Methoden-Workshop

Leading House „Economics of Education“/SKBF

Ben Jann und Rudi Farys

Universität Zürich, Rämistrasse 71, Raum KOL-F-123, 1.–3. Februar 2016

Analyse von Paneldaten

Übersicht

- Paneldaten
- Error-Components-Model
- Pooled-OLS und Between-Schätzer
- Fixed-Effects, First-Differences, LSDV
- Random-Effects-Modell
- FE versus RE und einige weitere Aspekte
- Hybrid-Modelle
- Dynamische Panelschätzer
- Logistische Regression mit Paneldaten

Einige Literaturhinweise

- Allison, Paul D. (2009). Fixed Effects Regression Models. Thousand Oaks, CA: Sage.
- Brüderl, Josef (2010). Kausalanalyse mit Paneldaten. S. 963-994 in: Henning Best und Christof Wolf (Eds.), Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag.
- Kapitel 8, 9, 18 aus: Cameron, A. Colin, Pravin K. Trivedi (2009). Microeconometrics Using Stata. College Station, TX: Stata Press.
- Halaby, Charles N. (2004). Panel Models in Sociological Research: Theory into Practice. Annual Review of Sociology 30:507-544.
- Kapitel 12 aus: Johnston, Jack, John DiNardo (1997). Econometric Methods. 4th edition. New York: McGraw-Hill.
- Schunck, Reinhard (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. The Stata Journal 13(1): 65-76.
- Kapitel 21/23 aus: Wooldridge, Jeffrey M. (2003). Introductory Econometrics. A Modern Approach. Mason, OH: Thomson South-Western.
- Kapitel 10, 11 und 15.8 aus: Wooldridge, Jeffrey M. (2010). Econometric Analysis of Cross Section and Panel Data (Second Edition). Cambridge, MA: The MIT Press.

Paneldaten

- Paneldaten sind Daten, bei denen pro Untersuchungseinheit mehrere Messungen zu verschiedenen Zeitpunkten vorliegen.
  - N = Anzahl Untersuchungseinheiten
  - T_i = Anzahl Messungen für Untersuchungseinheit i, $i = 1, \ldots, N$
    - ★ balanciertes Panel: $T_i = T$ für alle i
    - ★ unbalanciertes Panel: $T_i \neq T_j$ für mindestens ein Paar $i \neq j$
- Paneldaten werden i.d.R. mit Hilfe von Panelsurveys gewonnen, bei denen eine gleich bleibende Stichprobe von Personen z.B. jährlich befragt wird.
- Beispiele:
  - Schweizer Haushaltspanel (SHP)
  - Deutsches Sozio-oekonomisches Panel (SOEP)

Paneldaten: Datenstruktur

- Long-Format

ID	T	Y	X	Z	$\ldots$
1	1	y_{11}	x_{11}	z_{11}	$\ldots$
1	2	y_{12}	x_{12}	z_{12}	
1	3	y_{13}	x_{13}	z_{13}	
$\vdots$					
2	1	y_{21}	x_{21}	z_{21}	
2	2	y_{22}	x_{22}	z_{22}	
$\vdots$					
N	1	y_{N1}	x_{N1}	z_{N1}	
N	2	y_{N2}	x_{N2}	z_{N2}	
$\vdots$					

Paneldaten: Datenstruktur

- Wide-Format

ID	Y_1	Y_2	$\ldots$	X_1	X_1	$\ldots$	Z_1	Z_2	$\ldots$
1	y_{11}	y_{12}	$\ldots$	x_{11}	x_{12}	$\ldots$	z_{11}	z_{12}	$\ldots$
2	y_{21}	y_{22}	$\ldots$	x_{21}	x_{22}	$\ldots$	z_{21}	z_{22}	$\ldots$
$\vdots$									
N	y_{N1}	y_{N2}	$\ldots$	x_{N1}	x_{N2}	$\ldots$	z_{N1}	z_{N2}	$\ldots$

- Für die Analyse werden i.d.R. Daten im Long-Format benötigt.
- Wechsel von „wide“ zu „long“ und umgekehrt in Stata: `reshape`

Paneldaten: Vorteile und Nachteile

- Vorteile
  - Paneldaten sind informativer als Querschnittsdaten, da Veränderungen der Variablen über die Zeit gemessen werden können.
    - ★ Dies erlaubt beispielsweise die Trennung von Alters- und Kohorteneffekten.
  - Paneldaten geben Auskunft über die zeitliche Abfolge von Ereignissen, was für die Analyse von Kausalzusammenhängen wichtig ist.
  - Mit Paneldaten kann individuelle unbeobachtete Heterogenität kontrolliert werden, die bei der Analyse nicht-experimenteller Daten ein grosses Problem darstellt.
- Nachteile
  - Panel-Attrition: Durch systematische Ausfälle wird eine Panelstichprobe u.U. über die Zeit selektiver.
  - Reaktivität: Panel-Konditionierung, Lerneffekte etc.

Beispiel: Die Heiratsprämie

- Aus Querschnittsstudien ist bekannt, dass verheiratete Männer mehr verdienen als unverheiratete Männer gleichen Alters (die sog. „Heiratsprämie“).
- Hat die Heirat einen kausalen Effekt auf den Lohn? Wohl kaum. Es handelt sich wahrscheinlich um ein Artefakt aufgrund von Selbstselektion oder umgekehrter Kausalität.
  - „Fähigere“ Männer heiraten eher und haben gleichzeitig einen höheren Lohn.
  - Hohes Einkommen erhöht die Chancen auf dem Heiratsmarkt.
- Alternative Erklärungen wären, dass sich Männer besser bezahlte Jobs suchen, nachdem sie geheiratet haben, oder dass verheiratete Männer tatsächlich besser Arbeitsmarktchancen haben (z.B. höhere Lohnangebote erhalten oder eher befördert werden).
- Mit Querschnittsdaten kann das alles jedoch nur schwer auseinandergehalten werden.

