Prof. Dr. Peter von der Lippe Jan 2011 (Version 1 war Febr. 2010)

Ein Beispiel für die Anwendung von Methoden für Panel-Daten (Version 2)

Im Folgenden beschreiben wir Verfahren und Erfahrungen mit EViews im Zusammenhang mit einer empirischen Studie über Spillover Effekte von Auslandskapital in chinesischen Firmen.[1] Es ist ein typischer Fall eines Panels mit **large N - small T**.[2] Zunächst sind die Daten, die als Excel Tabelle vorliegen, in einen EViews workfile einzulesen, was nicht ganz einfach ist und in Abschn. 1 beschrieben wird. Wenn dann EViews die Daten und ihre Struktur bekannt sind kann man ein Pool-Objekt definieren (Abschn. 2). Wir zeigen dann (Abschn. 3) wie man Regressionsfunktionen schätzen kann und (Abschn. 4) Entscheidungen zwischen verschiedenen Modellvarianten treffen kann.

1. Import von Excel Daten in EViews um ein Workfile mit Panel Daten zu eröffnen:[3]

Die Excel Datei unseres Beispiels enthält Angaben in Spalten A bis M für N = 2947 Firmen (a, b, …) und T = 5 Jahre (von 2002 bis 2006). Es muss zunächst Art und Name des workfiles festgelegt werden

Obere Befehlsleiste **File-New-Workfile**, dann bei Abfrage nicht balanced panel wählen weil EViews nur eine Datenanordnung der N Objekte a, b, … bei T Daten, etwa den Jahren 2002, 2003, ... der folgenden Art a-02, a-03 ,… , b-02, b-03, … erlaubt (= ***stacked by cross section*** Anordnung). Wenn in Excel die Daten nicht so angeordnet sind, sondern ***stacked by date***, d.h. in einer Spalte a-02, b-02, c-02, …, a-03, b-03, c-03, … untereinander ist es sinnvoll wie folgt vorzugehen: Bei Workfile (wf) nicht balanced panel, sondern **File-New-Workfile**, **unstructured/undated** wählen, dem wf einen Namen geben und für number of observations die Zahl NT angeben Dann **Proc-Import- Read TextLotusExcel**, es kommt ein Fenster **open** das es erlaubt, die Excel Datei mit den Daten suchen; es ist auch der sheet name anzugeben (Name der Seite). "Open" heißt nicht, dass die Excel Datei jetzt sichtbar geöffnet wird, sondern nur dass EViews hier Daten sucht. Im Gegenteil, die Excel Datei selber darf nicht auf dem Bildschirm erscheinen und in diesem Sinne "geöffnet" sein wenn aus ihr die Daten herausgelesen werden sollen und man wird sie im Folgenden auch nicht zu sehen bekommen, so dass es gut ist, sich die Struktur der Tabelle zu notieren.

Nach anklicken von open ist anzugeben ob die Daten in Zeilen (= by series – series in rows) oder Spalten (= by observations – series in columns) – so im oben beschriebenen Fall - angeordnet sind und es ist der Beginn der ersten Datenreihe anzugeben (A2, nicht A1 angeben, EViews liest die erste Zeile korrekt als Variablennamen, gibt man A1 an, dann erzeugt EViews statt dessen die Variablen series01, serie02 usw.).

Achtung: EViews liest von links nach rechts alle Spalten, auch wenn sie leer sind.[4] Es ist also wichtig, die Struktur der Excel Tabelle genau zu kennen. Am besten sind in den ersten beiden Spalten A und B firm und year. Sie müssen als Variablen mit eingelesen werden und später bei einer Abfrage eingegeben werden. Man kann in einer Liste auch neue Namen (andere als in der Excel Tabelle) für die Variablen geben, die dann so im workfile (wf) benutzt werden. Es st wichtig, dass man sich hier nicht vertut und *für jede Spalte* (auch wenn sie leer ist oder später als Variable nicht gebraucht wird) einen Variablennamen vergibt, sonst erscheint später unter dem

---

[1] Dabei handelt es um die Auswertung von Daten des Kollegen Markus Taube.

