Dr. Christian Czymara

FORSCHUNGSPRAKTIKUM I UND II: LÄNGSSCHNITTDATENANALYSE IN R

Mundlak & Within-Between models
session vi

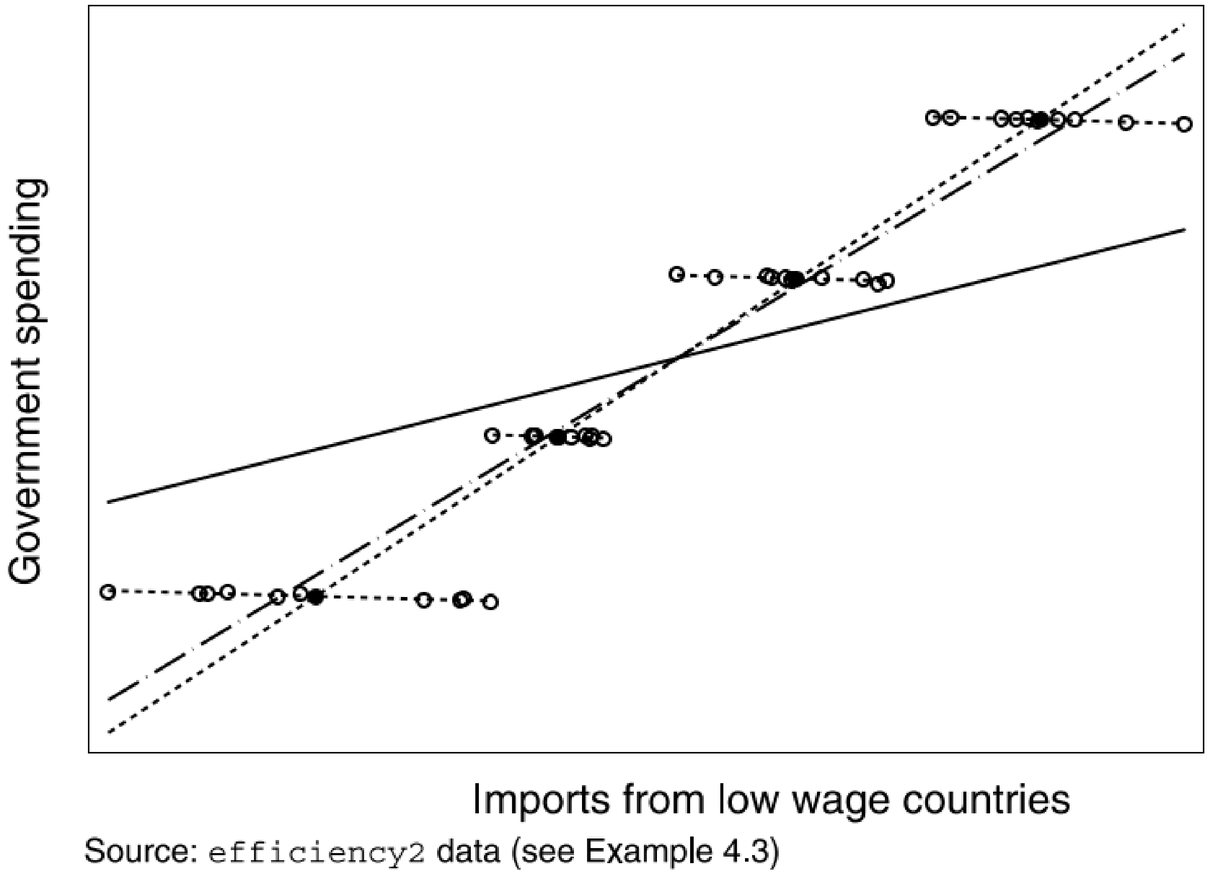
AGENDA

- So far: Fixed (i.e.: within) effects
- Today: Random & between effects
- Modeling within and between effects simultaneously in a RE framework

WITHIN & BETWEEN RELATIONSHIPS

EXAMPLE I

- RQ: Government spending and import (simulated data)
- Countries importing more have higher spending (dashed lines)
- But *within* each country, government spending is associated with lower import rates (dotted lines)
- The "total" relationship is a mix of within and between (solid line)

