Prof. Dr. Mitja Stadje
Björn Kriesche, M.Sc.

SS 2015
2. 7. 2015

Ökonometrie - Übungsblatt 6

Abgabe am **Montag**, dem 13. 7. **vor** Beginn der Übung

- Auf diesem letzten Blatt können 40 Punkte erzielt werden, wobei 30 Punkte schon 100% entsprechen. Es sind also 10 Bonuspunkte möglich.
- Um die Vorleistung zu bestehen sind demnach 90 von 190 Punkten nötig.
- Bitte bis zum 17.7. im Hochschulportal zur Vorleistung anmelden (sonst ist keine Teilnahme an der Klausur möglich).

Aufgabe 1 (2,5 Punkte)

Wir möchten ein lineares Modell aufstellen um die Abhängigkeit des jährlichen Einkommens von verschiedenen erklärenden Variablen zu betrachten. Wir haben (balancierte) Paneldaten gegeben, die am 31. Dezember 2009 und am 31. Dezember 2012 erhoben wurden. Das lineare Modell soll eine Dummyvariable zur Unterscheidung der Zeitperioden haben und um nicht beobachtbare zeitkonstante Effekte zu eliminieren soll ein Differenzenschätzer verwendet werden. Ist es möglich bzw. sinnvoll als erklärende Variable das Alter der befragten Personen zu verwenden? Begründe ausführlich!

Aufgabe 2 (3,5+5+2,5 Punkte)

Die Datei `crime.txt` enthält Daten aus einer Kriminalitätsstudie, die in den Jahren 1981 bis 1987 für 90 Counties in North Carolina erhoben wurden:

- *county* = Nummer des Counties für das Daten erhoben wurden
- *jahr* = Jahr in dem Daten erhoben wurden
- *rate* = durchschnittliche Anzahl Verbrechen pro Person
- *wverh* = geschätzte Wahrscheinlichkeit, dass man nach einem Verbrechen verhaftet wird
- *wverur* = geschätzte Wahrscheinlichkeit, dass man nach einer Verhaftung verurteilt wird
- $whaft$ = geschätzte Wahrscheinlichkeit, dass man nach einer Verurteilung eine Haftstrafe antreten muss
- *laenge* = durchschnittliche Länge einer Haftstrafe in Tagen
- *pol* = durchschnittliche Anzahl Polizisten pro Einwohner

(a) Lies den Datensatz in **R** ein und erstelle zusätzliche Dummyvariablen $j82, \ldots, j87$, die die verschiedenen Jahre des Beobachtungszeitraums repräsentieren. Betrachte das multivariate lineare Regressionsmodell

$$\log(rate_{it}) = \delta_1 + \delta_2 j82_t + \delta_3 j83_t + \delta_4 j84_t + \delta_5 j85_t + \delta_6 j86_t + \delta_7 j87_t + \beta_1 \log(wverh_{it}) + \beta_2 \log(wverur_{it}) + \beta_3 \log(whaft_{it}) + \beta_4 \log(laenge_{it}) + \beta_5 \log(pol_{it}) + a_i + u_{it},$$

wobei a_i die unbeobachtbaren zeitkonstanten Effekte und u_{it} die unbeobachtbaren zeitabhängigen Effekte (unkorreliert mit den erklärenden Variablen) darstellen. Gib zwei Beispielfaktoren an, die in a_i enthalten sein könnten. Schätze die Regressionsparemeter des Modells in **R** mittels gepoolter MKQ.

(b) Es kann durchaus sein, dass die zeitkonstanten Effekte mit einer oder mehreren erklärenden Variablen korreliert sind. Stelle per Hand das zugehörige Differenzenmodell auf um die zeitkonstanten Effekte zu eliminieren. Berechne den zugehörigen Differenzenschätzer indem du die Parameter des Differenzenmodells in **R** schätzt. Interpretiere alle Parameter, die sich nicht auf die Differenzen der Dummyvariablen beziehen.
Hinweis: durch `lm(y~0+x1+...)` lässt sich ein lineares Modell ohne Intercept aufstellen.

