Hoe analyseer je paneldata?

Jurjen Iedema - SCP j.iedema@scp.nl

NPSO 8 september 2016:
ontwikkelingen in de tijd - longitudinaal panelonderzoek

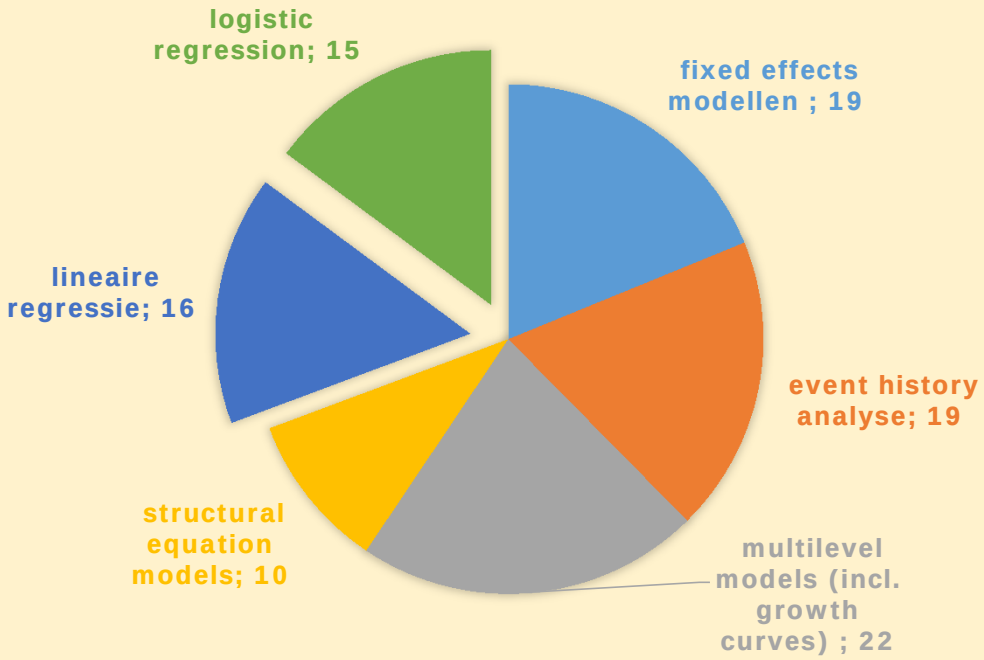
“Brandende vragen” van onderzoekers

- “Bij panelanalyse heb je longitudinale gegevens, hoe kan ik gebruik maken van het longitudinale karakter van de data?”
- “Ik zie bij panelanalyse het *fixed effects* en het *random effects* model voorbijkomen, wat is het verschil?”
- “Kun je met panelanalyse causale relaties vaststellen?”

Panelanalyse in de wetenschap

- **61%** gebruikten panel data in empirische artikelen van het *Journal of Marriage and the Family* tussen 2010 en 2014 (Young & Johnson, 2015).

- Cross-sectionele data en traditionele analyse-modellen verliezen terrein.



Constatering: panelanalyse is lastiger

- Analyse cross-sectionele data over meerdere jaren – *no problemo:* regressie met dummy's voor de jaren…
- Analyse longitudinale panel-data – mensen volgen in de tijd – *issues!*
- Je moet nadenken over:
  1. Corrigeren voor niet-onafhankelijkheid in de data (responsen zijn genest binnen respondenten).
  2. Hoe zit dat met between-subject (*tussen*-persoons) vergelijking en within-subject (*binnen*-persoons) vergelijking?
  3. Veelheid aan modellen! (bomen – bos)

Eerste stap: Literatuur!

- Allison, P.D. (1994). Using Panel Data to Estimate the Effects of Events. *Sociological Methods & Research (23),* 174-199.
- Allison, P.D. (Paul) (2005). *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. SAS Press.
- Allison, Paul D. 2009. *Fixed effects regression models.* London: Sage.
- Berrington, A., Smith, P.W.F. & Sturgis, P. *An Overview of Methods for the Analysis of Panel Data* (2006). ESRC National Centre for Research Methods Briefing Paper.
- Halaby, C. (2004). Panel Models in Sociological Research. *Annual Rev. of Sociology (30),* 507-544.
- Jozef Brüderl: Applied Panel Analysis http://www.ls3.soziologie.uni-muenchen.de/downloads/lehre/lehre_alt/panelanalysis-bruederl.pdf
- Johnson, D. (David) (2005). Two-Wave Panel Analysis: Comparing Statistical Methods for Studying the Effects of Transitions. *Journal of Marriage and Family (67),* 1061–1075.