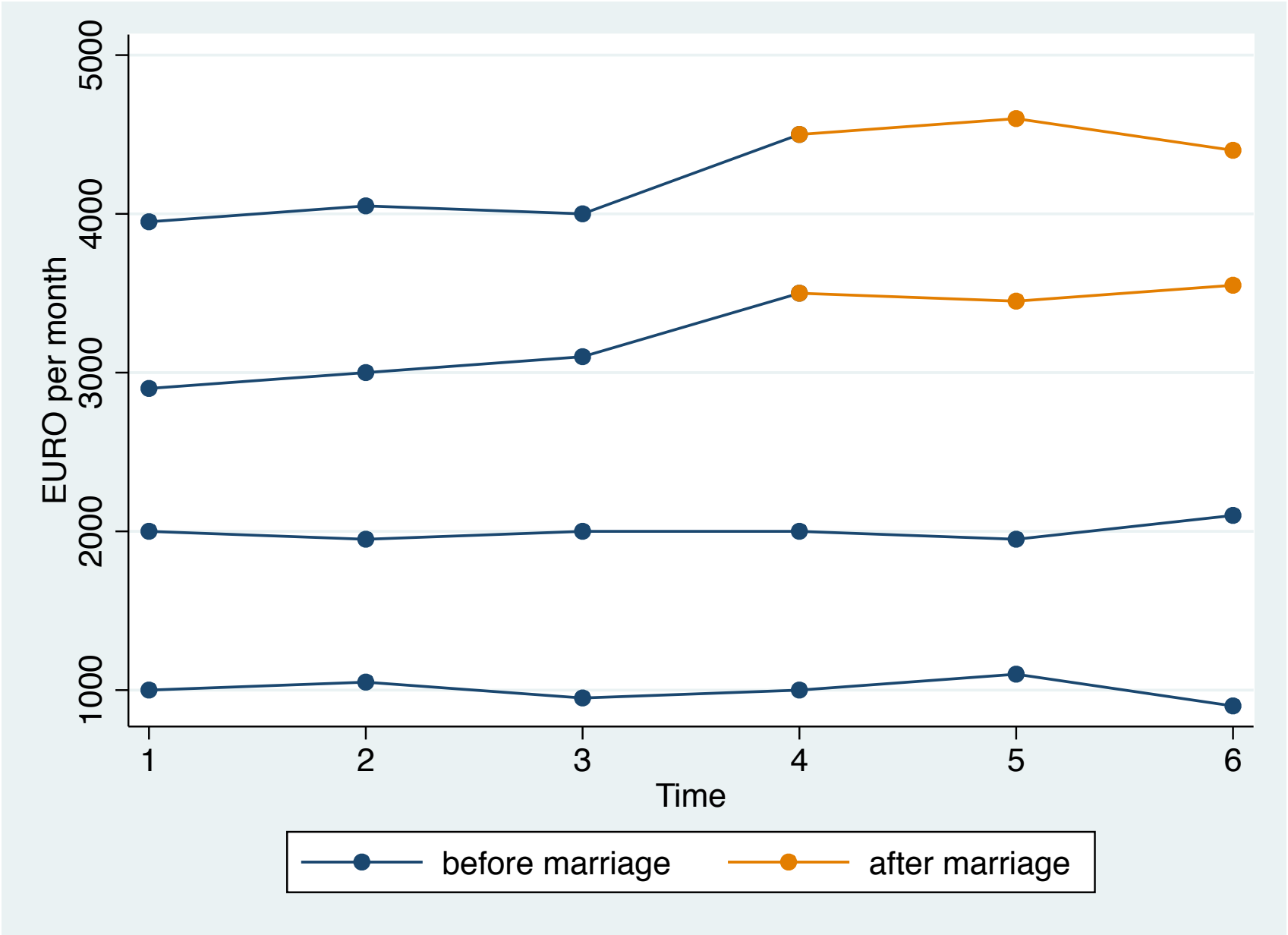
Beispiel: Die Heiratsprämie

```
. use "Panelanalyse.dta", clear
. sort id time
. list id time wage marr, separator(6)
```

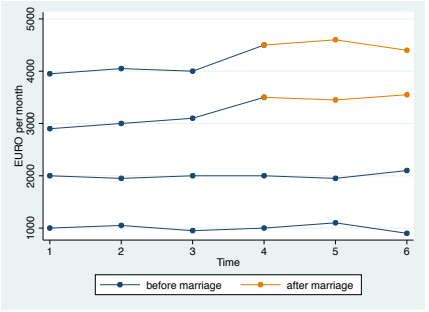
	id	time	wage	marr
1.	1	1	1000	0
2.	1	2	1050	0
3.	1	3	950	0
4.	1	4	1000	0
5.	1	5	1100	0
6.	1	6	900	0
7.	2	1	2000	0
8.	2	2	1950	0
9.	2	3	2000	0
10.	2	4	2000	0
11.	2	5	1950	0
12.	2	6	2100	0
13.	3	1	2900	0
14.	3	2	3000	0
15.	3	3	3100	0
16.	3	4	3500	1
17.	3	5	3450	1
18.	3	6	3550	1
19.	4	1	3950	0
20.	4	2	4050	0
21.	4	3	4000	0
22.	4	4	4500	1
23.	4	5	4600	1
24.	4	6	4400	1

Hypothetische Daten aus:
Brüderl, Josef (2005). Panel Data Analysis.
http://www2.sowi.uni-mannheim.de/lsssm/lehre.html

Beispiel: Die Heiratsprämie



Beispiel: Die Heiratsprämie




```
. two (scatter wage time if !(marr==1 & time>4), c(L)) ///
>     (scatter wage time if marr==1, c(L) psty(p4)) ///
>     , legend(order(1 "before marriage" 2 "after marriage"))
```

Beispiel: Die Heiratsprämie

- Querschnittseffekt zu Zeitpunkt 4

```
. regress wage marr if time==4, noheader
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	2500	707.1068	3.54	0.072	-542.4349	5542.435
_cons	1500	500	3.00	0.095	-651.3264	3651.326

- Effekt für Heiratende zwischen Zeitpunkt 3 und 4

```
. sort id time
. by id: generate byte married = marr[_N]
. regress wage marr if married & inlist(time,3,4), noheader
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	450	672.6812	0.67	0.572	-2444.314	3344.314
_cons	3550	475.6574	7.46	0.017	1503.411	5596.589

Beispiel: Die Heiratsprämie

- Das Beispiel zeigt, dass es wichtig ist, folgende beiden Betrachtungsweisen auseinander zu halten:
  - Betrachtung inter-individueller Differenzen in Y zwischen Untersuchungseinheiten mit unterschiedlichen X-Werten (Between-Betrachtung).
  - Betrachtung der intra-individuellen Veränderung von Y, wenn sich X verändert (Within-Betrachtung).
- Bei der Between-Betrachtung sind zwei Effekte miteinander vermischt.
  - Effekt der Selbstselektion bzw. unbeobachtete Heterogenität: Männer mit hohem Einkommen sind gleichzeitig diejenigen, die eher heiraten.
  - Kausaler Effekt des Heiratens auf das Einkommen.
- Mit Hilfe der Within-Betrachtung, können die beiden Effekte getrennt werden.

Error-Components-Modell

- Der Effekt unbeobachteter Heterogenität kann in einem Regressionsmodell durch Aufnahme eines personenspezifischen Fehlerterms berücksichtigt werden.
- Im einfachsten Fall lautet das Modell dann

$$Y_{it} = X'_{it}\beta + \alpha_i + \epsilon_{it}$$

  wobei α_i der personenspezifische „Fehler“ ist.
- Durch α_i werden die unterschiedlichen „Niveaus“ von Y für die verschiedenen Personen abgebildet.
- Beim Fehlerterm ϵ_{it} handelt es sich um weisses Rauschen mit den üblichen Annahmen. Insb. wird angenommen, dass ϵ_{it} nicht mit X_{it} und α_i zusammenhängt.

Error-Components-Modell

- Verschiedene Ansätze zur Schätzung des Error-Components-Modells:
  - Pooled-OLS
  - Fixed-Effects / First-Differences / LSDV
  - Random-Effects
- Die Verfahren unterscheiden sich hinsichtlich der Annahmen, die über α_i gemacht werden.
  - Random-Effects-Annahme (RE): Die personenspezifischen Fehler α_i hängen nicht mit X_{it} zusammen.
  - Fixed-Effects-Annahme (FE): Die personenspezifischen Fehler α_i hängen potentiell mit X_{it} zusammen.

Pooled-OLS (POLS)

- Ist die RE-Annahme gegeben, kann β ganz einfach durch eine OLS-Regression über alle Messungen konsistent geschätzt werden, also durch Anwendung von OLS auf

$$Y_{it} = X'_{it}\beta + \upsilon_{it}$$

  wobei implizit $\upsilon_{it} = \alpha_i + \epsilon_{it}$.

  - Auch eine OLS-Regression aufgrund von Querschnittsdaten würde in diesem Fall einen konsistenten Schätzer liefern, die gepoolte Schätzung ist jedoch effizienter.