(c) Oft ist es üblich das Differenzenmodell wie folgt zu ändern. Die Differenzen der Dummyvariablen werden ignoriert und stattdessen neue Dummyvariablen für die entsprechenden Zeiträume eingeführt und ein Intercept hinzugefügt (die Anzahl der Regressionsparameter bleibt die gleiche). Weise mit Hilfe von **R** nach, dass diese Änderung die Schätzer für $\beta_1, \ldots, \beta_5$ nicht beeinflusst. Überlege dir einen Vorteil und einen Nachteil, den diese Änderung haben könnte.

Aufgabe 3 (2+4,5+4,5+3,5+2 Punkte)
Die Datei `rental.txt` enthält Mietpreise und andere Daten für amerikanische Universitätsstädte aus den Jahren 1980 und 1990. Es handelt sich um balancierte Paneldaten mit folgenden Informationen:

- *stadt* = Nummer der Stadt für die Daten erhoben wurden
- *jahr* = Jahr in dem Daten erhoben wurden
- *bev* = Gesamtbevölkerung der Stadt im betreffenden Jahr
- *miete* = durchschnittlich gezahlte Miete pro Person in $
- *eink* = durchschnittliches Jahreseinkommen pro Person in $
- *stud* = Anteil Studenten an der Gesamtbevölkerung in %
- $j90$ = Dummyvariable (=1, wenn $jahr = 90$)

Wir betrachten das multivariate lineare Regressionsmodell

$$\log(miete_{it}) = \beta_0 + \delta_0 j90_t + \beta_1 \log(bev_{it}) + \beta_2 \log(eink_{it}) + \beta_3 stud_{it} + a_i + u_{it},$$

wobei a_i die unbeobachtbaren zeitkonstanten Effekte und u_{it} die unbeobachtbaren zeitabhängigen Effekte (unkorreliert mit den erklärenden Variablen) darstellen.

Zunächst wollen wir die Regressionsparameter mit Hilfe des Fixed-Effects-Schätzers bestimmen.

(a) Wir würden gern eine Dummyvariable, die angibt, ob sich die jeweilige Stadt im Norden oder Süden befindet, als weitere unabhängige Variable hinzufügen. Erkläre, warum das nicht so einfach möglich ist. Wie könnte man vorgehen um trotzdem Informationen über die unterschiedlichen Einflüsse der unabhängigen Variablen im Norden und Süden zu erhalten?

(b) Stelle zunächst per Hand die Within-Transformation auf und berechne den Fixed-Effects-Schätzer in **R** ohne die Funktion `lm` zu verwenden. Untersuche, ob die Mieten im Jahr 1990 signifikant höher sind (zum Niveau $\alpha = 0,01$) als im Jahr 1980.

Nun soll ein Random-Effects-Modell betrachtet werden.

(c) Welche zusätzliche Annahme muss gemacht werden, damit der Random-Effects-Schätzer effizient ist? Stelle das dazugehörige quasi-zeittransformierte Modell per Hand auf und schätze den Parameter λ (siehe Hinweise) in **R**. Ermittle den Random-Effects-Schätzer in **R**.

Wir wollen nun noch lernen, wie Paneldaten schneller in **R** bearbeitet werden können. Dazu soll die Funktion `plm` aus dem gleichnamigen Package verwenden werden (siehe Hinweise).

(d) Berechne die Regressionsparameter mit Hilfe von `plm` unter Nutzung gepoolter MKQ, des Differenzenschätzers, des Fixed-Effects-Schätzers und des Random-Effects-Schätzers. (Der von `plm` berechnete Random-Effects-Schätzer weicht leicht vom Ergebnis aus (c) ab, da **R** andere Methoden zur Schätzung von λ verwendet.) Vergleiche die Ergebnisse, insbesondere das Bestimmtheitsmaß, die Vorzeichen der Parameter und die Signifikanz der unabhängigen Variablen.

(e) Es soll ein Hausman-Test auf die Daten angewandt werden. Wie lautet die Nullhypothese und zu welchem Zweck wird der Test eingesetzt? Interpretiere das Ergebnis.