We willen graag causale uitspraken doen!

- Experiment nodig met random toewijzing aan condities
- Of RCT (Randomized Controlled Trial): een gerandomiseerd veldonderzoek
- Als dat niet gaat? Met panelanalyse volg je personen over de tijd. Je kunt ze dus met zichzelf vergelijken over de tijd (binnenpersoons). Veel variabelen blijven constant over tijd (geslacht, opleiding, beroep, hobby's, etc.).
- Globaal drie criteria voor causaliteit:
  - *de oorzaak gaat vooraf aan het gevolg*
  - *er is samenhang tussen oorzaak en gevolg*
  - *mogelijke andere verklaringen zijn uitgeschakeld.*

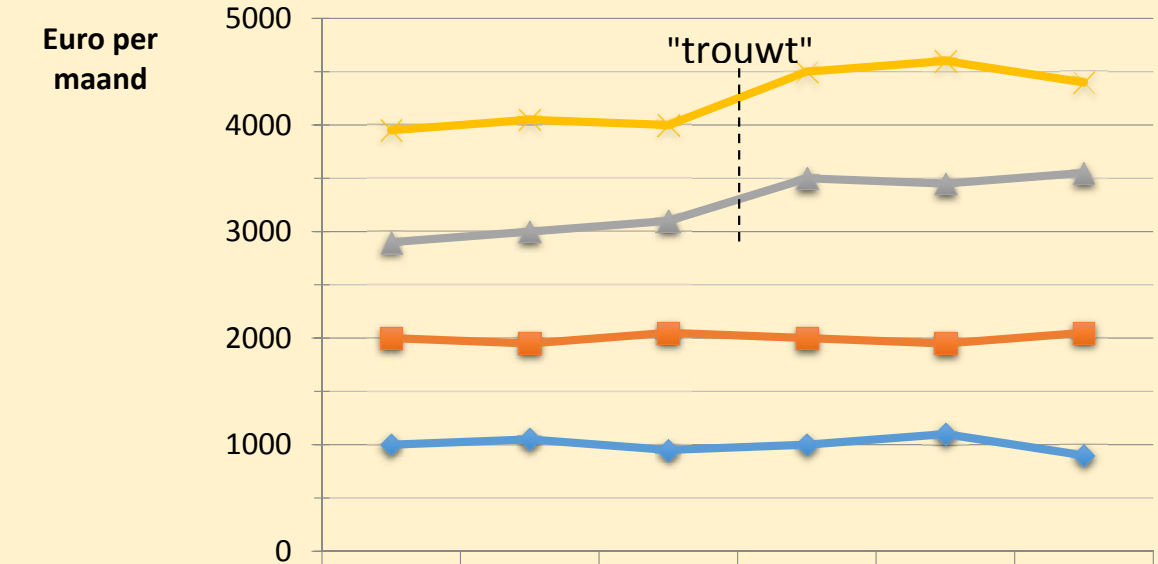
Voorbeeld panelanalyse: meestal “lange” data setup nodig

```
. list id time wage marr, separator(6)

     +-------------------------+       +-------------------------+
     | id   time   wage   marr |       | id   time   wage   marr |
     |-------------------------|       |-------------------------|
  1. |  1      1   1000      0 |   13. |  3      1   2900      0 |
  2. |  1      2   1050      0 |   14. |  3      2   3000      0 |
  3. |  1      3    950      0 |   15. |  3      3   3100      0 |
  4. |  1      4   1000      0 |   16. |  3      4   3500      1 |
  5. |  1      5   1100      0 |   17. |  3      5   3450      1 |
  6. |  1      6    900      0 |   18. |  3      6   3550      1 |
     |-------------------------|       |-------------------------|
  7. |  2      1   2000      0 |   19. |  4      1   3950      0 |
  8. |  2      2   1950      0 |   20. |  4      2   4050      0 |
  9. |  2      3   2050      0 |   21. |  4      3   4000      0 |
 10. |  2      4   2000      0 |   22. |  4      4   4500      1 |
 11. |  2      5   1950      0 |   23. |  4      5   4600      1 |
 12. |  2      6   2050      0 |   24. |  4      6   4400      1 |
     |-------------------------|       |-------------------------|
```