- Da α_i implizit in υ_{it} enthalten ist, entsteht Autokorrelation.
  - Dies sollte bei der Berechnung von Standardfehlern berücksichtigt werden (die Daten sind „geklumpt“ nach Untersuchungseinheiten).
  - Da OLS die Autokorrelation nicht berücksichtigt, existieren effizientere Schätzverfahren.

Pooled-OLS (POLS)

- Falls die RE-Annahme verletzt ist, ist der POLS-Schätzer verzerrt.
- Der Schätzer ist jedoch etwas weniger verzerrt als ein OLS-Schätzer aufgrund von Querschnittsdaten, da der POLS-Schätzer teilweise auf der Within-Betrachtung beruht.
- Beispiel Heiratsprämie: POLS-Schätzer über alle vier Zeitpunkte

```
. regress wage marr
```

Source	SS	df	MS
Model	15125000	1	15125000
Residual	22090000	22	1004090.91
Total	37215000	23	1618043.48

Number of obs	24
F(1, 22)	15.06
Prob > F	0.0008
R-squared	0.4064
Adj R-squared	0.3794
Root MSE	1002

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	1833.333	472.3678	3.88	0.001	853.7025	2812.964
_cons	2166.667	236.1839	9.17	0.000	1676.851	2656.482

Between-Schätzer

- Wie gesagt beruht der POLS-Schätzer auf Variation *zwischen* Personen sowie auf Variation „*innerhalb*“ von Personen über die Zeit.
- Einen reinen *Between-Schätzer* für Paneldaten, bei dem sämtliche Within-Variation ausgeblendet wird, erhält man durch Anwendung von OLS auf

$$\overline{Y}_i = \overline{X}_i'\beta + \alpha_i + \overline{\epsilon}_i$$

- Ähnlich wie der POLS-Schätzer, ist der Between-Schätzer nur dann konsistent, wenn die RE-Annahme zutrifft.
- Beispiel Heiratsprämie: Between-Schätzer

```
. sort id
. by id: egen m_wage = mean(wage)
. by id: egen m_marr = mean(marr)
. regress m_wage m_marr, noheader
```

m_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
m_marr	4500	426.4014	10.55	0.000	3615.698	5384.302
_cons	1500	150.7557	9.95	0.000	1187.352	1812.648

Within-Schätzer / Fixed-Effects-Schätzer

- Das andere Extrem, den Within-Schätzer, erhält man durch Anwendung von OLS auf within-transformierte Daten (i.e. alle Variablen werden mit Hilfe der personenspezifischen Mittelwerte zentriert).
- Begründung: Durch Abzug der Between-Gleichung

$$\overline{Y}_i = \overline{X}_i'\beta + \alpha_i + \overline{\epsilon}_i$$

vom Error-Components-Modell

$$Y_{it} = X_{it}'\beta + \alpha_i + \epsilon_{it}$$

erhält man

$$Y_{it} - \overline{Y}_i = (X_{it} - \overline{X}_i)'\beta + (\epsilon_{it} - \overline{\epsilon}_i)$$

Within-Schätzer / Fixed-Effects-Schätzer

- Beim Within-Schätzer werden also sämtliche Niveau-Unterschiede zwischen den Personen aus den Daten herausgerechnet, so dass nur noch die Within-Variation verbleibt.
- Der Within-Schätzer liefert auch dann einen konsistenten Schätzer für β, wenn die RE-Annahme verletzt ist. Dies erkennt man daran, dass durch die Zentrierung der Daten die personenspezifischen Fehler α_i „unter den Tisch“ fallen.
- Bei der Berechnung der Standardfehler muss allerdings berücksichtigt werden, dass durch die Zentrierung einige Freiheitsgrade verloren gehen ($df = N(T-1) - k$ anstatt $df = N \cdot T - k - 1$ bei einem balancierten Panel).

Within-Schätzer / Fixed-Effects-Schätzer

- Beispiel Heiratsprämie: Within-Schätzer

```
. generate c_wage = wage - m_wage
. generate c_marr = marr - m_marr
. regress c_wage c_marr, nocons dof(`=4 * (6-1) - 1´)
```

Source	SS	df	MS	Number of obs =	24
Model	750000	1	750000	F(1, 19) =	158.33
Residual	90000	19	4736.84211	Prob > F =	0.0000
				R-squared =	0.8929
Total	840000	24	35000	Adj R-squared =	0.8647
				Root MSE =	68.825

c_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
c_marr	500	39.73597	12.58	0.000	416.8317	583.1683

Within-Schätzer / Fixed-Effects-Schätzer

- Der Fixed-Effects-Schätzer ist äquivalent zum Durchschnitt der Koeffizienten, die man erhält, wenn man für jeden Fall ein eigenes Regressionsmodell schätzt (bei einem unbalancierten Panel handelt es sich um einen gewichteten Durchschnitt). Dies verdeutlicht, dass keine Between-Variation für die Schätzung verwendet wird.

```
. local b = 0
. forv i = 3/4 {
  2.     qui regress wage marr if id==`i´
  3.     local b = `b´ + _b[marr]
  4. }
. local b = `b´ / 2
. di `b´
500
```

First-Difference-Schätzer

- Der Within-Effekt kann auch durch First-Differencing geschätzt werden. Dabei werden intraindividuelle Differenzen in Y und X aufeinander regressiert, also

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + (\epsilon_{it} - \epsilon_{i,t-1})$$

- Beispiel:

```
. tsset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 6
                delta:  1 unit
. generate d_wage = wage - L.wage
(4 missing values generated)
. generate d_marr = marr - L.marr
(4 missing values generated)
. regress d_wage d_marr, nocons noheader
```

d_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d_marr	450	71.63504	6.28	0.000	300.0661	599.9339

First-Difference-Schätzer

- Bei nur zwei Messzeitpunkten sind der FE-Schätzer und der First-Difference-Schätzer identisch.
- Bei mehr als zwei Messzeitpunkten ist der FE-Schätzer effizienter, da mehr Information ausgenutzt wird.
- In dem Beispiel mit der Heiratsprämie werden beim First-Difference-Schätzer faktisch jeweils nur zwei Beobachtungen pro Fall verwendet (unmittelbar vor und nach der Heirat). Beim FE-Schätzer hingegen fliessen alle Beobachtungen vor und nach Heirat in die Schätzung ein (was die Schätzung stabilisiert).

LSDV-Schätzer

- Weiterhin lässt sich der Within-Schätzer durch Anwendung einer OLS-Regression berechnen, die für jede Person im Datensatz eine Dummy-Variable enthält (Least-Squares-Dummy-Variable-Schätzer). Die personenspezifischen Konstanten α_i werden so direkt geschätzt.
- Beispiel:

```
. regress wage marr ibn.id, nocons noheader
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	500	39.73597	12.58	0.000	416.8317	583.1683
id						
1	1000	28.09757	35.59	0.000	941.1911	1058.809
2	2000	28.09757	71.18	0.000	1941.191	2058.809
3	3000	34.41236	87.18	0.000	2927.974	3072.026
4	4000	34.41236	116.24	0.000	3927.974	4072.026

- Der LSDV-Schätzer ist identisch zum FE-Schätzer, kann aber bei grossem N rechentechnische Schwierigkeiten verursachen.

FE-Schätzer

- In der Praxis wird deshalb zumeist der FE-Schätzer eingesetzt.
- Beispiel:

```
. xtset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 6
                delta:  1 unit

. xtreg wage marr, fe
Fixed-effects (within) regression               Number of obs      =        24
Group variable: id                              Number of groups   =         4

R-sq:  within  = 0.8929                         Obs per group: min =         6
       between = 0.8351                                        avg =       6.0
       overall = 0.4064                                        max =         6