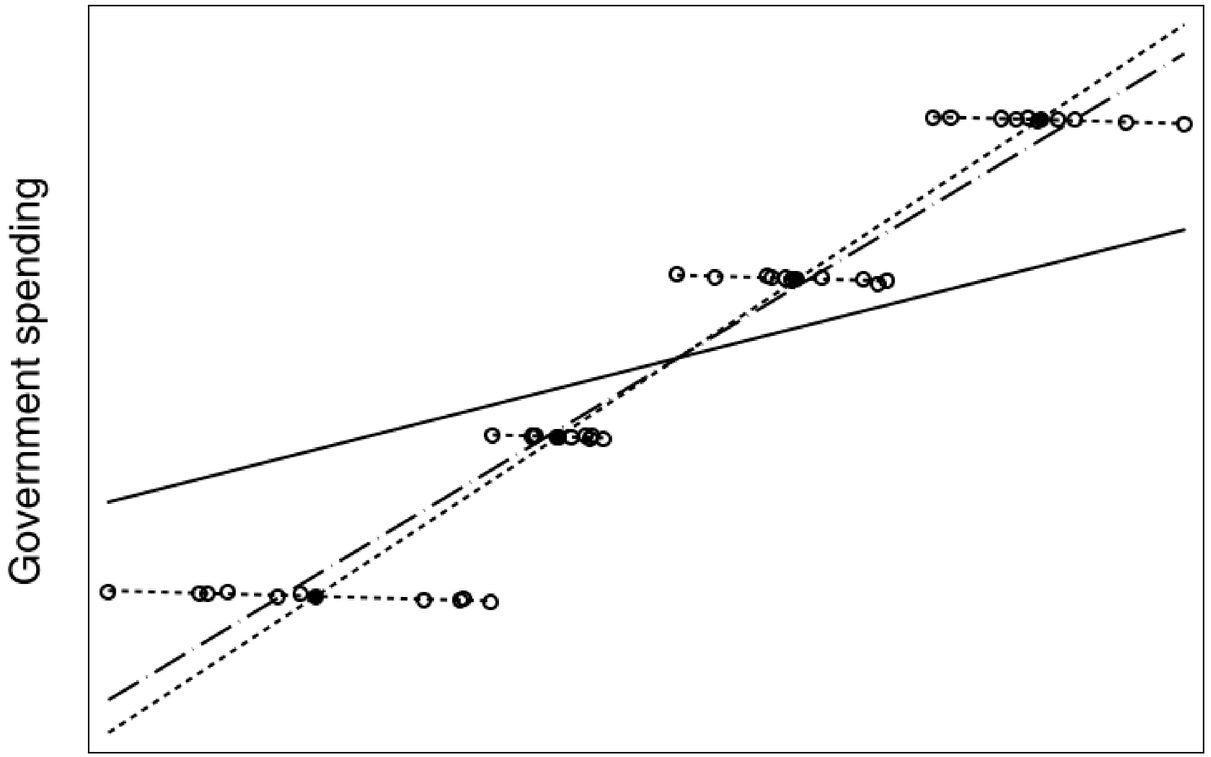


Source: `efficiency2` data (see Example 4.3)

Andreß, Golsch & Schmidt (2014): 161

EXAMPLE I

- Other country factors seem to distort the relationship *between* government spending and imports
- Unobserved heterogeneity on the country level
- Models only based on cross-sectional comparisons will lead to wrong conclusion that spending increases imports (*between*), when it actually seems to decrease it (*within*)



Source: `efficiency2` data (see Example 4.3)

Andreß, Golsch & Schmidt (2014): 161

RANDOM EFFECTS

WHAT IF UNOBSERVED HETEROGENEITY IS NOT PRESENT?