Stata reshape: van “lange” naar “wijde” data setup en vice versa. SPSS: varstocases

Leidt het huwelijk bij mannen tot een hoger inkomen?



	t1	t2	t3	t4	t5	t6
ongetrouwd	1000	1050	950	1000	1100	900
ongetrouwd	2000	1950	2050	2000	1950	2050
trouwt na tijdstip 3	2900	3000	3100	3500	3450	3550
trouwt na tijdstip 3	3950	4050	4000	4500	4600	4400

- Er is een causaal effect: een “trouwpremie”. De gehuwden verdienen €500 meer.
- Maar er is ook selectiviteit of zelfselectie: alleen de beter verdienende mannen trouwen.
- Hoe analyseer je dat?

Cross-sectionele of gepoolde regressie

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

      Source |       SS           df       MS      Number of obs   =         4
-------------+----------------------------------   F(1, 2)        =     12,50
       Model |     6250000         1     6250000   Prob > F        =    0,0715
    Residual |     1000000         2      500000   R-squared       =    0,8621
-------------+----------------------------------   Adj R-squared   =    0,7931
       Total |     7250000         3  2416666,67   Root MSE        =    707,11

------------------------------------------------------------------------------
        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
------------------------------------------------------------------------------
```

```
. regress  wage marr, vce(cluster id)   //correct panel-robust S.E.

Linear regression                               Number of obs     =         24
                                                F(1, 3)           =       7,81
                                                Prob > F          =     0,0681
                                                R-squared         =     0,4065
                                                Root MSE          =     1001,9

                                   (Std. Err. adjusted for 4 clusters in id)
------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |   1833,333   655,9178     2,80   0,068      -254,09    3920,757
       _cons |   2166,667   610,0389     3,55   0,038     225,2508    4108,083
------------------------------------------------------------------------------
```

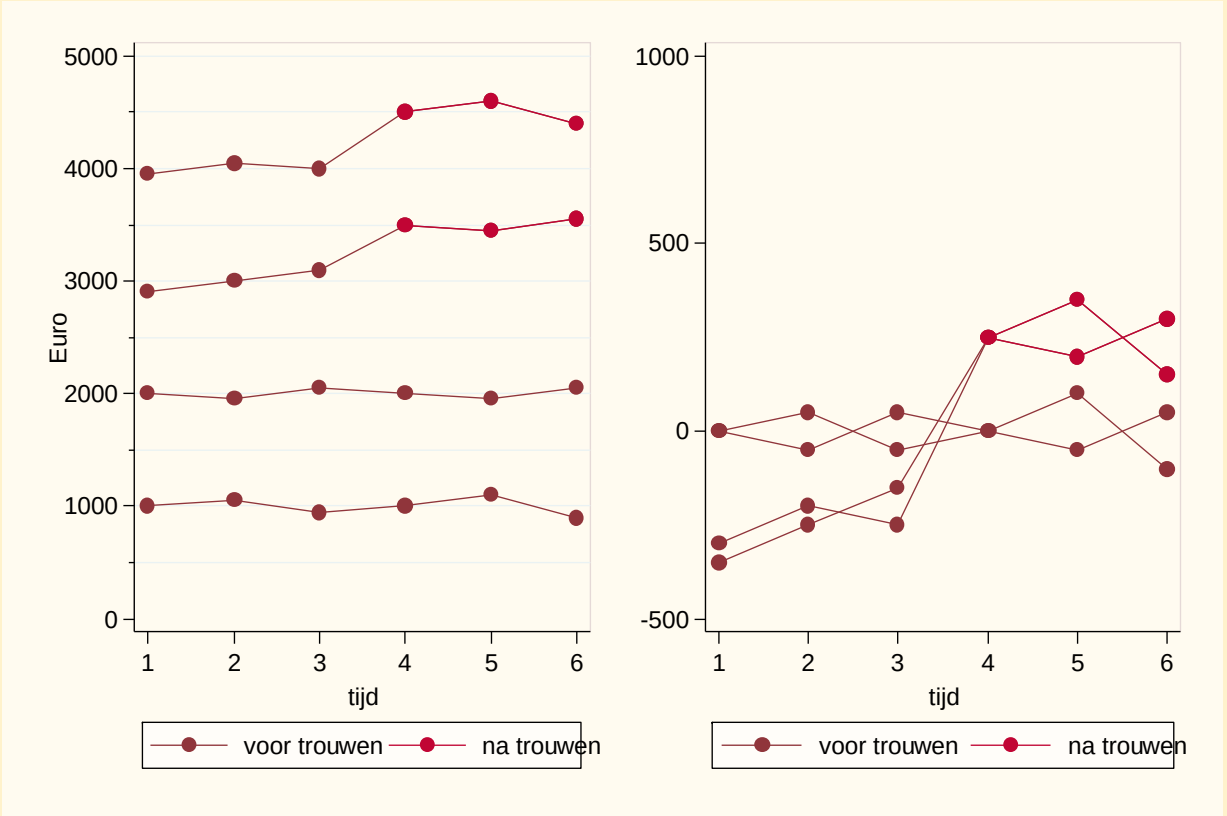
- Cross-sectionele regressie op tijdstip 4 laat een effect van €2500 zien.
- Gepoolde regressie over alle tijdstippen: trouwpremie van €1833.
- Dat is geen €500. Er gaat iets fout…

Fixed-effects regressie (FE) voor het *binnenpersoons* effect

FE (fixed effect regressie) past ***demeaning*** toe. Per variabele wordt het gemiddelde bepaald over de tijd en van de variabele afgetrokken: ***"centreren over de tijd".***

