FAKULTÄT FÜR SOZIALWISSENSCHAFTEN
Lehrstuhl für Statistik und
sozialwissenschaftliche Methodenlehre
Prof. Dr. Josef Brüderl

UNIVERSITÄT MANNHEIM

Longitudinal Data Analysis

RatSWD Nachwuchsworkshop

Vorlesung von Josef Brüderl

25. August, 2009

Longitudinal Data Analysis

- Traditional definition
  - Statistical methods for analyzing data with a time dimension
  - Trend data, event history data, panel data
- Modern definition (Cameron/Trivedi, Microeconometrics)
  - Cross-sectional analysis: inference from between-subject comparison
  - Longitudinal analysis: inference from within-subject comparison
- According to the modern definition
  - are trend data always cross-sectional
  - is traditional event-history analysis also cross-sectional
  - only panel data (repeated observation of the same persons) allow for longitudinal analysis

Panel Data

i	t	y	x
1	1	y_{11}	x_{11}
1	2	y_{12}	x_{12}
2	1	y_{21}	x_{21}
2	2	y_{22}	x_{22}
⋮			
N	1	y_{N1}	x_{N1}
N	2	y_{N2}	x_{N2}

- Repeated measures of one or more variables on one or more persons
- Macroeconomics, Political Science
  - Unit of analysis: countries
  - N small, T large
  - Repeated cross-sectional time-series
- Microeconomics, Sociology
  - Unit of analysis: persons
  - N large, T small
  - Mostly from panel surveys
    - Also from cross-sectional surveys by retrospective questions

Advantages of Panel Data

- Panel data allow for higher precision
  - Due to the higher number of cases (pooling data, N•T)
  - However, in this respect trend data would be even better
- Panel data allow to study individual dynamics
  - Transitions into and out of states (e.g. poverty)
  - Individual growth curves (e.g. wage, materialism, intelligence)
    - Cohort or age effect?
  - Procedure: including age/cohort dummies
- They provide information on the time-ordering of events
  - Causal inference gains strength
  - Procedure: careful data preparation (lags)
- They allow for unobserved heterogeneity
  - Procedure: special statistical models (the rest of this lecture)

Panel Data and Causal Inference I

- Counterfactual approach to causality (Rubin‘s model)

$$Y_{i,t_0}^{T} - Y_{i,t_0}^{C}$$

- With cross-sectional data (between estimation)

$$Y_{i,t_0}^{T} - Y_{j,t_0}^{C}$$

  - Assumption of unit homogeneity (no unobserved heterogeneity)
  - Assumption of conditional independence (no reverse causality)

- With panel data I (within estimation)

$$Y_{i,t_1}^{T} - Y_{i,t_0}^{C}$$

  - Problem: period effects, maturation

- With panel data II (difference-in-differences estimator, DID)

$$(Y_{i,t_1}^{T} - Y_{i,t_0}^{C}) - (Y_{j,t_1}^{C} - Y_{j,t_0}^{C})$$

Panel Data and Causal Inference II

The two major problems in Social Research	Solution with experimental design	Solution with panel design
Self-selection (leading to unobserved heterogeneity)	Randomization	Within estimation (before-after comparison)
Reverse Causality (treatment depends on Y)	Controlled treatment	No simple solution (e.g. no time-varying unobserved heterog.)