                                                F(1,19)            =    158.33
corr(u_i, Xb)  = 0.5164                         Prob > F           =    0.0000

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         500   39.73597    12.58   0.000     416.8317    583.1683
       _cons |        2500   17.20618   145.30   0.000     2463.987    2536.013
-------------+----------------------------------------------------------------
     sigma_u |  1290.9944
     sigma_e |   68.82472
         rho |   .99716595   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(3, 19) =  1548.15                 Prob > F = 0.0000
```

Random-Effects-Modell

- Ein Nachteil des FE-Modells ist, dass keine Effekte für personenkonstante Variablen (z.B. Geschlecht) geschätzt werden können.
- Weiterhin ergibt sich häufig das Problem, dass eine Variable über die Zeit sehr viel weniger variiert als zwischen Personen, und so der FE-Schätzer im Vergleich zum gepoolten Modell wenig effizient ist.
- Aus diesen Gründen wird zur Schätzung des Error-Components-Modells oft das sog. Random-Effects-Modell verwendet.

Random-Effects-Modell

- Beim RE-Modell wird der personenspezifische Fehler α_i als Zufallsvariable mit Varianz σ^2_α aufgefasst. Weiterhin wird angenommen, dass α_i nicht mit X korreliert ist (RE-Annahme).
- Der Parametervektor β kann unter diesen Bedingungen mit Hilfe der gepoolten OLS-Regression (POLS) konsistent geschätzt werden. Aufgrund der Autokorrelation ist POLS jedoch nicht effizient.
- Eine effiziente Schätzung ist mit GLS (Generalized Least Squares) möglich. Dies entspricht der Anwendung von OLS auf die gemäss

$$Y_{it} - \hat{\theta}_i \overline{Y}_i = (X_{it} - \hat{\theta}_i \overline{X}_i)'\beta + (1 - \hat{\theta}_i)\alpha_i + (\epsilon_{it} - \hat{\theta}_i \bar{\epsilon}_i)$$

transformierten Daten mit

$$\hat{\theta}_i = 1 - \sqrt{\frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_\epsilon + T_i \hat{\sigma}^2_\alpha}}$$

Random-Effects-Modell

- Beispiel:

```
. xtset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 6
                delta:  1 unit

. xtreg wage marr, re theta
Random-effects GLS regression                   Number of obs      =        24
Group variable: id                              Number of groups   =         4

R-sq:  within  = 0.8929                         Obs per group: min =         6
       between = 0.8351                                        avg =       6.0
       overall = 0.4064                                        max =         6

                                                Wald chi2(1)       =    121.76
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
theta          = .96026403
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
marr	503.1554	45.59874	11.03	0.000	413.7835	592.5273
_cons	2499.211	406.038	6.16	0.000	1703.391	3295.031
sigma_u	706.54832					
sigma_e	68.82472					
rho	.99060052	(fraction of variance due to u_i)				

Random-Effects-Modell

- Man erkennt, dass der RE-Schätzer in Abhängigkeit von $\hat{\theta}$ irgendwo zwischen dem FE-Schätzer und dem POLS-Schätzer liegen wird.
  - Bei $\hat{\theta} = 0$ erhalten wir den POLS-Schätzer.
  - Bei $\hat{\theta} = 1$ erhalten wir den FE-Schätzer.
- Falls die Annahme, dass α nicht mit X zusammenhängt, verletzt ist, weist der RE-Schätzer eine Verzerrung auf. Die Verzerrung ist umso kleiner ...
  - ...je geringer der Zusammenhang zwischen α und X
  - ...je grösser σ^2_α im Vergleich zu σ^2_ϵ
  - ...je grösser T

Random-Effects versus Fixed-Effects

- In den meisten Anwendungen muss davon ausgegangen werden, dass die Korrelation zwischen α_i und X nicht gleich null ist. Der FE-Schätzer ist aus diesem Grund i.d.R. vorzuziehen.
- Nur wenn diese RE-Annahme nicht verletzt ist, ist der RE-Schätzer besser, da er mehr Information ausnützt und deshalb eine höhere Effizienz aufweist als der FE-Schätzer.
- Die RE-Annahme kann mit einem sog. Hausman-Test geprüft werden. Es wird dabei der „immer“ konsistente (aber u.U. ineffiziente) FE-Schätzer mit dem „manchmal“ konsistenten (aber dann effizienten) RE-Schätzer verglichen:

$$H = (\hat{\beta}^{\text{FE}} - \hat{\beta}^{\text{RE}})'[\hat{V}(\hat{\beta}^{\text{FE}}) - \hat{V}(\hat{\beta}^{\text{RE}})]^{-1}(\hat{\beta}^{\text{FE}} - \hat{\beta}^{\text{RE}}) \overset{a}{\sim} \chi^2(k)$$

Hausman-Test

Heiratsprämie mit den Daten der Schweizerischen Arbeitskräfteerhebung (SAKE) 1991-2008 (rotierendes Panel); Auswahl: Männer im Alter von 25-45

```
. xtset hhnr
       panel variable:  hhnr (unbalanced)

. xtreg lnwage married alter, re

Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0105                         Obs per group: min =         1
       between = 0.0473                                        avg =       2.3
       overall = 0.0431                                        max =         5

                                                Wald chi2(2)       =   2373.01
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
married	.0222534	.0041884	5.31	0.000	.0140443	.0304625
alter	.0174676	.0003835	45.54	0.000	.0167159	.0182193
_cons	3.096188	.0135052	229.26	0.000	3.069718	3.122657
sigma_u	.4208396					
sigma_e	.25924995					
rho	.72490382	(fraction of variance due to u_i)				

```
. estimates store re
```

Hausman-Test

```
. xtreg lnwage married alter, fe

Fixed-effects (within) regression               Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.0464                                        avg =       2.3
       overall = 0.0421                                        max =         5

                                                F(2,48750)         =    264.04
corr(u_i, Xb)  = -0.0321                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0005383   .0070654     0.08   0.939      -.01331    .0143865
       alter |   .0199373   .0008758    22.76   0.000     .0182207    .0216539
       _cons |   3.038927   .0310898    97.75   0.000     2.977991    3.099863
-------------+----------------------------------------------------------------
     sigma_u |  .46830167
     sigma_e |  .25924995
         rho |  .76542217   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(37469, 48750) =     6.24        Prob > F = 0.0000

. estimates store fe

. hausman fe re
                 ---- Coefficients ----
              |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
              |       fe           re         Difference          S.E.
--------------+----------------------------------------------------------------
     married  |    .0005383     .0222534       -.0217151        .0056901
       alter  |    .0199373     .0174676        .0024697        .0007874
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       22.34
                Prob>chi2 =      0.0000
```

Robuste Standardfehler?

- Durch die Berücksichtigung der personenspezifischen Komponente wird das Problem korrelierter Fehler im FE- und im RE-Modell grösstenteils behoben. Trotzdem ist es ratsam, cluster-robuste Standardfehler zu verwenden, da noch ein Rest an Autokorrelation vorhanden sein kann.