```
. egen      mwage = mean(wage), by(id)
. egen      mmarr = mean(marr), by(id)
. generate  wwage = wage - mwage
. generate  wmarr = marr - mmarr
. regress   wwage wmarr
```

id	time	wage	marr	mwage	mmarr	wwage	wmarr
2	1	2000	0	2000	0	0	0
2	2	1950	0	2000	0	-50	0
2	3	2050	0	2000	0	50	0
2	4	2000	0	2000	0	0	0
2	5	1950	0	2000	0	-50	0
2	6	2050	0	2000	0	50	0
3	1	2900	0	3250	0,5	-350	-0,5
3	2	3000	0	3250	0,5	-250	-0,5
3	3	3100	0	3250	0,5	-150	-0,5
3	4	3500	1	3250	0,5	250	0,5
3	5	3450	1	3250	0,5	200	0,5
3	6	3550	1	3250	0,5	300	0,5



Fixed-effects regressie (FE) (*binnenpersoons* effect)

- Op de voor tijd gecentreerde variabelen kun je in principe een gewone regressie toepassen.
- Maar Stata's xtreg is makkelijker.
- Vóór het xtreg commando moet je wel de paneldata instellen met xtset: `xtset id time`

```
. egen      mwage = mean(wage), by(id)

. egen      mmarr = mean(marr), by(id)

. generate  wwage = wage - mwage

. generate  wmarr = marr - mmarr

. regress   wwage wmarr
```

Source	SS	df	MS		
Model	750000	1	750000	Number of obs	= 24
Residual	85000	22	3863,63636	F(1, 22)	= 194,12
				Prob > F	= 0,0000
Total	835000	23	36304,3478	R-squared	= 0,8982
				Adj R-squared	= 0,8936
				Root MSE	= 62,158

wwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wmarr	500	35,88703	13,93	0,000	425,5749	574,4251
_cons	0	12,68798	0,00	1,000	-26,31326	26,31326

```
. xtreg wage3 marr, fe

Fixed-effects (within) regression               Number of obs     =         24
Group variable: id                              Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0,4732                                         min =          6
     between = 0,8000                                         avg =        6,0
     overall = 0,3958                                         max =          6

                                                F(1,19)           =      17,07
corr(u_i, Xb)  = 0,5013                         Prob > F          =     0,0006
```

wage3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	500	121,0336	4,13	0,001	246,6738	753,3262
_cons	2625	52,40907	50,09	0,000	2515,307	2734,693
sigma_u	1163,6867					
sigma_e	209,63628					
rho	,96856668	(fraction of variance due to u_i)				

F test that all u_i=0: F(3, 19) = 138,42 Prob > F = 0,0000

Fixed-effects regressie (*binnenpersoons* effect):

- laat predictoren die niet over tijd variëren uit de analye vallen
- laat respondenten die niet over tijd variëren uit de analye vallen

- gebruikt dus alleen info die over tijd varieert
- de logica is: iets wat ***niet*** over tijd varieert (een constante dus), kan over tijd ***niet*** iets beïnvloeden

Het between subjects model (BE) (voor de volledigheid)