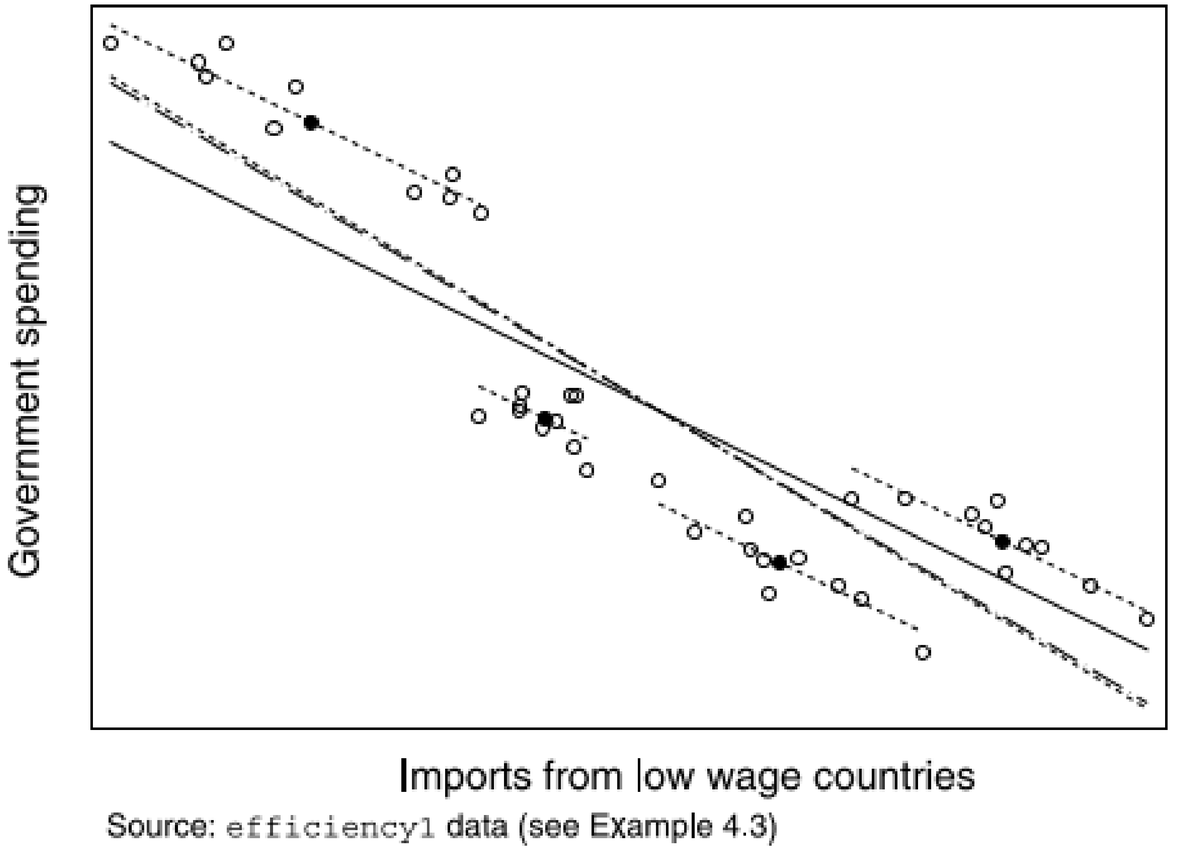
- FE automatically control u_i
- What if u_i does not correlate with x?
- What if there are no time-constant confounders?
- OLS?
  - Biased coefficients?
  - Biased standard errors?

EXAMPLE FROM TEXTBOOK

- Same RQ as before but different data
- In this scenario, government spending is associated with lower import rates within each country (dotted lines)
- And the same holds also between countries (dashed lines)
- Countries importing more have less spending

→Slope of all lines rather similar

→Analysis of within-, between variance, or a combination of both all tell the same story



Andreß, Golsch & Schmidt (2014): 161

STATISTICAL (IN)DEPENDENCE OF PANEL DATA

- If u_i is not important why not use POLS?
- POLS assumes statistical independence of observations (every data points carries new information)
- With panel data, each unit contributes several data points (repeated measurements)

→Inflated sample size

→Underestimation of standard errors (“too significant” effects)

STATISTICAL (IN)DEPENDENCE OF PANEL DATA

- Information of one individual still (assumed) independent of information of other individuals
  - Between individuals
  - Just like cross-sectional data
- But: Information of one individual at time point t not independent from the information of this individual at $t-1$
  - Within individuals
  - Between time points

SERIAL CORRELATION

- Correlation of a variable with itself over time
- For example: unemployment rate in Germany 2016 probably not independent from its rate in 2015
- Put differently, the value of the unemployment rate in 2015 does not carry completely new information when you know the value of 2016
- Similarly, error term at t likely to correlate with error term at $t - 1$
- Also called *autocorrelation* or *serial dependence*

