Panel Data Analysis

Josef Brüderl, University of Mannheim, March 2005

This is an introduction to panel data analysis on an applied level using Stata. The focus will be on showing the "mechanics" of these methods. This you will find in no textbook. Panel analysis textbooks are generally full of formulas. For those who see from formulas the "mechanics" this is fine. But most students will need a more basic explanation.

That many students did not fully understand the "mechanics" of panel data analysis, I (also Halaby 2004) infer from many published studies. Researchers use panel data methods (i.e. RE-models) that are not able to fully exploit the main advantage of panel data: getting rid of unobserved heterogeneity.

For those who are interested in the statistical details I recommend:

- Wooldridge, J. (2003) Introductory Econometrics: A Modern Approach. Thomson. Chap. 13, 14. (easy introduction)
- Wooldridge, J. (2002) Econometric Analysis of Cross Section and Panel Data. MIT Press. (more advanced, but very thorough)

Nice introductions on the main issues are:

- 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.

My approach is the "modern" econometric approach. The "classic" approach ("dynamic" panel models) is described in:

- Finkel, S. (1995) Causal Analysis with Panel Data. Sage.

A nice introduction to the increasingly popular "growth curve model" is:

- Singer, J., and J. Willett (2003) Applied Longitudinal Data Analysis. Oxford.

Panel Data

Panel data are repeated measures of one or more variables on one or more persons (repeated cross-sectional time-series).

Mostly they come from panel surveys, however, you can get them also from cross-sectional surveys by retrospective questions.

Panel data record "snapshots" from the life course (continuous or discrete dependent variable). Insofar they are less informative than event-history data.

Data structure ("long" format, $T = 2$):

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}

Benefits of panel data:

- They are more informative (more variability, less collinearity, more degrees of freedom), estimates are more efficient.
- They allow to study individual dynamics (e.g. separating age and cohort effects).
- They give information on the time-ordering of events.
- They allow to control for individual unobserved heterogeneity.

Since unobserved heterogeneity is **the** problem of non-experimental research, the latter benefit is especially useful.

Panel Data and Causal Inference

According to the counterfactual approach on causality (Rubin's model), the causal effect of a treatment (T) is defined by (individual i, time t_0, before treatment C):

$$Y^T_{i,t_0} - Y^C_{i,t_0}.$$

This obviously is not estimable. With cross-sectional data we estimate (between estimation)

$$Y^T_{i,t_0} - Y^C_{j,t_0}.$$

This only provides the true causal effect if the assumption of unit homogeneity (no unobserved heterogeneity) holds. With panel data we can improve on this, by using (within estimation)

$$Y^T_{i,t_1} - Y^C_{i,t_0}.$$

Unit homogeneity here is needed only in an intrapersonal sense! However, period effects might endanger causal inference. To avoid this, we could use (difference-in-differences estimator)

$$(Y^T_{i,t_1} - Y^C_{i,t_0}) - (Y^C_{j,t_1} - Y^C_{j,t_0}).$$

Example: Does marriage increase the wage of men?

Many cross-sectional studies have shown that married men earn more. But is this effect causal? Probably not. Due to self-selection this effect might be spurious. High-ability men select themselves (or are selected) into marriage. In addition, high ability-men earn more. Since researchers usually have no measure of ability, there is potential for an omitted variable bias (unobserved heterogeneity).

I have constructed an artificial dataset, where there is both selectivity and a causal effect:

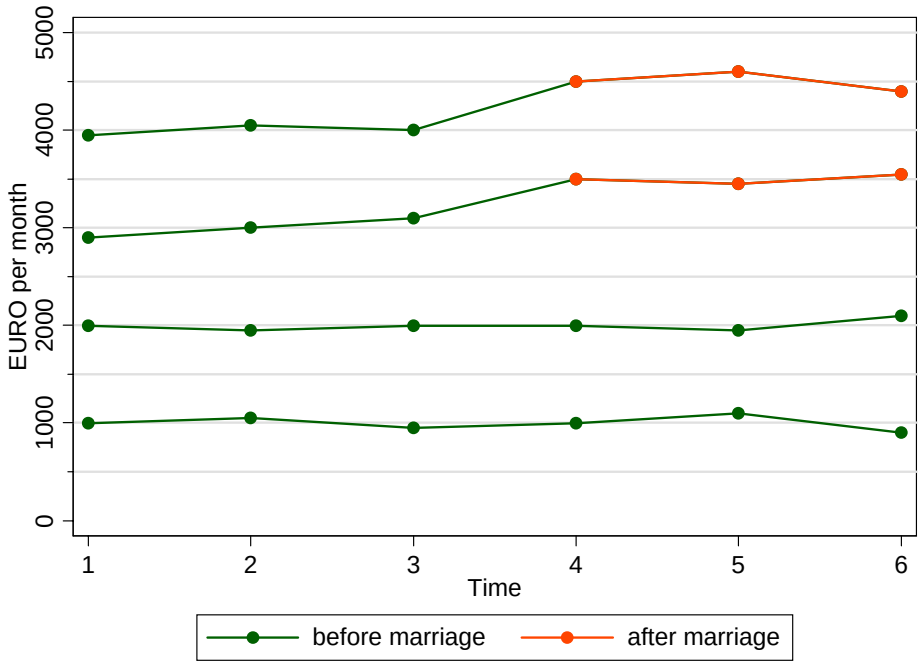
```
. list id time wage marr, separator(6)
```

	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

We observe four men ($N = 4$) for six years ($T = 6$). Wage varies a little around 1000, 2000, 3000, and 4000 € respectively. The two high-wage men marry between year 3 and 4 (selectivity). This is indicated by "marr" jumping from 0 to 1. (This specification implies a lasting effect of marriage).

Note that with panel data we have two sources of variation: between and within persons.

```
twoway
  (scatter wage time, ylabel(0(1000)5000, grid)
   ymtick(500(1000)4500, grid) c(L))
  (scatter wage time if marr==1, c(L)),
   legend(label(1 "before marriage") label(2 "after marriage"))
```



Identifying the Marriage Effect

Now, what is the effect of marriage in these data? Clearly there is self-selection: the high-wage men marry. In addition, however, there is a causal effect of marriage: After marriage there is a wage increase (whereas there is none, for the "control group").

More formally, we use the *difference-in-differences* (DID) approach. First, compute for every man the mean wage before and after marriage (for unmarried men before year 3.5 and after). Then, compute for every man the difference in the mean wage before and after marriage (before-after difference). Take the average of married and unmarried men. Finally, the difference of the before-after difference of married and unmarried men is the causal effect:

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

In this example, marriage causally increases earnings by 500 €. (At least this is what I wanted it to be, but due to a "data construction error" the correct DID-estimator is 483 €. Do you find the "error"?).

The DID-approach mimics what one would do with experimental data. In identifying the marriage effect we rely on a 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

What would be the result of a cross-sectional regression at $T = 4$

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

$\widehat{\beta}_1 = 2500$ (this is the difference of the wage mean between married and unmarried men). The estimate is highly misleading!

It is biased because of unobserved heterogeneity (also called: omitted variables bias): unobserved ability is in the error term, therefore u_{i4} and x_{i4} are correlated. The central regression assumption is, however, that the X-variable and the error term are uncorrelated (exogeneity). Endogeneity (X-variable correlates with the error term) results in biased regression estimates. *Endogeneity is the fundamental problem of non-experimental research*. Endogeneity can be the consequence of three mechanisms: unobserved heterogeneity (self-selection), simultaneity, and measurement error (for more on this see below).

The nature of the problem can be seen in the Figure from above: the cross-sectional OLS estimator relies totally on a *between-person comparison*. This is misleading because persons are self-selected. In an experiment we would assign persons randomly.

With cross-sectional data we could improve on this only, if we would have measured "ability". In the absence of such a measure our conclusions will be wrong. *Due to the problem of unobserved heterogeneity the results of many cross-sectional studies are highly disputable*.

What to Do?

The best thing would be to conduct an experiment. If this is not possible collect at least longitudinal data. As we saw above, *with panel data it is possible to identify the true effect, even in the presence of self-selection*!