```
. xtreg lnwage married alter, fe vce(cluster hhnr)
Fixed-effects (within) regression               Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.0464                                        avg =       2.3
       overall = 0.0421                                        max =         5

                                                F(2,37469)         =    229.16
corr(u_i, Xb)  = -0.0321                        Prob > F           =    0.0000

                                (Std. Err. adjusted for 37470 clusters in hhnr)
------------------------------------------------------------------------------
             |               Robust
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0005383   .0076393     0.07   0.944    -.0144348    .0155114
       alter |   .0199373   .0009457    21.08   0.000     .0180836     .021791
       _cons |   3.038927   .0333828    91.03   0.000     2.973496    3.104358
-------------+----------------------------------------------------------------
     sigma_u |  .46830167
     sigma_e |  .25924995
         rho |  .76542217   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Interaktion mit zeitkonstanten Variablen im FE-Modell

- Obwohl im FE-Modell keine Effekte von zeitkonstanten Variablen geschätzt werden können, lassen sich *Interaktionseffekte* zwischen zeitveränderlichen und zeitkonstanten Variablen sehr wohl modellieren.
  - Z.B. kann so geprüft werden, ob sich der Effekt einer Variablen je nach Bildungsgrad unterscheidet.

```
. gen alterXeduc = alter * educyrs
. xtreg lnwage married alter alterXeduc, fe
Fixed-effects (within) regression               Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0113                         Obs per group: min =         1
       between = 0.1038                                        avg =       2.3
       overall = 0.0970                                        max =         5

                                                F(3,48749)         =    186.31
corr(u_i, Xb)  = 0.0803                         Prob > F           =    0.0000
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.0002657	.0070634	0.04	0.970	-.0135787	.0141101
alter	.0148436	.0012716	11.67	0.000	.0123513	.0173359
alterXeduc	.0003727	.0000675	5.52	0.000	.0002405	.0005049
_cons	3.054843	.0312136	97.87	0.000	2.993664	3.116022
sigma_u	.45502922					
sigma_e	.25917151					
rho	.75505247	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(37469, 48749) =     5.11        Prob > F = 0.0000
```

Trend-/Periodeneffekte im FE-Modell

- Beim FE-Schätzer werden nur Fälle berücksichtigt, bei denen sich X verändert. Alle anderen Beobachtungen werden quasi ignoriert.
- Die Beobachtungen ohne Veränderung von X können aber trotzdem wichtig sein, um Trends oder Reifungsprozesse zu kontrollieren.
- Häufig empfiehlt es sich deshalb, entsprechende Trend- oder Periodenvariablen in das Modell aufzunehmen (z.B. Dummies für die Befragungsjahre).
- Dies führt dazu, dass auch Personen ohne Veränderung von X zur Schätzung beitragen (im einfachsten Fall erhält man so den Difference-in-Difference-Schätzer).

Trend-/Periodeneffekte im FE-Modell

```
. xtreg lnwage married alter i.jahr, fe
Fixed-effects (within) regression               Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0131                         Obs per group: min =         1
       between = 0.0027                                        avg =       2.3
       overall = 0.0022                                        max =         5

                                                F(19,48733)        =     34.00
corr(u_i, Xb)  = -0.1816                        Prob > F           =    0.0000
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	-.000601	.007064	-0.09	0.932	-.0144466	.0132446
alter	.0033592	.0042902	0.78	0.434	-.0050496	.011768
jahr						
1992	.044849	.0095601	4.69	0.000	.0261111	.0635869
1993	.0343588	.0126207	2.72	0.006	.0096222	.0590955
1994	.0499114	.01624	3.07	0.002	.0180808	.081742
1995	.0876958	.0201844	4.34	0.000	.0481342	.1272574
1996	.0888028	.024358	3.65	0.000	.0410608	.1365448
1997	.091642	.0284916	3.22	0.001	.035798	.1474859
1998	.1131903	.0326849	3.46	0.001	.0491276	.1772531
1999	.135103	.0367893	3.67	0.000	.0629956	.2072104
2000	.1624754	.0410994	3.95	0.000	.08192	.2430308
2001	.2068566	.0452739	4.57	0.000	.1181191	.295594
2002	.2400234	.0494354	4.86	0.000	.1431293	.3369175
2003	.2594175	.0535379	4.85	0.000	.1544825	.3643524
2004	.2711599	.0577477	4.70	0.000	.1579737	.384346
2005	.2773405	.0619571	4.48	0.000	.1559038	.3987773
2006	.2893532	.066222	4.37	0.000	.1595572	.4191492
2007	.3053256	.0704626	4.33	0.000	.1672181	.4434331

Einige weitere Aspekte

- Auch ein FE-Schätzer kann verzerrt sein:
  - Wenn systematische Schocks auftreten, nachdem sich X verändert hat (Periodeneffekte; kann kontrolliert werden, siehe oben).
  - Falls es zeitveränderliche unbeobachtete Variablen gibt, die sowohl X als auch Y beeinflussen.
  - Simultanität bzw. umgekehrte Kausalität.
  - Messfehler: Der durch Messfehler verursachte Attenuation-Bias kann in einem FE-Modell akzentuiert auftreten, weil durch die Differenzierung der Anteil der Messfehler an der genutzten Varianz von X zunimmt.
- Generalisierung von FE-Schätzern:
  - Mit dem FE-Modell wird immer ein ATT geschätzt (Treatment effect on the treated). Das heisst, der geschätzte Effekt bezieht sich nur auf die Subpopulation der Personen mit Veränderung in X. Bei Effekt-Heterogenität kann man die Ergebnisse also nur bedingt auf die Gesamtpopulation verallgemeinern.

Hybrid-Modell

- Will man (deskriptive) Effekte von personenkonstanten Merkmalen schätzen, aber für zeitveränderliche Variablen nicht auf die Vorteile des FE-Modells verzichten, kann man sich mit einem sog. Hybrid-Modell behelfen.
- Dazu werden personenspezifischen Mittelwerte der zeitveränderlichen Variablen berechnet und die Variablen entsprechend zentriert. Es werden dann die Variablen mit den personenspezifischen Mittelwerten sowie die zentrierten Variablen in ein RE-Modell aufgenommen. Die Within- und Between-Effekte der zeitveränderlichen Variablen werden dadurch separiert.
- Alternativ kann man auch die Variablen mit den personenspezifischen Mittelwerten und die nicht-zentrierten Variablen in das Modell aufnehmen. Beide Varianten führen inhaltlich zum gleichen Ergebnis.

Hybrid-Modell

- Beispiel: Heiratsprämie für Männer
  - SAKE-Daten 1991-2008 (rotierendes Panel)
  - Männer im Alter von 25-45 Jahren
- Pooled-OLS

```
. reg lnwage married alter
```

Source	SS	df	MS	Number of obs =	86222
Model	857.201882	2	428.600941	F(2, 86219) =	1946.77
Residual	18981.9295	86219	.220159471	Prob > F =	0.0000
				R-squared =	0.0432
Total	19839.1313	86221	.23009628	Adj R-squared =	0.0432
				Root MSE =	.46921

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.0337687	.0034056	9.92	0.000	.0270938	.0404437
alter	.0166827	.0002959	56.38	0.000	.0161028	.0172627
_cons	3.135317	.0103355	303.35	0.000	3.115059	3.155574

Hybrid-Modell

- RE-Schätzer

```
. xtset hhnr
       panel variable:  hhnr (unbalanced)

. xtreg lnwage married alter, re theta
```

```
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0105                         Obs per group: min =         1
       between = 0.0473                                        avg =       2.3
       overall = 0.0431                                        max =         5

                                                Wald chi2(2)       =   2373.01
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

theta

min	5%	median	95%	max
0.4755	0.4755	0.6006	0.7344	0.7344

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
married	.0222534	.0041884	5.31	0.000	.0140443	.0304625
alter	.0174676	.0003835	45.54	0.000	.0167159	.0182193
_cons	3.096188	.0135052	229.26	0.000	3.069718	3.122657
sigma_u	.4208396					
sigma_e	.25924995					
rho	.72490382	(fraction of variance due to u_i)				