[2] Bei dem hier beispielhaft behandelten workfile (wf) mit dem Namen "MT1" geht es um N = 2947 Firmen aus dem Verarbeitenden Gewerbe (in Shanghai) und T=5 Jahre, nämlich die Jahre 2002 bis 2006 (die Variable year ist auch Gegenstand der Auswertung), es ist also NT = 14735. Wie an den screen shoots später zu sehen sein wird ist das panel jedoch nicht vollständig "balanced" so dass insgesamt nur 14719 Beobachtungen zur Verfügung standen.

[3] Diesen Hinweis verdanke ich Herrn Jens Mehrhoff, Deutsche Bundesbank.

[4] EViews erkennt also die Variablen und gibt sie in der Reihenfolge wie bei Excel (also z.B. stacked by date) in den wf ein.

Namen x und y die Spalte G und H statt F und G. Nachträgliches Umbenennen oder einlesen einer vergessenen Variable ist schwierig.

Es muss jetzt noch festgelegt werden wie sie als Querschnitt und Zeitreihe zu verstehen sind. Dazu **Proc-Structure/resize…- Dated Panel** anklicken und im Dialogfenster "cross section identifier" als "Identifier-Variable "firm" angeben[5] und unter "Data series" die Zeitvariable (z.B. year) angeben. Damit ist die Excel Tabelle in einem Pool und man kann Panel-Methoden anwenden. Bei frequency kann man die Voreinstellung auto detect stehen lassen. Ähnlich beim ersten und letzten Datum. Man kann dort aber auch die Jahreszahlen (im Beispiel 2002 und 2005) angeben

Es empfiehlt sich dann mit object – new object – group eine Gruppe zu definieren und zu benennen, die alle Variablen enthält und sich dann mit view-spreadsheet anzusehen ob die Variablen richtig eingelesen sind. Man sollte diese **Datenprüfung unbedingt vornehmen**, denn es ist sehr schwer nachträglich an den Daten im wf noch etwas zu ändern.

2. Pool-Objekt

Man kann auch mit object – new object – pool ein sog. Pool Objekt bilden und benennen. Dabei wird man aufgefordert "cross section identifier", also Namen für die Einheiten in den Querschnitten anzugeben, was bei N = 2947 Firmen natürlich keinen Sinn macht, d.h. es ist nicht sinnvoll, von dieser Möglichkeit im (hier vorliegenden) Fall von large N – small T Gebrauch zu machen.

3. Erste Berechnungen: Poolregressionen für alle oder einige Jahre (Perioden)

Die Definition einer alle Variablen umfassenden Gruppe empfiehlt sich auch bei der Durchführung von Schätzungen weil dann beim Aufruf von **group - proc - make equation** bereits alle Variablen aufgelistet sind, die man als Regressoren gebrauchen könnte. Im equation Menü (wird erreicht über Proc. oder mit Estimate im Arbeitsfenster) erscheint dann neben Specification auch Panel-Options womit man dann z.B. fixed effects und random effects Modelle bilden kann.

Alternativ kann auch man über Quick (Hauptmenü, oberen Befehlsleiste) – Estimate equation vorgehen und auch so die Panel Optionen erhalten.

a) Poolregression (alle Jahre)

Rechnet man einfach mit den - wie oben beschrieben - importierten Variablen, dann ist das eine pool-regression, das restriktivste Modell mit gleichen Regressions-Koeffizienten α und β für alle Einheiten (firms) und alle Perioden (years), weshalb man auch von einem restringierten Modell sprechen kann. Es läuft darauf hinaus, einfach alle NT Beobachtungen für jede Variable in einen Topf zu werfen und nicht bei den cross sections nach Jahren zu differenzieren. Ob diese Vorgehensweise gerechtfertigt ist hängt vom im Folgenden beschriebenen F Test ab.