SERIAL CORRELATION

- First order case: Pearson's correlation of y with $t-1$ lag of same variable

→ Example: correlation of y with $L1.y$

→ $r = 0.78; n = 6$

ID	Year	y	$L1.y$
1	2009	0.04	-
1	2010	0.58	0.04
1	2011	0.88	0.58
2	2009	0.22	-
2	2010	0.51	0.22
2	2011	0.66	0.51
3	2009	0.08	-
3	2010	0.43	0.08
3	2011	0.92	0.43

SOLUTION TO SERIAL CORRELATION DUE TO U_i

- Eliminate between variance completely → demeaning / Fixed Effects-transformation → Fixed Effects
- Eliminate only share of between variance related to serial correlation → quasi-demeaning → Random Effects
- Both solve serial correlation due to u_i
- Neither solve serial correlation due to e_{it}

QUASI-DEMEANING

- Fixed Effects-Transformation (demeaning): $(y_{it} - \bar{y}_{i.}) = \beta(x_{it} - \bar{x}_{i.}) + (e_{it} - \bar{e}_{i.})$
- Completely eliminates time constant part
- Random effects transformation (quasi-demeaning): only subtract a part of the unit-specific mean

→ Which part? That which produces serial correlation

→ $(y_{it} - \theta\bar{y}_{i.}) = \beta(x_{it} - \theta\bar{x}_{i.}) + (e_{it} - \theta\bar{e}_{i.}) + \gamma(z_i - \theta z_i) + (u_i - \theta u_i)$

- θ: Demeaning parameter

ASSUMPTIONS

	POLS	RE	FE
Omitted Variable Bias			
Not in e_{it} (strict exogeneity): $cov(e_{it}, x) = 0$	✓	✓	✓
Not in u_i (RE assumption): $cov(u_i, x) = 0$	✓	✓	✗
Serial correlation			
Not in e_{it}: $cov(e_{it}, e_{is}) = 0$	✓	✓	✓
Not in ε_{it}: $corr(\varepsilon_{it}, \varepsilon_{is}) = var(u_i) = 0$	✓	✗	✗

BETWEEN EFFECTS

BETWEEN EFFECTS

- Variance of time-varying y_{it} can be decomposed in
  1. Within variance → $y_{it} - \bar{y}_i$
  2. Between variance → $\bar{y}_i - \bar{\bar{y}}$
- Fixed Effects eliminate all between variance (2.)
- The opposite: Eliminate all within variance (1.)
- … And estimate effects solely based on between variance (2.)

BETWEEN EFFECTS

- Cross-sectional analysis usually only use between unit variance
- How do you model between variation with panel data?
- Remember that time-stable differences between units are captured by unit-specific means ($\bar{y}_{i.}$, $\bar{x}_{i.}$)

→ $\bar{y}_{i.} = \beta_0 + \beta_1 \bar{x}_{1i.} + \cdots + \beta_k \bar{x}_{ki.} + \gamma_1 z_{1i} + \cdots + \gamma_l z_{li} + u_i$

- $\bar{e}_{i.} = u_i$

BE VS. RE VS. FE

- BE ≠ RE
- BE focus on *between* variation only
- FE focus on *within* variation only
- RE and POLS are a *mixture* of BE and FE
- BE > POLS > RE > FE; or BE < POLS < RE < FE (if FE ≠ BE)
- RE and POLS have lowest standard errors because they draw upon the most information (within and between)
- With longer panels, BE have highest standard errors because they draw upon least information (one observation per unit)

USING RE TO COMBINE FE AND BE

FE VS. RE

- FE are *unbiased* if model is correctly specified with respect to *time-varying* characteristics (time-constant aspects automatically controlled)
- FE estimates *less efficient* because between variance not used
- RE are *unbiased* if model is correctly specified with respect to *time-varying* <u>*and*</u> *time-constant* characteristics
- … But more efficient because they draw upon within and between variation
- First and foremost ensure that the model is correctly specified
- … then worry about standard errors

FE AND BE IN ONE MODEL

- RE may include x and z
- z variables yield BE
- RE of x are mixture of FE and BE
- Idea: Decompose the total effect of x into within part and between part