In the following, we will discuss several regression methods for panel data. The focus will be on showing, how these methods succeed in identifying the true causal effect. We will continue with the data from our artificial example.

Pooled-OLS

We pool the data and estimate an OLS regression (pooled-OLS)

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

$\widehat{\beta}_1 = 1833$ (the mean of the red points minus the mean of the green points). This estimate is still heavily biased because of unobserved heterogeneity (u_{it} and x_{it} are correlated). This is due to the fact that pooled-OLS also relies on a between comparison. Compared with the cross-sectional OLS the bias is lower, however, because pooled-OLS also takes regard of the within variation.

Thus, *panel data alone do not remedy the problem of unobserved heterogeneity*! One has to apply special regression models.

Error-Components Model

These regression models are based on the following modelling strategy. We decompose the error term in two components: A person-specific error v_i and an idiosyncratic error ϵ_{it},

$$u_{it} = v_i + \epsilon_{it}.$$

The model is now (we omit the constant, because it would be collinear with v_i):

$$y_{it} = \beta_1 x_{it} + v_i + \epsilon_{it}.$$

The person-specific error does not change over time. Every person has a fixed value on this latent variable (fixed-effects). v_i represents person-specific time-constant unobserved heterogeneity. In our example v_i could be unobserved ability (which is constant over the six years).

The idiosyncratic error varies over individuals and time. It should fulfill the assumptions for standard OLS error terms.

The assumption of pooled-OLS is that x_{it} is uncorrelated both with v_i and ϵ_{it}.

First-Difference Estimator

With panel data we can "difference out" the person-specific error ($T = 2$):

$$\begin{aligned} y_{i2} &= \beta_1 x_{i2} + v_i + \epsilon_{i2} \\ y_{i1} &= \beta_1 x_{i1} + v_i + \epsilon_{i1}. \end{aligned}$$

Subtracting the second equation from the first gives:

$$\Delta y_i = \beta_1 \Delta x_i + \Delta \epsilon_i,$$

where "Δ" denotes the change from $t = 1$ to $t = 2$. This is a simple cross-sectional regression equation in differences (without constant). β_1 can be estimated consistently by OLS, if ϵ_{it} is uncorrelated with x_{it} (first-difference (FD) estimator).

The big advantage is that the fixed-effects have been cancelled out. Therefore, we no longer need the assumption that v_i is uncorrelated with x_{it}. *Time-constant unobserved heterogeneity is no longer a problem.*

Differencing is also straightforward with more than two time-periods. For $T = 3$ one could subtract period 1 from 2, and period 2 from 3. This estimator, however, is not efficient, because one could also subtract period 1 from 3 (this information is not used). In addition, with more than two time-periods the problem of serially correlated $\Delta\epsilon_i$ arises. OLS assumes that error terms across observations are uncorrelated (no autocorrelation). This assumption can easily be violated with multi-period panel data. Then S.E.s will be biased. To remedy this one can use GLS or Huber-White sandwich estimators.

Example

We generate the first-differenced variables:

```
tsset id time                        /* tsset the data
generate  dwage = wage - L.wage      /* L. is the lag-operator
generate  dmarr = marr - L.marr
```

Then we run an OLS regression (with no constant):

```
. regress   dwage dmarr, noconstant
  Source |       SS       df       MS          Number of obs =      20
---------+------------------------------       F(  1,    19) =   39.46
   Model |      405000     1      405000        Prob > F      =  0.0000
Residual |      195000    19  10263.1579        R-squared     =  0.6750
---------+------------------------------       Adj R-squared =  0.6579
   Total |      600000    20       30000        Root MSE      =  101.31
------------------------------------------------------------------------------
   dwage |      Coef.   Std. Err.      t    P>|t|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
   dmarr |         450   71.63504     6.28   0.000  300.0661    599.9339
------------------------------------------------------------------------------
```