```
. * BE (between effects)
. regress mwage mmarr
```

Source	SS	df	MS		
Model	30375000	1	30375000	Prob > F	0,0000
Residual	6000000	22	272727,273	R-squared	0,8351
				Adj R-squared	0,8276
Total	36375000	23	1581521,74	Root MSE	522,23

Number of obs	=	24			
F(1, 22)	=	111,38			

mwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mmarr	4500	426,4014	10,55	0,000	3615,698	5384,302
_cons	1500	150,7557	9,95	0,000	1187,352	1812,648

```
. xtreg wage marr, be

Between regression (regression on group means)  Number of obs     =         24
Group variable: id                              Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0,8982                                         min =          6
     between = 0,8351                                         avg =        6,0
     overall = 0,4065                                         max =          6

                                                F(1,2)            =      10,13
sd(u_i + avg(e_i.))=  707,1068                  Prob > F          =     0,0862
```

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	4500	1414,214	3,18	0,086	-1584,87	10584,87
_cons	1500	500	3,00	0,095	-651,3264	3651,326

Het between-effects model ***negeert*** de factor tijd en kijkt alleen naar de *over tijd gepoolde* gemiddelden.

Vandaar dat een regressie op de over tijd gepoolde gemiddelden mwage en mmarr hetzelfde oplevert als het speciale commando xtreg, be

Tenslotte het random-effects model (RE)

- Het random-effects model is een (matrix gewogen) gemiddelde van het fixed-effects (within) en het between-effects model.

  [Je kunt de theta opvragen: xtreg, re theta . Zit deze dicht bij 1 dan geeft het RE vooral het within-effect weer; dicht bij 0 vooral het between-effect]

- De coëfficiënt van een tijdsvariërende variabele – bijv. trouwen – drukt bij het RE-model dus een soort gemiddeld effect uit van het ***binnenpersoons*** (within subjects) effect van trouwen op inkomen en het ***tussenpersoons*** (between subjects) effect van trouwen op inkomen.

Het random-effects model (RE)

```
. xtreg wage marr, re theta

Random-effects GLS regression                   Number of obs     =         24
Group variable: id                              Number of groups  =          4

R-sq:                                           Obs per group:
     within  = 0,8982                                         min =          6
     between = 0,8351                                         avg =        6,0
     overall = 0,4065                                         max =          6

                                                Wald chi2(1)      =     128,82
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0,0000
theta          = ,96138358
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
marr	502,9802	44,31525	11,35	0,000	416,1239	589,8365
_cons	2499,255	406,0315	6,16	0,000	1703,448	3295,062
sigma_u	706,57936					
sigma_e	66,885605					
rho	,99111885	(fraction of variance due to u_i)				

Het random-effects model schat het effect van trouwen hier op €503.

Dat klopt aardig.

Aan de theta van 0,96 is te zien dat het hier vooral een fixed effect (binnenpersoons) betreft.

Verschillen fixed- en random-effect model

- Interpretatie van effecten verschillend: *FE geeft binnenpersoons effect. RE houdt het midden tussen het binnenpersoons en het tussenpersoons effect.*
- FE ***corrigeert automatisch*** voor alle niet-tijdsvariërende predictoren, gemeten of ongemeten!
- Echter, **niet-tijdsvariërende predictoren vallen uit** het FE-model. Je krijgt geen effect te zien (bijv. opleiding). Wat niet verandert kan volgens FE geen causaal effect hebben op iets anders. Qua duidelijke, “causale” interpretatie een pluspunt van FE.
- FE gebruikt alleen variatie binnen individuen – *the ‘within-group’ estimator.* Deze variatie is soms maar een klein deel van het totaal en kan dus leiden tot een lagere precisie.
- Als aan assumpties voldaan is, is de RE schatting ***efficiënter*** dan die van FE.

Verschillen fixed- en random-effects model

- Met een fixed-effect (FE) of random-effect (RE) model wordt op verschillende manieren gecontroleerd voor individuele verschillen:
  - In een fixed-effect model wordt dit gedaan met ***demeaning***, in feite een apart intercept voor iedere respondent (alleen efficiënter). *This can be considered a* ***`fixed-effects' model*** *because the regression line is raised or lowered by a fixed amount for each individual.*
  - In een random-effect model is er niet “een apart intercept per individu”, maar een random intercept met een bepaalde variantie *V*. [RE = random intercept multilevel model]. Een voordeel is dat het minder vrijheidsgraden kost (efficiënter dan FE). Een nadeel is dat er **geen correlatie mag zijn tussen predictoren in het model en de variantie *V* (het endogeniteitsprobleem)**.
  - Deze assumptie kan worden getoetst met de Hausman test.