Hinweise:

- Im Random-Effects-Modell kann ein Schätzer $\hat{\lambda}$ für λ wie folgt bestimmt werden. Zuerst wird das Modell mit gepooltem MKQ berechnet und die empirischen Residuen $\hat{v}_{it}$ für $i = 1, \dots, n$ und $t = 1, \dots, T$ bestimmt. Dann gilt, dass

$$\hat{\lambda} = 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_a^2}} \quad \text{wobei} \quad \hat{\sigma}_a^2 = \frac{1}{\frac{nT(T-1)}{2} - (k+1)} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{v}_{it}\hat{v}_{is}$$

  und $\hat{\sigma}_u^2 = s^2 - \hat{\sigma}_a^2$. Dabei ist k wie immer die Anzahl der unabhängigen Variablen und s^2 die geschätzte Varianz der Residuen $\hat{v}_{it}$.

- Um die **R**-Funktion `plm` zu verwenden muss zunächst das package `plm` installiert und aktiviert werden (siehe Blatt 3, Aufgabe 4). Es sei `daten` ein data frame, welcher balancierte Paneldaten enthält und in der ersten Spalte stets einen Indikator für die verschiedenen Elemente des Querschnitts und in der zweiten Spalte einen Indikator für den aktuellen Zeitpunkt im Panel enthält (so wie die Daten in `rental.txt`). Dann kann ein lineares Modell durch den Befehl `panel = plm(y~1+x1+x2+x3, data=daten, model="...")` angepasst werden, wobei die Option `model` folgende Werte annehmen kann: `pooling` (gepoolte MKQ), `fd` (Differenzenschätzer), `within` (Fixed-Effects-Schätzer) und `random` (Random-Effects-Schätzer). Seien `panel1` und `panel2` zwei solche Paneldaten-Modelle mit Fixed-Effects-Schätzer und Random-Effects-Schätzer, dann kann mittels `phtest(panel1, panel2)` ein Hausman-Test durchgeführt werden.

Aufgabe 4 (1+1+2+4+2 Punkt)

Die folgende Aaufgabe soll ohne die Verwendung von **R** bearbeitet werden. Mit Hilfe eines einfachen linearen Modells und den unten gegebenen Daten soll untersucht werden, ob der Besitz eines eigenen PCs einen Einfluss auf den erzielten Notendurchschnitt im Studium hat. Wir betrachten dazu das lineare Modell $avg = \beta_0 + \beta_1 pc + u$.

Notendurchschnitt (*avg*)	3,3	2,1	3,0	1,9	1,4	2,1	2,6	2,2	2,4	1,3	1,2
Besitz eines eigenen PCs (*pc*)	0	0	0	0	1	1	1	1	1	1	1
PC-Stipendium erhalten (*grant*)	0	0	0	0	0	0	1	1	1	1	1

(a) Würdest du in diesem Modell Endogenität erwarten? Begründe!

(b) Ein Vorschlag für eine mögliche Instrumentvariable wäre das monatliche Einkommen der Studenten. Wäre dies eine gute Wahl für eine Instrumentvariable? Begründe!

(c) Vor ein paar Jahren hat die Universität Stipendien vergeben, damit sich Studenten einen eigenen PC kaufen können. Dabei wurden die unterstützten Studenten rein zufällig ausgewählt. Würdest du vermuten, dass die Dummyvariable, die angibt, ob der jeweilige Student ein Stipendium erhalten hat, eine geeignete Instrumentvariable ist? Begründe.

(d) Wir möchten die in (c) beschriebene Variable *grant* als Instrumentvariable für *pc* nutzen. Überprüfe, ob diese tatsächlich einen Einfluss auf den Besitz eines eigenen PCs hat, indem du das lineare Modell $pc = \pi_0 + \pi_1 grant + v$ aufstellst, die Regressionskoeffizienten schätzt und zum Niveau $\alpha = 0,05$ testest, ob sich π_1 signifikant von null unterscheidet.

(e) Berechne den Instrumentvariablenschätzer $\hat{\beta}_{IV}$ für β_1 und den daraus resultierenden Schätzer $\hat{\beta}_0$ für β_0.