Hybrid-Modell

- FE-Schätzer

```
. xtreg lnwage married alter, fe
Fixed-effects (within) regression               Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.0464                                        avg =       2.3
       overall = 0.0421                                        max =         5

                                                F(2,48750)         =    264.04
corr(u_i, Xb)  = -0.0321                        Prob > F           =    0.0000

------------------------------------------------------------------------------
     lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
    married |   .0005383   .0070654     0.08   0.939      -.01331    .0143865
      alter |   .0199373   .0008758    22.76   0.000     .0182207    .0216539
      _cons |   3.038927   .0310898    97.75   0.000     2.977991    3.099863
------------+----------------------------------------------------------------
    sigma_u |  .46830167
    sigma_e |  .25924995
        rho |  .76542217   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(37469, 48750) =     6.24        Prob > F = 0.0000
```

Hybrid-Modell

- RE-Schätzer mit zentrierten Variablen (Within-Schätzer)

```
. bysort hhnr: center married alter, mean
. xtreg lnwage c_married c_alter, re
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0000                         Obs per group: min =         1
       between = 0.0000                                        avg =       2.3
       overall = 0.0018                                        max =         5

                                                Wald chi2(2)       =    532.19
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	.0005383	.0070381	0.08	0.939	-.0132561	.0143327
c_alter	.0199373	.0008724	22.85	0.000	.0182273	.0216472
_cons	3.732782	.0024552	1520.38	0.000	3.72797	3.737594
sigma_u	.43358892					
sigma_e	.25924995					
rho	.73664604	(fraction of variance due to u_i)				

Hybrid-Modell

- RE-Schätzer mit zentrierten Variablen und Mittelwerten (Within- und Between-Schätzer)

```
. xtreg lnwage c_married m_married c_alter m_alter, re
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.0475                                        avg =       2.3
       overall = 0.0433                                        max =         5

                                                Wald chi2(4)       =   2396.19
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	.0005383	.0070368	0.08	0.939	-.0132536	.0143301
m_married	.0347906	.0052161	6.67	0.000	.0245673	.0450139
c_alter	.0199373	.0008723	22.86	0.000	.0182277	.0216469
m_alter	.0166842	.0004304	38.77	0.000	.0158407	.0175277
_cons	3.116765	.0149435	208.57	0.000	3.087476	3.146054
sigma_u	.4208396					
sigma_e	.25924995					
rho	.72490382	(fraction of variance due to u_i)				

Hybrid-Modell

- Alternative Formulierung des Hybrid-Modells (der Effekt von `m_*` entspricht der Differenz zwischen Within- und Between-Schätzer)

```
. xtreg lnwage married m_married alter m_alter, re
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.0475                                        avg =       2.3
       overall = 0.0433                                        max =         5

                                                Wald chi2(4)       =   2396.19
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
married	.0005383	.0070368	0.08	0.939	-.0132536	.0143301
m_married	.0342523	.0087592	3.91	0.000	.0170846	.0514201
alter	.0199373	.0008723	22.86	0.000	.0182277	.0216469
m_alter	-.0032531	.0009727	-3.34	0.001	-.0051594	-.0013467
_cons	3.116765	.0149435	208.57	0.000	3.087476	3.146054
sigma_u	.4208396					
sigma_e	.25924995					
rho	.72490382	(fraction of variance due to u_i)				

Hybrid-Modell

- Hybrid-Modell inklusive personenkonstante Variablen

```
. xtreg lnwage c_married m_married c_alter m_alter educyrs ausland, re
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0068                         Obs per group: min =         1
       between = 0.2035                                        avg =       2.3
       overall = 0.1893                                        max =         5

                                                Wald chi2(6)       =   9625.17
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	.0005037	.0070491	0.07	0.943	-.0133124	.0143197
m_married	.0809681	.0048439	16.72	0.000	.0714742	.090462
c_alter	.0170054	.0008749	19.44	0.000	.0152907	.0187201
m_alter	.0146784	.0003952	37.14	0.000	.0139038	.0154531
educyrs	.0631823	.0007827	80.72	0.000	.0616482	.0647163
ausland	-.0808968	.0043587	-18.56	0.000	-.0894397	-.0723538
_cons	2.404607	.016681	144.15	0.000	2.371912	2.437301
sigma_u	.3757801					
sigma_e	.25919076					
rho	.67762522	(fraction of variance due to u_i)				

Hybrid-Modell

- Wieso haben sich die Within-Schätzer bei Aufnahme der „personenkonstanten“ Variablen verändert? Die Variablen sind im vorliegenden Fall nicht wirklich personenkonstant!

```
. bysort hhnr: center educyrs ausland, mean
. sum c_educyrs m_educyrs c_ausland m_ausland
```

Variable	Obs	Mean	Std. Dev.	Min	Max
c_educyrs	86222	-1.15e-18	.3667774	-6.8	7.4
m_educyrs	86222	12.47437	2.616576	8	17.5
c_ausland	86222	5.15e-20	.0676392	-.8	.8
m_ausland	86222	.3451671	.4705894	0	1