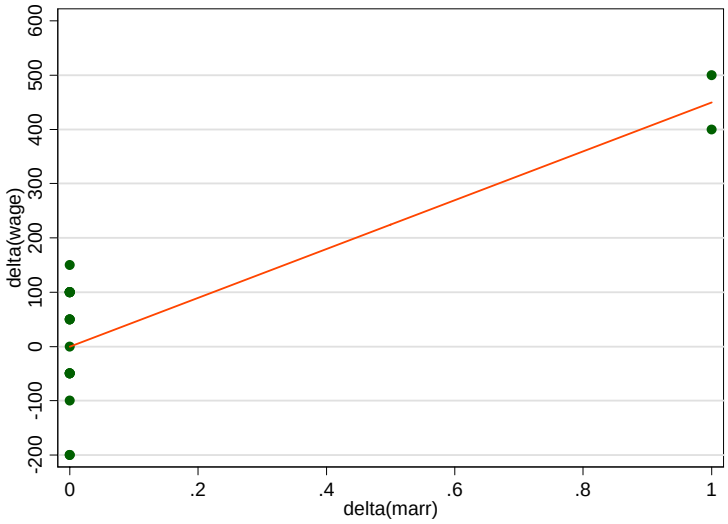
Next with robust S.E. (Huber-White sandwich estimator):

```
. regress   dwage dmarr, noconstant cluster(id)
Regression with robust standard errors               Number of obs =      20
                                                     F(  1,     3) =  121.50
                                                     Prob > F      =  0.0016
                                                     R-squared     =  0.6750
Number of clusters (id) = 4                          Root MSE      =  101.31
------------------------------------------------------------------------------
             |               Robust
   dwage |      Coef.   Std. Err.      t    P>|t|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
   dmarr |         450   40.82483    11.02   0.002  320.0772    579.9228
------------------------------------------------------------------------------
```

We see that the FD-estimator identifies the true causal effect (almost). Note that the robust S.E. is much lower.

However, the FD-estimator is inefficient because it uses only the wage observations immediately before and after a marriage to estimate the slope (i.e., two observations!). This can be seen in the following plot of the first-differenced data:

```
twoway  (scatter dwage dmarr)
        (lfit dwage dmarr, estopts(noconstant)),
        legend(off) ylabel(-200(100)600, grid)
        xtitle(delta(marr)) ytitle(delta(wage));
```



Within-person change

The intuition behind the FD-estimator is that it no longer uses the between-person comparison. It uses only within-person changes: If X changes, how much does Y change (within one person)? Therefore, in our example, unobserved ability differences between persons no longer bias the estimator.

Fixed-Effects Estimation

An alternative to differencing is the within transformation. We start from the error-components model:

$$y_{it} = \beta_1 x_{it} + \nu_i + \epsilon_{it}.$$

Average this equation over time for each i (between transformation):

$$\overline{y}_i = \beta_1 \overline{x}_i + \nu_i + \overline{\epsilon}_i.$$

Subtract the second equation from the first for each t (within transformation):

$$y_{it} - \overline{y}_i = \beta_1 (x_{it} - \overline{x}_i) + \epsilon_{it} - \overline{\epsilon}_i.$$

This model can be estimated by pooled-OLS (fixed-effects (FE) estimator). The important thing is that again the ν_i have disappeared. We no longer need the assumption that ν_i is uncorrelated with x_{it}. *Time-constant unobserved heterogeneity is no longer a problem.*

What we do here is to "time-demean" the data. Again, only the within variation is left, because we subtract the between variation. But here all information is used, the within transformation is more efficient than differencing. Therefore, this estimator is also called the within estimator.

Example

We time-demean our data and run OLS:

```
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              Number of obs =      24
-------+------------------------------           F(  1,    22) =  183.33
 Model |      750000     1      750000           Prob > F      =  0.0000
Resid  |       90000    22  4090.90909           R-squared     =  0.8929
-------+------------------------------           Adj R-squared =  0.8880
 Total |      840000    23  36521.7391           Root MSE      =   63.96

------------------------------------------------------------------------------
 wwage |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-------+----------------------------------------------------------------
 wmarr |         500   36.92745    13.54   0.000   423.4172    576.5828
 _cons |           0   13.05582     0.00   1.000  -27.07612    27.07612
------------------------------------------------------------------------------
```