FE AND BE IN ONE MODEL

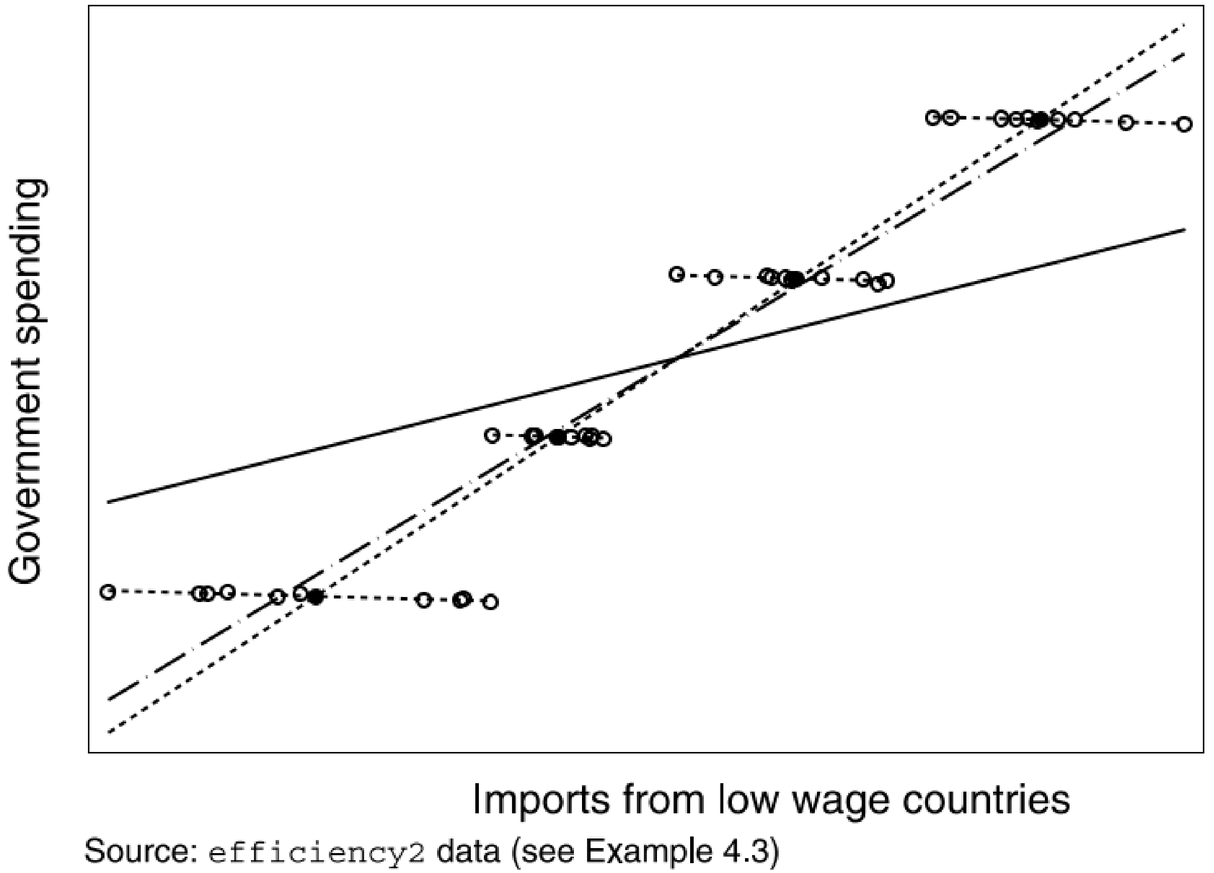
- BE for x variables are captured by their unit-specific means ($\bar{x}_i$)

→Include x and $\bar{x}_i$ to RE model (to be able to include time-constant variables)

- This will control between-unit differences
- The effects of original x variables are thus FE

EXAMPLE

- Simulated data
- Short dashed lines: FE = -0.19
- Solid line: RE = 1.74
- Long dashed line: POLS = 4.14
- Dotted line: BE = 4.53
- Circles: observations
- Black dots: Unit-specific means
- Government spending: $spend$
- Imports from low wage countries: $lowwage$



Source: `efficiency2` data (see Example 4.3)

Andreß, Golsch & Schmidt (2014): 161

INCLUDING $\bar{x}_i$ INTO RE MODEL

- $spend_{it} = 24.66 - 0.19 * lowwage_{it} + 4.72 * \overline{lowwage}_{i.}$
- Model replicates FE estimate (-0.19) for $lowwage$ when $\overline{lowwage}$ included
- $\overline{lowwage}$ nets out time stable differences in import between countries
- $\overline{lowwage}$ yields the *difference between the BE and FE* of $lowwage$ (4.72)
- *Hybrid model type 1* (Andreß et al. 2014)
- Or: *Mundlak model* (e. g.: Bell et al. 2019)