b) getrennte Querschnitte

Will man sich die cross section Daten (also alle N firms) für nur eine Periode etwa 2006 ansehen und z.B. hierfür eine Regressionsfunktion rechnen, dann sollte man das sample auf 2006 2006 (Anfang und Ende angeben) setzen. Man kann dies für jedes Jahr tun und erhält dabei auch jeweils die Sum squared resid (SSR) für die entsprechende Regressionsgleichung. Die Summe der SSR über alle T Jahre ist die unrestricted SSR ($S_{\hat{u}\hat{u}}$) im Unterschied zur restricted SSR ($S^0_{\hat{u}\hat{u}}$). Ist die F verteilte Prüfgröße

[5] Es war deshalb sehr wichtig, auch diese Spalte A im Excel-Spreadsheet als Variable mit einzulesen (genauso wie das Jahr [year] als Spalte B).

$$F = \frac{(S^0_{\hat{u}\hat{u}} - S_{\hat{u}\hat{u}})/k(T-1)}{S_{\hat{u}\hat{u}}/(N-k)T} \sim F_{k(T-1),(N-k)T}$$

(wobei im Modell k Regressionskoeffizienten zu schätzen sind) kleiner als der kritische Wert, dann ist es gerechtfertigt die Daten zu "poolen" und im Sinne von Teil a dieses Abschnitts zu verfahren.

4. Ablauf der Modellanalyse

Es empfiehlt sich, zunächst mit einer Schätzung eines Modells mit "random effects" zu beginnen. Die entsprechende Option ist unter "Panel Options" zu wählen. Man kann dann mit **view – fixed/random effects testing – correlated random effects** die Angemessenheit dieses Modells mit dem Hausman Test (bzw. Hausman Wu Test) überprüfen (man kann sich dabei wie immer an dem prob. value orientieren

---

Ablaufschema

Stufe1: H_0: Störgrößen nicht mit Regressoren korreliert (Hausman Test: Ausschließen von Endogenität, bei E-Views : "correlated random effects")

H_0 abgelehnt: random effects nicht angebracht, mit **fixed** effects rechnen

H_0 angenommen: **random** effects modell ist angepasst

Stufe2: H_0: Prüfen ob (H_0) $\alpha_1 = ... = \alpha_T = \alpha$ (ungleiche intercepts sind redundant) mit (E-Views) "Redundant fixed effects" Test"

H_0 angenommen: fixed effects sind redundant FE-modell ist nicht angepasst

H_0 abgelehnt: fixed effects tatsächl. unterschiedl. also FE model ist angepasst

wenn Modell gewählt robuste Varianzschätzung vornehmen (z. B. nach White)

pool regression (mit OLS) rechnen (d.h. keine cross-section oder time specific coeff.)

---

1) Vergleich fixed vs random[6]

Der **Hausman Test** vergleicht einen konsistenten aber nicht erwartungstreuen Schätzer etwa den Instrumentvariablen (IV) Schätzer mit einem inkonsistenten aber effizienten Schätzer (etwas den OLS Schätzer). Er wird v.a. angewandt um die bei OLS geforderte Exogenität der Regressoren zu prüfen. Verwerfen von H_0 - wenn die χ^2 verteilte Prüfgröße größer als der kritische Wert ist - bedeutet dass die Variablen nicht exogen sind (also mit u kontemporär korreliert sind).