- With panel data we can tackle one of the two major problems of Social Research

Panel Data and Causal Inference III

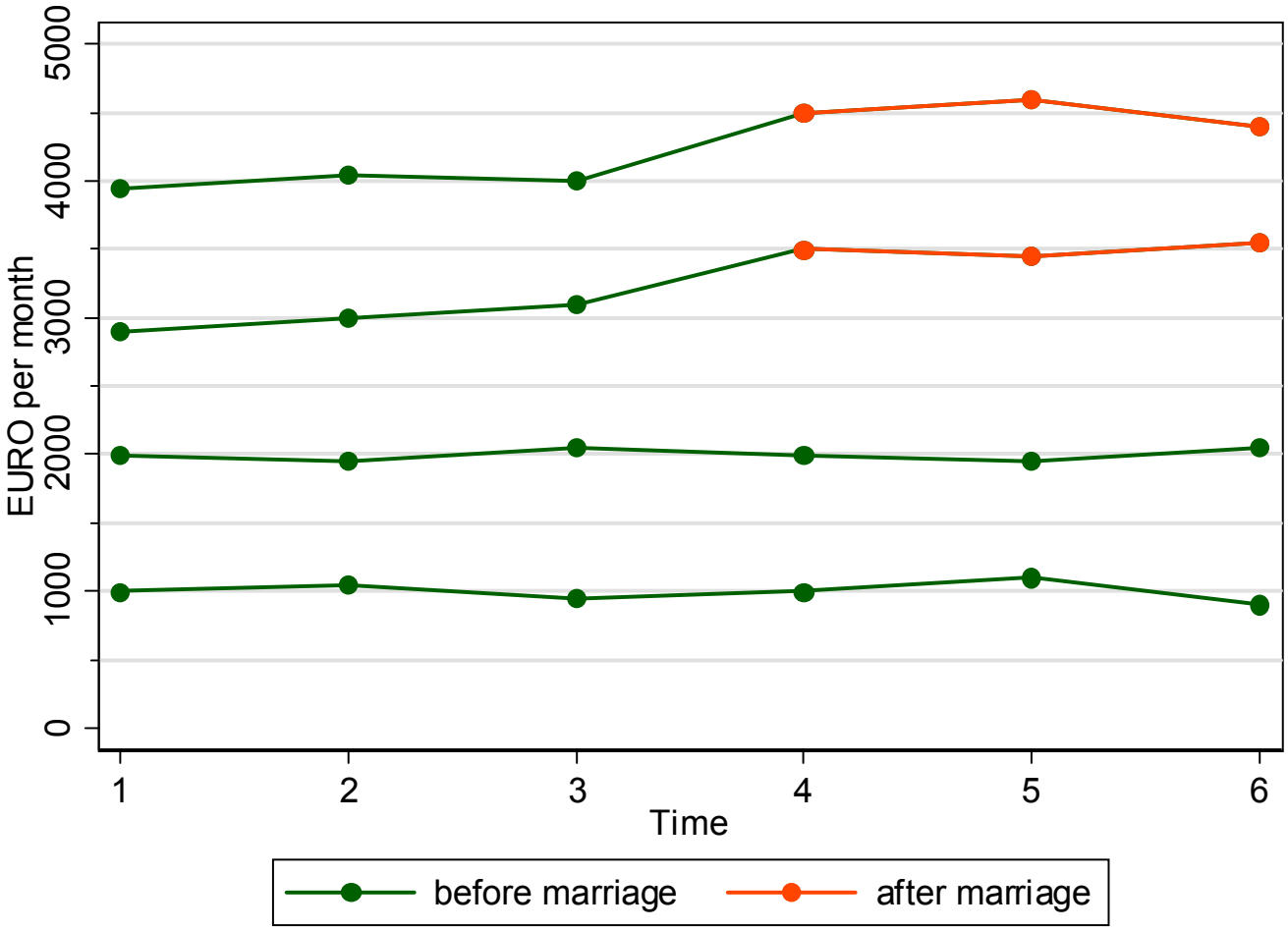
- No self-selection
  - Bivariate analysis suffices
- Self-selection only on observables
  - Cross-sectional regression provides unbiased estimates
  - Even better: Cross-sectional propensity-score matching
- Self-selection also on unobservables
  - Cross-sectional IV-estimation provides unbiased estimates under very strong assumptions
  - Panel regression (fixed-effects regression) provides unbiased estimates under much weaker assumptions

Example: Marriage-Premium for Men?

- Fabricated data (“Wage Premium.dta”): long-format

```
. 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 |
     |-------------------------|      |-------------------------|
```

Example: Marriage-Premium for Men?



There is a causal effect: a marriage-premium

And there is selectivity: Only high wage men marry

Example: Computing the Marriage-Premium

- These data are like experimental data
  - Treatment and control group
  - Before-after comparison
- Compute the DID-estimator

$$\frac{(4500-4000)+(3500-3000)}{2}-\frac{(2000-2000)+(1000-1000)}{2}=500$$

- The marriage-premium is 500 €
- Within-person comparison (the before-after difference)
- To rule out the possibility of maturation or period effects we compare the within-difference of married (treatment) and unmarried (control) men

The Fundamental Problem of Non-Experimental Research

- Result of a cross-sectional regression at T=4:

$$y_{i4} = \beta_0 + \beta_1 x_{i4} + u_{i4}$$

  - Between-comparison: compare wages of married and unmarried men

$$\hat{\beta}_1 = \frac{4500 + 3500}{2} - \frac{2000 + 1000}{2} = 2500$$

- A cross-sectional regression is highly misleading!
  - The bias is due to unobserved heterogeneity
    - High-wage men self-select into marriage
  - Technically: endogeneity (x_{i4} and u_{i4} are correlated)
- Self-selection is the fundamental problem of non-experimental research
  - Most cross-sectional regression results are therefore highly disputable!