Hybrid-Modell

```
. bysort hhnr (jahr): generate educyrs0 = educyrs[1]
. bysort hhnr (jahr): generate ausland0 = ausland[1]
. xtreg lnwage c_married m_married c_alter m_alter educyrs0 ausland0, re
```

```
Random-effects GLS regression                   Number of obs      =     86222
Group variable: hhnr                            Number of groups   =     37470

R-sq:  within  = 0.0107                         Obs per group: min =         1
       between = 0.2017                                        avg =       2.3
       overall = 0.1882                                        max =         5

                                                Wald chi2(6)       =  10099.57
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	.0005383	.0070279	0.08	0.939	-.0132362	.0143128
m_married	.0859785	.0048473	17.74	0.000	.0764779	.095479
c_alter	.0199373	.0008712	22.89	0.000	.0182298	.0216447
m_alter	.0143027	.0003948	36.23	0.000	.013529	.0150764
educyrs0	.0686706	.0008264	83.09	0.000	.0670509	.0702904
ausland0	-.0940611	.0045653	-20.60	0.000	-.1030088	-.0851134
_cons	2.354227	.0169145	139.18	0.000	2.321075	2.387379
sigma_u	.37641253					
sigma_e	.25924995					
rho	.67825982	(fraction of variance due to u_i)				

Dynamische Panel-Modelle

- Die Paneldatenmodelle, die wir bisher besprochen haben, werden als „statisch“ bezeichnet, da die Variablen jeweils zu gleichen Zeitpunkt zueinander in Beziehung gesetzt werden.
- Eine Erweiterung sind so genannte „dynamische“ Modelle, bei denen mit zeitverzögerten Variablen gearbeitet wird.
- Ein dynamisches Standardmodell lautet

$$Y_{it} = \gamma Y_{it-1} + X'_{it}\beta + \alpha_i + \epsilon_{it}$$

- Das Modelle macht Sinn, wenn man davon ausgeht, dass der Wert von Y zu $t-1$ einen *kausalen* Effekt auf den Wert von Y zu t hat (true state dependence).

Dynamische Panel-Modelle

- Bei dynamischen Modellen ergeben sich erhebliche schätztechnische Probleme.
- Beispielsweise ist eine RE-Schätzung verzerrt, weil Y_{it-1} von α_i abhängt. Die RE-Annahme ist also per Definition verletzt.
- Auch eine Fixed-Effects-Schätzung führt nicht zum Ziel.
- Für die Schätzung dynamischer Modelle wird deshalb auf IV-Methoden zurückgegriffen.
  - Es wird dabei zuerst differenziert, um die personenspezifischen Effekte loszuwerden, also

$$(Y_{it} - Y_{it-1}) = \gamma(Y_{it-1} - Y_{it-2}) + (X_{it} - X_{it-1})'\beta + (\epsilon_{it} - \epsilon_{it-1})$$

  - Da Y_{it-1} und ϵ_{it-1} korreliert sind, wird für $(Y_{it-1} - Y_{it-2})$ ein Instrument benötigt, das nicht mit $(\epsilon_{it} - \epsilon_{it-1})$ zusammenhängt.

Dynamische Panel-Modelle

- Im einfachsten Fall wird Y_{it-2} als Instrument verwendet (Anderson-Hsiao-Schätzer). Y_{it-2} ist ein valides Instrument, da es mit $(Y_{it-1} - Y_{it-2})$ zusammenhängt, jedoch nicht mit $(\epsilon_{it} - \epsilon_{it-1})$.
- Alternativ wurde $(Y_{it-2} - Y_{it-3})$ als Instrument vorgeschlagen, was sich in Monte-Carlo-Studien aber als weniger weniger effizient erwies.
- Beliebt ist weiterhin der GMM-Schätzer von Arellano und Bond, bei dem verzögerte Niveaus und Differenzen von Y wie auch exogenen X-Variablen als Instrumente verwendet werden.

- Diese Methoden beruhen alle auf der Annahme, dass keine Autokorrelation zweiter Ordnung besteht (was nicht wirklich vernünftig getestet werden kann). Ansonsten sind die Instrumente nicht valide. In der Praxis ist zu beobachten, das die Methoden nur selten zu robusten Ergebnissen führen.

Logistische Regression mit Paneldaten

- Bislang haben wir uns nur mit linearen Modellen bzw. Modellen mit einer kontinuierlichen Variablen beschäftigt.
- Panel-Schätzer sind aber auch für nichtlineare Modelle verfügbar.
- Ein wichtiger Fall ist die logistische Regression, für die eine Fixed-Effects-Schätzung möglich ist.
- FE-Schätzung steht weiterhin für Zähldatenmodelle (Poisson, Negativ-Binomial-Modell) und für die Cox-Regression (durch sog. Stratifizierung; d.h. für jede Person ist ein eigener unspezifizierter Verlauf der Baseline-Hazarzdrate zugelassen) zur Verfügung.

Logistische Regression mit Paneldaten

- Das Logit-Modell mit personenspezifischen Fehlern ist gegeben als

$$\Pr(Y_{it} = 1) = \frac{\exp(X'_{it}\beta + \alpha_i)}{1 + \exp(X'_{it}\beta + \alpha_i)}$$

- Wie kann das Modell geschätzt werden?
- Erste Idee: Dummy-Variablen für die Personen ins Modell aufnehmen (analog zu LSDV).
  - Dies führt jedoch zu verzerrten Schätzern aufgrund des so genannten Incidential-Parameters-Problem bei Maximum-Likelihood-Schätzern. (Die Anzahl Parameter steigt direkt mit der Stichprobengrösse, was eine grundlegende Bedingung der asymptotischen Theorie hinter MLE verletzt.)
  - Der Ansatz mit Dummy-Variablen ist deshalb nur sinnvoll, wenn N fix ist und T gegen „unendlich“ geht.

Logistische Regression mit Paneldaten

- Eine konsistente FE-Schätzung ist mit der *bedingten* ML-Methode möglich, bei der α_i durch Konditionierung auf $\sum_t Y_{it}$ kontrolliert wird.
  - FE-Logit (`xtlogit, fe`) ist formal identisch mit Conditional-Logit (`clogit`).
  - FE-Logit hat den grossen Vorteil, dass keine RE-Annahme getroffen werden muss. Das heisst, wie bei der linearen FE-Regression ist der Schätzer konsistent, auch wenn (zeitkonstante) unbeobachtete Heterogenität vorliegt.
  - Zu beachten ist, dass Fälle, bei denen Y immer 0 oder immer 1 ist, keine Information zur bedingten Likelihood beitragen. Der Datensatz wird dadurch insb. in kurzen Panels u.U. drastisch reduziert.
  - Anders als bei der linearen FE-Regression können die fixen Effekte α_i nicht konsistent quantifiziert werden. Dies ist ein Problem, wenn man Marginaleffekte berechnen möchte.

Logistische Regression mit Paneldaten

- Alternativ steht auch für Logit ein RE-Modell mit $\alpha_i \sim N(0, \sigma_\alpha^2)$ zur Verfügung. Eine weitere Alternative ist das GEE (generalized estimating equations) bzw. das Population-Average-Modell, bei dem α_i ignoriert, die Abhängigkeit der Beobachtungen jedoch direkt über die Kovarianzmatrix der Fehlerterme modelliert wird.
  - RE-Logit wie auch das PA-Logit beruhen jedoch auf der RE-Annahme (α_i ist unabhängig von X) und liefern verzerrte Ergebnisse, wenn die Annahme verletzt ist.
  - RE-Logit schätzt subjektspezifische Effekte, PA-Logit schätzt Populationsdurchschnitts-Effekte. Der Unterschied hängt mit der Varianz von α zusammen. Es gilt approximativ:

$$\beta^{PA} \approx \frac{\beta^{RE}}{\sqrt{0.346 V(\alpha_i) + 1}}$$

Logistische Regression mit Paneldaten

- Wie im linearen Modell lassen sich auch bei der logistischen Regression FE und RE zu einem Hybrid-Modell kombinieren.
  - Man zentriere dazu die zeitveränderlichen Kovariaten an den personenspezifischen Mittelwerten und nehme dann die zentrierten Variablen und die Mittelwerte in ein RE-Modell auf. Für die Schätzung der Effekte der zentrierten Variablen wird so nur die „within“-Information verwendet und man erhält ähnliche Ergebnisse wie im FE-Logit.
- Linear-Probability-Modell (LPM): Nicht zuletzt kann auch die Verwendung linearer Panelmodelle eine Option sein. Abgebildet werden dadurch direkt die durchschnittlichen Marginaleffekte auf die Wahrscheinlichkeit das Y_{it} den Wert 1 annimmt.

Logistische Regression mit Paneldaten: Beispiel

- Pooled-Logit

```
. logit lfp married kinder0 kinder7 educyrs0 ausland0, nolog cluster(hhnr)
Logistic regression                                     Number of obs   =     115206
                                                        Wald chi2(5)    =    5923.16
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -51569.862                       Pseudo R2       =     0.1301

                                  (Std. Err. adjusted for 46341 clusters in hhnr)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |  -1.133491   .0335349   -33.80   0.000    -1.199218   -1.067764
     kinder0 |  -.7483326   .0148443   -50.41   0.000    -.7774269   -.7192383
     kinder7 |  -.2759604    .013482   -20.47   0.000    -.3023846   -.2495362
    educyrs0 |   .0649452   .0053709    12.09   0.000     .0544185    .0754718
    ausland0 |  -.0405798   .0254881    -1.59   0.111    -.0905356     .009376
       _cons |   2.068385   .0722702    28.62   0.000     1.926738    2.210032
------------------------------------------------------------------------------
```

Logistische Regression mit Paneldaten: Beispiel

- PA-Logit

```
. xtset hhnr
       panel variable:  hhnr (unbalanced)

. xtlogit lfp married kinder0 kinder7 educyrs0 ausland0, pa nolog robust

GEE population-averaged model                   Number of obs      =    115206
Group variable:                         hhnr    Number of groups   =     46341
Link:                                  logit    Obs per group: min =         1
Family:                             binomial                   avg =       2.5
Correlation:                   exchangeable                    max =         5
                                                Wald chi2(5)       =   6542.88
Scale parameter:                           1    Prob > chi2        =    0.0000