Im Zusammenhang fixed vs random gilt: random effects verlangt, dass die individuellen Effekte α_i für die Einheiten ui = 1,…, N unkorreliert sind mit den Regressoren. Im Fall von omitted variables kann das evtl. nicht erfüllt sein. Der Test basiert darauf, dass gilt

	unter Geltung von H_0*	wenn H_0 nicht gilt
OLS**	konsistent	konsistent
GLS	konsistent	nicht konsistent

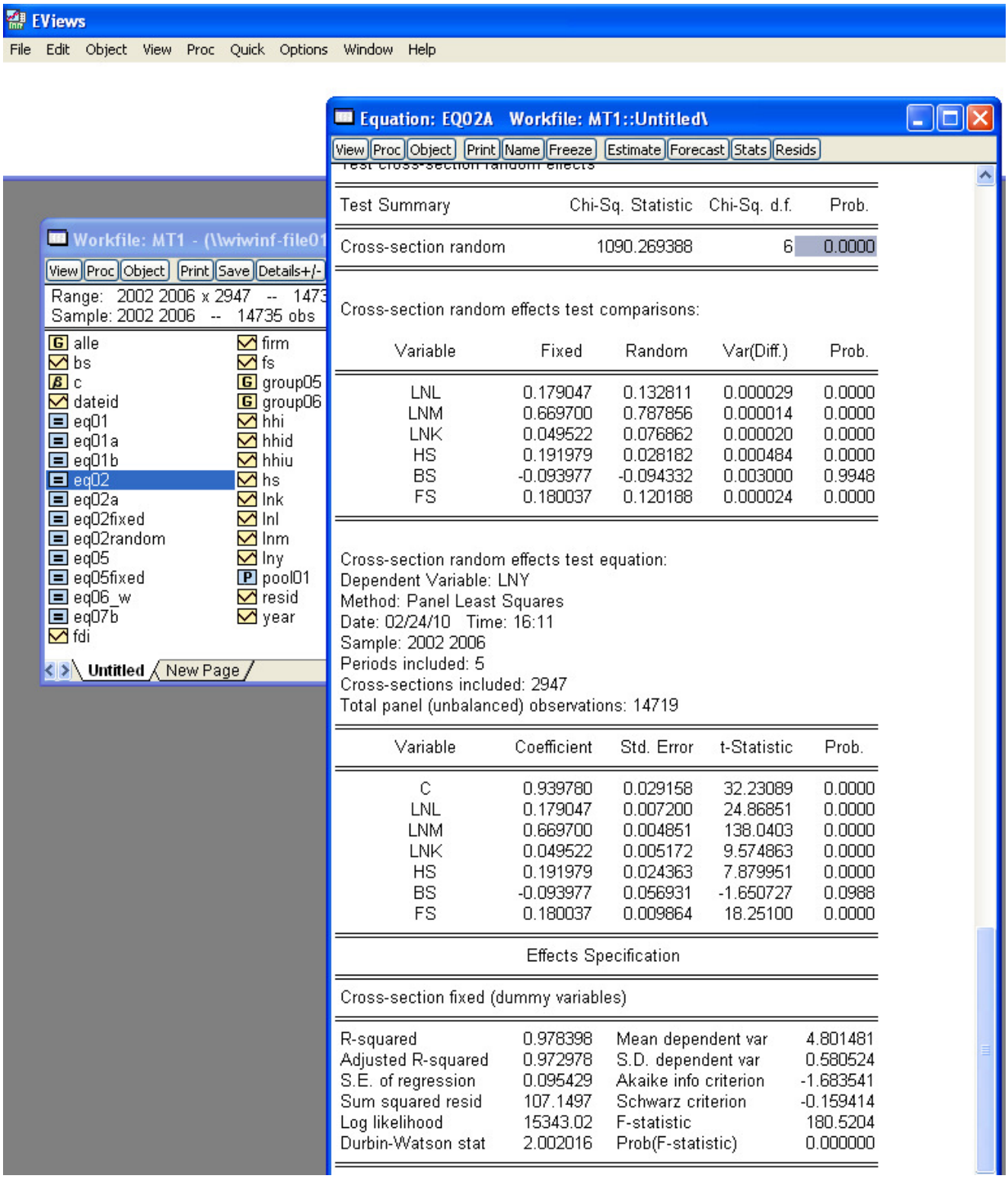
* d.h. unkorrelierte Effekte (anzunehmen bei random effects)
** im Rahmen des LSDV Modells

[6] zu einer Darstellung des Hausmann Wu Tests bei dieser Anwendung vgl. W. H. Greene, Econometric Analysis, 4. ed., p. 576f.