INCLUDING $\bar{x}_i$ AND $\ddot{x}_{it}$ INTO RE MODEL

- $\ddot{x}_{it} = x_{it} - \bar{x}_i$
- $spend_{it} = 24.66 - 0.19 * \ddot{lowwage}_{it} + 4.53 * \overline{lowwage}_{i.}$
- $\ddot{lowwage}_{it}$ is the demeaned variable
- Demeaned variables yield FE (-0.19) because, remember, this is the Fixed Effects-Transformation
- $\overline{lowwage}_{i.}$ now yields BE (4.53)
- *Hybrid model type 2* (Andreß et al. 2014)
- Or: *Random Effect Within-Between model (REWB)* (Bell et al. 2019)

INCLUDING $\bar{x}_i$ AND $\ddot{x}_{it}$ INTO RE MODEL

- $spend_{it} = 24.66 - 0.19 * \ddot{lowwage}_{it} + 4.53 * \overline{lowwage}_{i.}$
- $\bar{x}_i$ and $\ddot{x}_{it}$ are orthogonal (uncorrelated), so are $\ddot{x}_{it}$ and u_i
- Effects of $\ddot{x}_{it}$ thus not biased due to u_i
- Effects of $\bar{x}_i$ might be correlated with u_i and therefore biased (cross-sectional effects)

BENEFITS OF HYBRID MODELING

- FE effects for x and BE for z as well as for x all within one RE model
- Control for u_i but still estimate effects of time-constant variables
- Test for differences between BE and FE estimates

EXAMPLE: JOHNSON & WU (2002): AN EMPIRICAL TEST OF CRISIS, SOCIAL SELECTION, AND ROLE EXPLANATIONS OF THE RELATIONSHIP BETWEEN MARITAL DISRUPTION AND PSYCHOLOGICAL DISTRESS

RESEARCH QUESTION

- Are higher distress levels of the divorced a result of divorce or of social selection?
- Divorce between t_1 and t_2 should increase individual distress

→Longitudinal variation of family status and distress *within individuals*

- When comparing divorced and married individuals, divorced ones should have higher levels of distress

→Cross-sectional variation of family status and distress *between individuals*

→But the divorced likely to differ from the married also in other relevant characteristics

- Are they all measured in the data / can we control them?
- If not: Problem of unobserved heterogeneity

WITHIN AND BETWEEN EFFECTS

- Cross-sectional differences between individuals
  - Based on between variation
  - Between Effects (BE)
  - Likely to be plagued by unobserved heterogeneity
- Longitudinal differences within individuals
  - Based on within variation
  - Fixed Effects (FE) / Within Effects (WE)
  - Automatically controlling unobserved heterogeneity
  - Individuals are “their own controls”
  - Only possible with panel data

DATA

- Outcome: actual stress level (index)
- 2,033 individuals observed 1980, 1983, 1988, 1992
- $n = 1{,}166$, $t = 4$ (original data: unbalanced panel)

VARIABLES

Variable	Label	Time-constant
psydis	psychological distress	
socsel	social selection (divorce experienced before beginning of study)	✓
divorce	divorced	
widow1	widowed	
cohab1	cohabiting	
ager	age	
sexr	gender	✓
educr	educational	

TIME-CONSTANT "EFFECTS"?

- Again, research question: Are higher distress levels of the divorced a result of …
  - Divorce? → Time-varying
  - Or social selection? → Time-constant
- A causal effect of divorce?
  - FE of divorce controls for social selection
  - But other unmeasured time-varying aspects might still be relevant

→Compare size of divorce and of selection effect

→Both need to be estimated

RESULTS

- `socsel` & `sexr`: time-constant
- Original version of time-varying variables
- Unit-specific means
- Demeaned versions