Hausman test

```
. quietly: xtreg wage marr, fe

. est store fe

. quietly: xtreg wage marr, re theta

. est store re

. hausman re fe
```

	——— Coefficients ———			
	(b) re	(B) fe	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
marr	502,9802	500	2,980234	21,73967

```
                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(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                  =        0,02
        Prob>chi2 =      0,8910
```

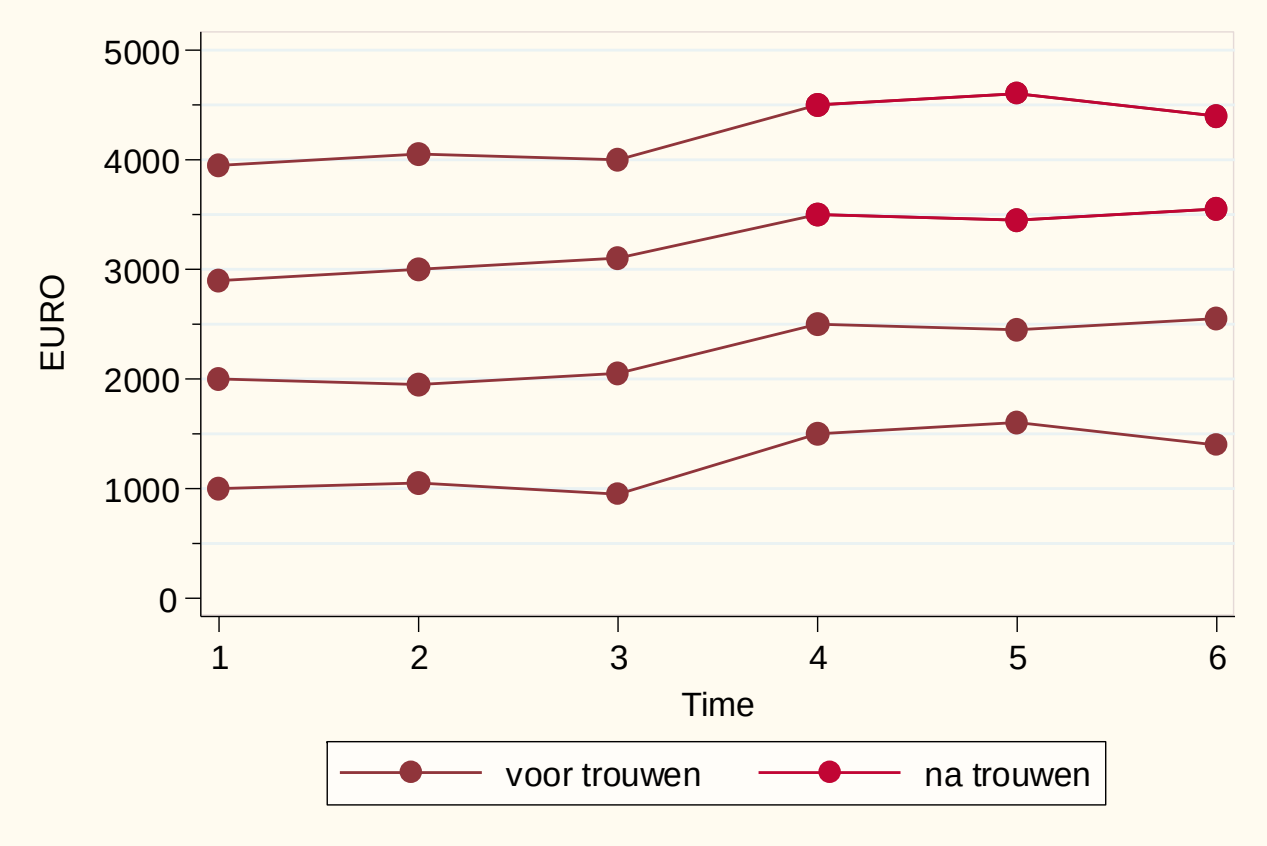
Met het commando estimates store naam kun je de resultaten van een analyse opslaan en verderop bij de hausman test aangeven.

De test is hier niet sign. (dit is in het wild zeldzaam…)

Is de Hausman test significant dan wordt het RE model verworpen en val je dus terug op het FE model.

