0 = -12.62$.
Interpretieren Sie die geschätzten *fixed effects* über die Zeit. Vergleichen Sie die Anzahl der tödliche Verkehrsunfälle in den Bundesstaaten Kalifornien, New Mexico und Mississippi in den Jahren 1982 und 1988 mit denen des Bundesstaates Alabama in den Jahren 1982 und 1988, wobei alle restlichen Regressoren als konstant angenommen werden.