No Solution: Pooled-OLS

- Pool the data and estimate an OLS regression (POLS)

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$$

- The result is $\hat{\beta}_1 = 1833$
  - This is the mean of the red points minus the mean of the green points
  - The bias is still heavy
  - POLS also relies on a between comparison. It is thus biased due to unobserved heterogeneity: x_{it} and u_{it} are correlated
- Panel data per se do not remedy the problem of unobserved heterogeneity!
  - One has to use appropriate methods of analysis

A Solution: Panel Data and Within-Estimation

- One has to construct a regression model that relies on the before-after comparison (like DID)
- Starting point: error-components model
  - Person-specific error v_i, idiosyncratic error ε_{it}

$$u_{it} = \nu_i + \varepsilon_{it}$$

  - Error-components model

$$y_{it} = \beta_1 x_{it} + \nu_i + \varepsilon_{it}$$

  - v_i represents person-specific time-constant unobserved heterogeneity (fixed-effects)
    (in our example v_i could be unobserved ability)
- Pooled-OLS has to assume that x_{it} is uncorrelated with both error-components

Fixed-Effects Regression

- How can we get rid of the fixed-effects?
- Within transformation
  - "Time-demeaning" the data

$$y_{it} = \beta_1 x_{it} + \nu_i + \varepsilon_{it} \quad (1)$$

Average over t for each i $\quad \bar{y}_i = \beta_1 \bar{x}_i + \nu_i + \bar{\varepsilon}_i \quad (2)$

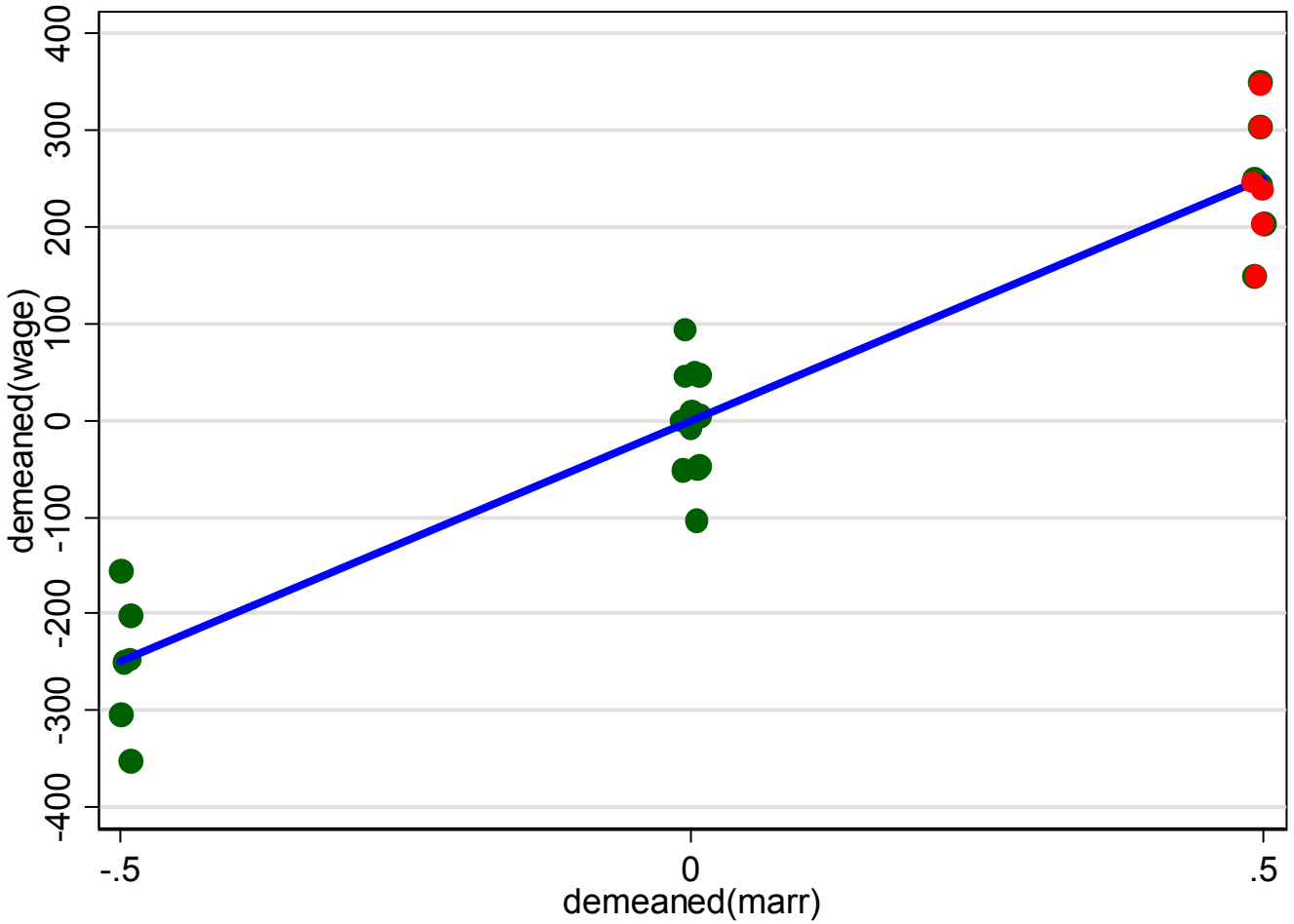
Substract (2) from (1) $\quad y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + \varepsilon_{it} - \bar{\varepsilon}_i \quad (3)$