RE is efficiënter (maakt gebruik van alle respondenten), maar bij sign. Hausman test is FE beter, want consistent.

Aangepast voorbeeld: periode effect valt samen met trouwen



Er is nu een periode-effect op wave 4: **iedereen verdient €500 meer**.

Trouwen (van bovenste 2 mannen) heeft hier **geen effect** meer.

Hoe gaat het fixed-effects model hiermee om?

xtreg wage3 marr, fe

```
. xtreg wage3 marr, fe

Fixed-effects (within) regression               Number of obs      =        24
Group variable: id                              Number of groups   =         4

R-sq:                                           Obs per group:
     within  = 0,4732                                         min =         6
     between = 0,8000                                         avg =       6,0
     overall = 0,3958                                         max =         6

                                                F(1,19)            =     17,07
corr(u_i, Xb)  = 0,5013                         Prob > F           =    0,0006

------------------------------------------------------------------------------
       wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         500   121,0336     4,13   0,001     246,6738    753,3262
       _cons |        2625   52,40907    50,09   0,000     2515,307    2734,693
-------------+----------------------------------------------------------------
     sigma_u |  1163,6867
     sigma_e |  209,63628
         rho |  ,96856668   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(3, 19) = 138,42                     Prob > F = 0,0000
```

Het gaat mis!

Er wordt niet gecontroleerd voor het periode-effect.

De oplossing is om het periode-effect te modeleren m.b.v. **dummy's voor tijd**.

Tweeweg FE-regressie

xtreg wage3 marr i.time, fe

wage3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marr	4,41e-13	58,24824	0,00	1,000	-124,93	124,93
time						
1	0	(base)				
2	50	50,44445	0,99	0,338	-58,19259	158,1926
3	62,5	50,44445	1,24	0,236	-45,69259	170,6926
4	537,5	58,24824	9,23	0,000	412,57	662,43
5	562,5	58,24824	9,66	0,000	437,57	687,43
6	512,5	58,24824	8,80	0,000	387,57	637,43
_cons	2462,5	35,66961	69,04	0,000	2385,996	2539,004
sigma_u	1290,9944					
sigma_e	71,33923					
rho	,99695572	(fraction of variance due to u_i)				

F test that all u_i=0: F(3, 14) = 1178,95 Prob > F = 0,0000

Nu klopt het resultaat: het effect van trouwen is €0.

In een FE-regressie altijd de tijd modeleren. (met een periode of leeftijdeffect).

(RE geeft in dit geval vergelijkbare resultaten.)

fixed-effect panel regressie modellen in Stata voor diverse afhankelijke variabelen

- In Stata (versie 14) zijn fixed-effect modellen te berekenen voor afhankelijke variabelen van verschillend niveau:

	Afhankelijke variabele	Stata syntax
1	Interval	xtreg depvar indepvars , i(idno) fe
2	Dichotoom logit en probit	xtlogit depvar indepvars , i(idno) fe xtprobit depvar indepvars , i(idno) fe
3	Telvariabele poisson en negatief binomiaal	xtpoisson depvar indepvars , i(idno) fe xtnbreg depvar indepvars , i(idno) fe
4	Categoriaal / multinomial logit	femlogit depvar indepvars , group(idno)

Het fixed-effects model als basis

- Het fixed-effects panel regressie model (FE) vormt dus het juiste uitgangspunt en heeft de meest causale interpretatie

- Alleen bij een niet-significante Hausman test mag je het random-effects model gebruiken. De RE-coëfficiënten zullen dan vrijwel nauwelijks afwijken van die van FE (anders was Hausman sign.)

Antwoord op de “brandende vragen”

- “Bij panelanalyse heb je longitudinale gegevens, hoe kan ik gebruik maken van het longitudinale karakter van de data?”
  Gebruik het fixed-effects model (of random-effects bij een niet-significante Hausman test)

- “Ik zie bij panelanalyse het *fixed effects* en het *random effects* model voorbijkomen, wat is het verschil?”
  FE geeft binnenpersoons effect.
  RE houdt het midden tussen het binnenpersoons en het tussenpersoons effect.

- “Kun je met panelanalyse causale relaties vaststellen?”
  Altijd lastige vraag! Afhankelijke van je definitie van causaliteit en het kunnen uitschakelen van alternatieve verklaringen kom je hiermee verder dan met traditionele modellen. Econometristen hechten duidelijk meer waarde aan deze aanpak.