                                  (Std. Err. adjusted for clustering on hhnr)
------------------------------------------------------------------------------
             |               Robust
         lfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   -1.001202   .0276847   -36.16   0.000    -1.055463   -.9469405
     kinder0 |   -.6742328   .0123599   -54.55   0.000    -.6984578   -.6500077
     kinder7 |   -.2571368   .0112474   -22.86   0.000    -.2791812   -.2350923
    educyrs0 |    .0550801   .0048319    11.40   0.000     .0456096    .0645505
    ausland0 |   -.0625342   .0229819    -2.72   0.007    -.1075778   -.0174905
       _cons |    2.008692   .0642319    31.27   0.000        1.8828    2.134585
------------------------------------------------------------------------------
```

Logistische Regression mit Paneldaten: Beispiel

- RE-Logit

```
. xtlogit lfp married kinder0 kinder7 educyrs0 ausland0, re nolog

Random-effects logistic regression              Number of obs      =    115206
Group variable: hhnr                            Number of groups   =     46341

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       2.5
                                                               max =         5

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(5)       =   5618.76
Log likelihood  = -40379.443                    Prob > chi2        =    0.0000
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
married	-2.115862	.0594128	-35.61	0.000	-2.232309	-1.999415
kinder0	-1.645666	.0288523	-57.04	0.000	-1.702215	-1.589117
kinder7	-.6112318	.0257198	-23.77	0.000	-.6616417	-.5608218
educyrs0	.1187836	.0101339	11.72	0.000	.0989214	.1386457
ausland0	-.1941615	.0508473	-3.82	0.000	-.2938205	-.0945026
_cons	4.867177	.1349644	36.06	0.000	4.602652	5.131703
/lnsig2u	2.623627	.0220809			2.58035	2.666905
sigma_u	3.712902	.0409921			3.633422	3.79412
rho	.8073341	.0034346			.8005128	.8139764

```
Likelihood-ratio test of rho=0: chibar2(01) =  2.2e+04 Prob >= chibar2 = 0.000

. di _b[married] / sqrt(0.346 * e(sigma_u)^2 + 1)
-.88085778

. di _b[kinder0] / sqrt(0.346 * e(sigma_u)^2 + 1)
-.68510978
```

Logistische Regression mit Paneldaten: Beispiel

- FE-Logit

```
. xtlogit lfp married kinder0 kinder7, fe nolog
note: multiple positive outcomes within groups encountered.
note: 41044 groups (94927 obs) dropped because of all positive or
      all negative outcomes.

Conditional fixed-effects logistic regression   Number of obs      =     20279
Group variable: hhnr                            Number of groups   =      5297

                                                Obs per group: min =         2
                                                               avg =       3.8
                                                               max =         5

                                                LR chi2(3)         =    918.33
Log likelihood  = -7103.7052                    Prob > chi2        =    0.0000
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
married	-1.018733	.1188769	-8.57	0.000	-1.251728	-.7857386
kinder0	-1.112417	.0460113	-24.18	0.000	-1.202597	-1.022237
kinder7	-.3075244	.0476478	-6.45	0.000	-.4009123	-.2141364

Logistische Regression mit Paneldaten: Beispiel

- Hybrid-Logit

```
. bysort hhnr: center married kinder0 kinder7, mean
. xtlogit lfp c_married c_kinder0 c_kinder7 ///
>     m_married m_kinder0 m_kinder7 educyrs0 ausland0, re nolog
Random-effects logistic regression              Number of obs      =    115206
Group variable: hhnr                            Number of groups   =     46341
Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       2.5
                                                               max =         5
Integration method: mvaghermite                 Integration points =        12
                                                Wald chi2(8)       =   5173.95
Log likelihood  = -40235.442                    Prob > chi2        =    0.0000
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	-1.135551	.11494	-9.88	0.000	-1.36083	-.9102731
c_kinder0	-1.202055	.0462381	-26.00	0.000	-1.29268	-1.11143
c_kinder7	-.3658058	.046613	-7.85	0.000	-.4571655	-.274446
m_married	-2.286048	.0698937	-32.71	0.000	-2.423037	-2.149059
m_kinder0	-1.915535	.0383108	-50.00	0.000	-1.990623	-1.840447
m_kinder7	-.6195026	.0317648	-19.50	0.000	-.6817604	-.5572448
educyrs0	.1193711	.0104422	11.43	0.000	.0989048	.1398375
ausland0	-.1696696	.0522801	-3.25	0.001	-.2721368	-.0672024
_cons	5.188984	.1421182	36.51	0.000	4.910437	5.46753
/lnsig2u	2.674917	.0225812			2.630659	2.719175
sigma_u	3.809349	.0430098			3.725978	3.894587
rho	.8151864	.003402			.8084254	.8217615

```
Likelihood-ratio test of rho=0: chibar2(01) =  2.3e+04 Prob >= chibar2 = 0.000
```

Logistische Regression mit Paneldaten: Beispiel

- FE-LPM

```
. xtreg lfp married kinder0 kinder7, fe
Fixed-effects (within) regression               Number of obs      =    115206
Group variable: hhnr                            Number of groups   =     46341

R-sq:  within  = 0.0202                         Obs per group: min =         1
       between = 0.1393                                        avg =       2.5
       overall = 0.1251                                        max =         5

                                                F(3,68862)         =    473.97
corr(u_i, Xb)  = 0.1507                         Prob > F           =    0.0000
```

lfp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	-.0522877	.005987	-8.73	0.000	-.0640223	-.0405531
kinder0	-.103531	.0030629	-33.80	0.000	-.1095342	-.0975278
kinder7	-.0289731	.0030103	-9.62	0.000	-.0348732	-.023073
_cons	.8889493	.0045088	197.16	0.000	.8801121	.8977865
sigma_u	.34904163					
sigma_e	.24362634					
rho	.67241124	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(46340, 68862) =     4.59        Prob > F = 0.0000
```

Logistische Regression mit Paneldaten: Beispiel

- Marginaleffekte des Hybrid-Modells (für $\alpha_i = 0$)

```
. est restore hybrid
(results hybrid are active now)
. margins, predict(pu0) dydx(c_married c_kinder0 c_kinder7)
Average marginal effects                        Number of obs   =     115206
Model VCE    : OIM
Expression   : Pr(lfp=1 assuming u_i=0), predict(pu0)
dy/dx w.r.t. : c_married c_kinder0 c_kinder7
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	-.0679083	.0069279	-9.80	0.000	-.0814868	-.0543298
c_kinder0	-.0718853	.0028852	-24.92	0.000	-.0775401	-.0662305
c_kinder7	-.0218759	.0028044	-7.80	0.000	-.0273725	-.0163793

Logistische Regression mit Paneldaten: Beispiel

- Stata 14: Marginaleffekte des Hybrid-Modells (mit Integration über α_i)

```
. quietly melogit lfp c_married c_kinder0 c_kinder7 ///
>     m_married m_kinder0 m_kinder7 educyrs0 ausland0 || hhnr:

. margins, dydx(c_married c_kinder0 c_kinder7)

Average marginal effects                        Number of obs     =    115,206
Model VCE    : OIM

Expression   : Marginal predicted mean, predict()
dy/dx w.r.t. : c_married c_kinder0 c_kinder7
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_married	-.0583049	.0057105	-10.21	0.000	-.0694973	-.0471124
c_kinder0	-.0608568	.0024625	-24.71	0.000	-.0656832	-.0560305
c_kinder7	-.0186534	.0023313	-8.00	0.000	-.0232226	-.0140842