Variable	mundlak	rewb
socsel	.11794022	.11794022
divorce	.16958134***	
widow1	.10211989	
cohab1	-.34746066***	
ager	.00871517***	
educr	-.0234401*	
sexr	-.03514288	-.03514288
mdivorce	.40298295*	.57256429**
mwidow1	.14102847	.24314836
mcohab1	-.41290445	-.76036512*
mager	-.00937816***	-.00066299
meducr	-.01518868	-.03862878***
ddivorce		.16958134***
dwidow1		.10211989
dcohab1		-.34746066***
dager		.00871517***
deducr		-.0234401*
_cons	.47015118***	.47015118***

legend: * p<0.05; ** p<0.01; *** p<0.001

RESULTS

- Between effects (BE)
- Fixed / within effects (FE / WE)
- Differences between BE and FE
- `divorce`
  - Between effect: 0.57**
  - Fixed effect: 0.17***
  - Difference of BE and FE:
    0.57 - 0.17 ≈ -0.4*

Variable	mundlak	rewb
socsel	.11794022	.11794022
divorce	.16958134***	
widow1	.10211989	
cohab1	-.34746066***	
ager	.00871517***	
educr	-.0234401*	
sexr	-.03514288	-.03514288
mdivorce	.40298295*	.57256429**
mwidow1	.14102847	.24314836
mcohab1	-.41290445	-.76036512*
mager	-.00937816***	-.00066299
meducr	-.01518868	-.03862878***
ddivorce		.16958134***
dwidow1		.10211989
dcohab1		-.34746066***
dager		.00871517***
deducr		-.0234401*
_cons	.47015118***	.47015118***

legend: * p<0.05; ** p<0.01; *** p<0.001

TESTING DIFFERENCES BETWEEN WE AND BE

DIFFERENCES IN WE AND BE

- RE are unbiased if BE and WE are essentially the same
- Mundlak model automatically tests for differences between WE and BE (effects of $\bar{x}_i$)
- REWB
  - Effects of $\bar{x}_i$ =BE
  - Effects of $\ddot{x}_{it}$ =WE
  - Test whether $BE = WE$, or $BE - WE = 0$
- Both types of tests numerically equivalent

TESTING FOR DIFFERENCES IN BE AND WE

- Test whether $BE - WE = 0$
- Either for single parameters
  - Mundlak: Test if coefficient of $\bar{x}_i = 0$ (automatically done in regression output)
  - REWB: Test whether difference in coefficients of $\ddot{x}_{it}$ and $\bar{x}_i$ is zero
- … or the model in total
  - Mundlak: Test if coefficient of *all* $\bar{x}_i = 0$
  - REWB: Test whether differences in coefficients of *all* $\ddot{x}_{it}$ and $\bar{x}_i$ are zero
- Overall tests and Hausman test are asymptotically equivalent

DIFFERENCES OF BE AND FE

- If BE and WE do not differ, there are no time-constant confounders in the BE model
- If BE and WE differ, BE are plagued by unobserved heterogeneity
- However, BE might still be interesting, as they might be proxies for "a range of unmeasured social processes, which might include those omitted variables themselves" (Bell et al. 2019: 1059 f.)
- E. g.: "Effect" of ethnicity (time-constant)
  - Not direct causal effect of particular genes
  - Rather, effects of unmeasured social and cultural factors that are related to ethnicity
  - BE can help understanding patterns in the world, but needs theoretical knowledge

DIFFERENCES OF BE AND FE

- There is some relevant but unmeasured time-constant characteristic that is not in the model
- However, can be seen as something of substantive interest
- Opens opportunity for theoretical speculation: W*hat* is it that might be different between units that is relevant? How did prior (cross-sectional) research deal with this issue?
- Depending on time span of panel: BE may indicate "historical" differences which are not captured by rather short-term over time variation

SUMMARY

- Causal claims can more easily be defended with effects based purely on within variation → FE
- In some cases, differences between FE and RE are only marginal
  - Long panel: Relative share of within variation tends to increase
  - "Sluggish" data
  - Good controls: If all confounders are measured, all relevant differences between units can be statistically controlled, no need to turn to within variance

SUMMARY

- Hybrid models combine virtues of FE and RE models
- WE and BE all in one model
- Estimates of z
- Allows to test for differences between WE and BE
  - Are there differences between the two?
  - Test whether these differences are statistically significant (t-test in linear case)
  - Quantify how large the difference is

LITERATURE

- Bell, Fairbrother & Jones (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity* 53 (2). 1051 - 1074.
- Study applying REWB: *Lancee & Sarrasin (2015). Educated preferences or selection effects? A longitudinal analysis of the impact of educational attainment on attitudes towards immigrants. European Sociological Review, 31(4), 490-501*