Ablehnung von H_0 bedeutet also dass man mit dem random effects model die cross section effects nicht konsistent schätzt und das fixed effects model die bessere Wahl ist.

Die Situation nach Durchführung des Hausman Test zeigt der screenshot der Abb. 1. Dem oberen Teil ist das Testergebnis zu entnehmen (H_0 ist danach abzulehnen) und es folgen dann die Schätzwerte für die Parameter nach beiden Modellen (fixed und random) und im unteren Teil des fixed effects Modell mit den üblichen Angaben auch zur Güte der Schätzung.

Abbildung 1



Man erkennt z.B. dass der Regressor BS (hier "backward spillover") in der Produktionsfunktion

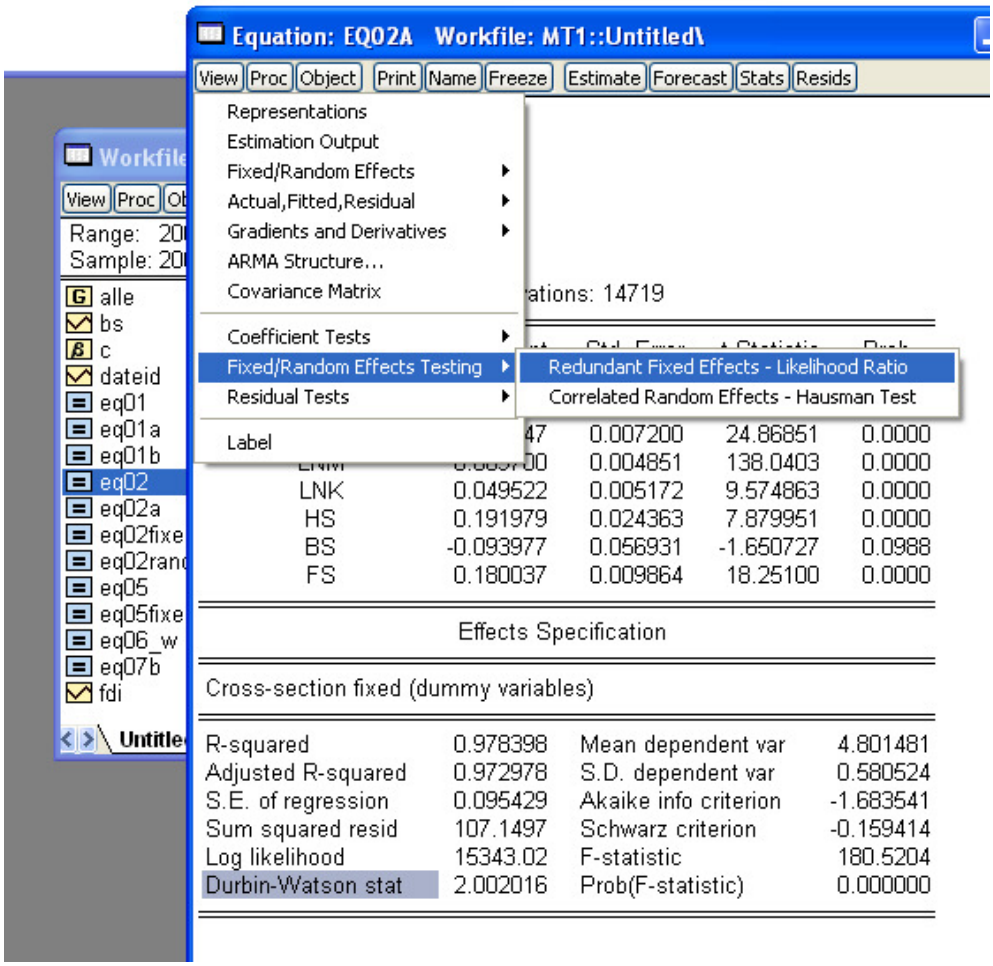
$$\ln Y_{it} = \alpha + \beta_1 \ln L_{it} + \beta_2 \ln M_{it} + \beta_3 \ln K_{it} + \beta_4 HS_t + \beta_5 BS_t + \beta_6 FS_{it} + u_{it}$$