  - Only within variation is left
  - Pooled OLS (FE-estimator) unbiased, if $Cov(x_{it}, \varepsilon_{it}) = 0$
  - However, $Cov(x_{it}, v_i) \neq 0$ is allowed
    Time-constant unobserved heterogeneity is no longer a problem

Example: Fixed-Effects Regression

```
. xtreg wage marr, fe

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

R-sq:  within  = 0.8982                         Obs per group: min =          6
       between = 0.8351                                        avg =        6.0
       overall = 0.4065                                        max =          6

                                                F(1,19)            =     167.65
corr(u_i, Xb)  = 0.5164                         Prob > F           =     0.0000

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         500   38.61642    12.95   0.000     419.1749    580.8251
       _cons |        2500    16.7214   149.51   0.000     2465.002    2534.998
-------------+----------------------------------------------------------------
     sigma_u |  1290.9944
     sigma_e |  66.885605
         rho |  .99732298   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

“Mechanics” of a FE-Regression



- Those, never marrying are at X=0. They contribute nothing to the regression.
- The slope is determined by the wages of those marrying only:
  It is the difference in the mean wage before and after marriage.

Summary of FE-Estimation

- Panel data and within estimation (DID, FE-regression) can remedy the problem of unobserved heterogeneity
- However, with FE-regressions we cannot estimate the effects of time-constant covariates. These are all cancelled out by the within transformation.
- This reflects the fact that panel data do not help to identify the causal effect of a time-constant covariate!
- The "within logic" applies only with time-varying covariates
  - Something has to “happen” (the effects of events)
  - Only then a before-after comparison is possible

An Example: Male Marital Wage Premium

- Mikrozensus Panel 1996-1999 (Campus-File)
- Analysesample
  - Balanced Sample: nur Personen mit 4 Beobachtungen (bis auf MV)
  - Männer, die 1996 18-40 Jahre alt sind und 1996 ledig sind
- Abhängige Variable
  - Natürlicher Logarithmus des Netto-Monatslohnes (Intervallmitte imputiert)
- Unabhängige Variable
  - Heirats-Dummy (Verheiratet)
- Kontrollvariablen
  - Alterseffekt: Alter und Alter^2
  - Periodeneffekt: Jahres-Dummies
- Panel-robuste Standardfehler

An Example: Male Marital Wage Premium

	POLS	RE-Modell	FE-Modell
Verheiratet	0.19***	0.06	0.00
Alter	0.30***	0.30***	0.31***
$Alter^2$ / 100	-0.42***	-0.43***	-0.42***
Personen	712	712	712
Personenjahre	2636	2636	2636
R^2	0.28	0.11	0.11

Further Readings

- Lecture Notes by Josef Brüderl on Panel and EH Analysis
  - http://www.sowi.uni-mannheim.de/lessm/lehre.html
- Textbooks
  - Wooldridge, J. (2003) Introductory Econometrics. Thomson.
  - Cameron, A.C. and P.K. Trivedi (2005) Microeconometrics.
- Panel Data Analysis
  - Allison, P.D. (2005) Fixed Effects Regression Methods for Longitudinal Data Using SAS. SAS Press.
  - Allison, P.D. (1994) Using Panel Data to Estimate the Effects of Events. Sociological Methods & Research 23: 174-199.
  - Halaby, C. (2004) Panel Models in Sociological Research. Annual Rev. of Sociology 30: 507-544.
- EHA with repeated events
  - Allison, P.D. (1996) Fixed-Effects Partial Likelihood for Repeated Events. Sociological Methods & Research 25: 207-222.