die hier geschätzt wurde nicht signifikant auf dem 5% Niveau ist.

2. redundante Effekte

Im nächsten Schritt ist zwischen dem jetzt zu schätzenden fixed effect Modell und der Poolregression (gleiche Koeffizienten α für alle N Firmen im Querschnitt und keine period dummies) zu entscheiden. Dazu ist beim geschätzten fixed effects Modell zu wählen **view – fixed/random effects testing – redundant fixed effects testing** wie das in Abb. 2 gezeigt wird.

Abbildung 2



Im vorliegenden Fall sind nur cross section effects (also verschiedene α Koeffizienten für die N Firmen gewählt worden. Man sieht das an der Auflistung unter "Effects Specification". Man kann auch "period fixed effects" (also für die T Perioden) wählen indem man unter "period" nicht none sondern fixed effects wählt. Bei der Schätzung gem. **Abb. 3** ist das geschehen.

Das Testergebnis (H_0 Effekte sind nicht "redundant") wird dargestellt in Form einer Tabelle mit den Ergebnissen von F-Test, zum Beispiel wie folgt:

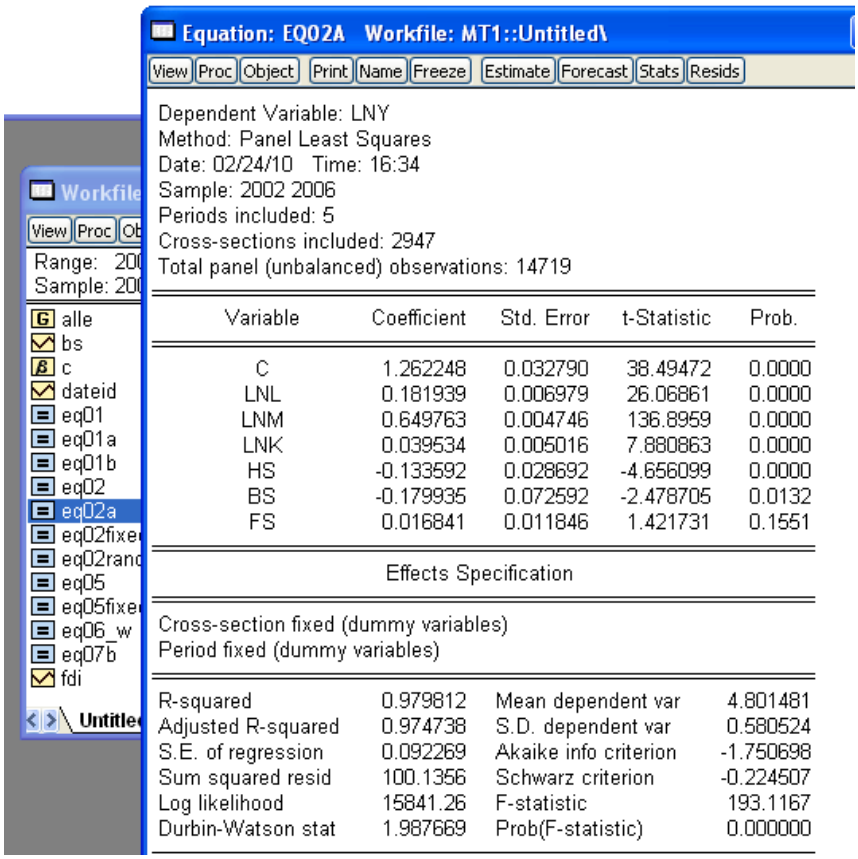
Redundant Fixed Effects Tests
Equation: EQ02A
Test cross-section and period fixed effects

Effects Test	Statistic	d.f.	Prob.
Cross-section F	4.360909	(2946,11762)	0.0000
Cross-section Chi-square	10866.27032	2946	0.0000
Period F	205.968421	(4,11762)	0.0000
Period Chi-square	996.491797	4	0.0000
Cross-Section/Period F	4.539727	(2950,11762)	0.0000
Cross-Section/Period Chi-square	11188.65809	2950	0.0000

Dabei wird zunächst die Berechtigung von nur cross-section und nur period effects und dann von beiden gemeinsam getestet. Wird H_0 verworfen sind die Effekte relevant und *nicht* "redundant".

Abb. 3 zeigt ein fixed effect Modell in dem beide Arten von Effekten unterstellt wurden. Der Regressor FS (forward spillover) ist offenbar nicht signifikant und die Schätzqualität ist ansonsten gemessen an $\overline{R}^2$, DW usw. nicht schlecht.

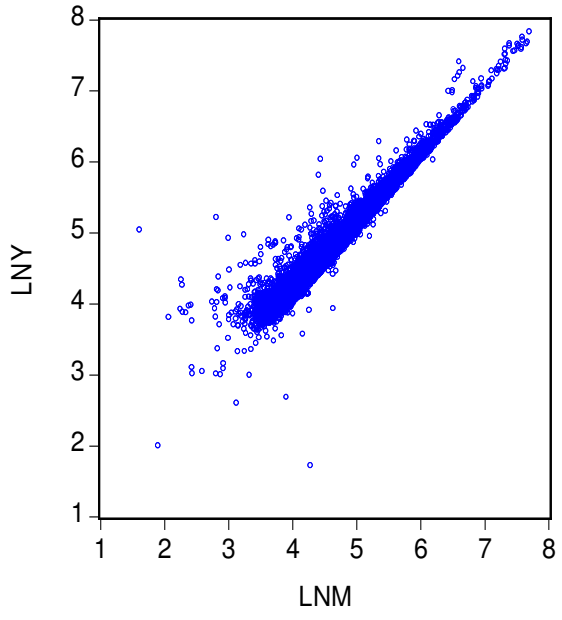
Abbildung 3



Die cross-section specific coefficients (Liste von 2947 zum Intercept zu addierenden positiven oder negativen Koeffizienten) erhält man unter view; sie ist aber in der Regel nicht von Interesse. Anders dagegen die entsprechende Liste der time dummies bei kleinem T.
Man erhält sie mit **view – fixed/random effects – period fixed effects**. Im Beispiel des screen shots erhält man damit die folgende Tabelle.

	DATEID	Effect
1	2002-01-01	-0.026855
2	2003-01-01	0.002319
3	2004-01-01	0.051392
4	2005-01-01	0.027832
5	2006-01-01	0.042969

Es mag auch nützlich sein, sich die Plausibilität mancher Berechnung durch eine Graphik klar zu machen. Die Abb. unten zeigt ein Streuungsdiagramm für alle 14719 observations.



Einführung von Dummies für die einzelnen Perioden

Anstelle einer Schätzung mit period fixed dummies kann man auch für die einzelnen Jahre Dummy Variablen (z.B. bei t = 5 Jahren 2002 bis 2006 wie hier im Beispiel die Verabredung von 4 dummies t3, …, t6) definieren und diese dann in die Regressionsgleichung eingeben.

Dazu ist wie folgt vorzugehen: Man geht mit **quick - generate series** in das Gleichungsfenster und schreibt z.B.

t3 = 0*(year<>2003)+1*(year=2003) und t4, t5, t6 analog.