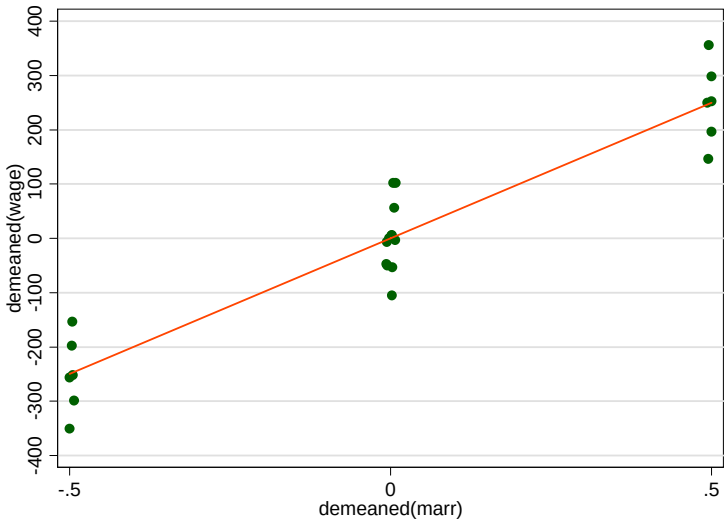
The FE-estimator succeeds in identifying the true causal effect! However, OLS uses $df = N \cdot T - k$. This is wrong, since we used up another N degrees of freedom by time-demeaning. Correct is $df = N \cdot (T-1) - k$. `xtreg` takes regard of this (and demeans automatically):

```
. xtreg wage marr, fe

Fixed-effects (within) regression       Number of obs      =        24
Group variable (i): 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
```

The S.E. is somewhat larger. R^2-within is the explained variance for the demeaned data. This is the one that you would want to report. `sigma_u` is $\hat{\sigma}_v$ and `sigma_e` is $\hat{\sigma}_\epsilon$. By construction of our dataset the estimated person-specific error is much larger. (There is a constant in the output, because Stata after the within transformation adds back the overall wage mean, which is 2500 €.)

An intuitive feeling for what the FE-estimator does, gives this plot:

```
twoway  (scatter wwage wmarr, jitter(2))
        (lfit wwage wmarr),
        legend(off) ylabel(-400(100)400, grid)
        xtitle(demeaned(marr)) ytitle(demeaned(wage));
```



All wage observations of those who married contribute to the slope of the FE-regression. Basically, it compares the before-after wages of those who married. However, it does not use the observations of the "control-group" (those, who did not marry). These are at $X = 0$ and contribute nothing to the slope of the FE-regression.

Dummy Variable Regression (LSDV)

Instead of demeaning the data, one could include a dummy for every i and estimate the first equation from above by pooled-OLS (least-squares-dummy-variables-estimator). This provides also the FE-estimator (with correct test statistics)! In addition, we get estimates for the v_i which may be of substantive interest.

```
. tabulate id, gen(pers)

. regress wage marr pers1-pers4, noconstant

  Source |       SS       df       MS          Number of obs =      24
---------+------------------------------       F(  5,    19) = 8550.00
   Model |   202500000     5    40500000       Prob > F      =  0.0000
Residual |       90000    19  4736.84211       R-squared     =  0.9996
---------+------------------------------       Adj R-squared =  0.9994
   Total |   202590000    24     8441250       Root MSE      =  68.825

------------------------------------------------------------------------------
    wage |      Coef.   Std. Err.      t    P>|t| [95% Conf. Interval]
---------+--------------------------------------------------------------------
    marr |         500   39.73597    12.58   0.000 416.8317    583.1683
   pers1 |        1000   28.09757    35.59   0.000 941.1911    1058.809
   pers2 |        2000   28.09757    71.18   0.000 1941.191    2058.809
   pers3 |        3000   34.41236    87.18   0.000 2927.974    3072.026
   pers4 |        4000   34.41236   116.24   0.000 3927.974    4072.026
------------------------------------------------------------------------------
```

The LSDV-estimator, however, is practical only when N is small.

Individual Slope Regressions

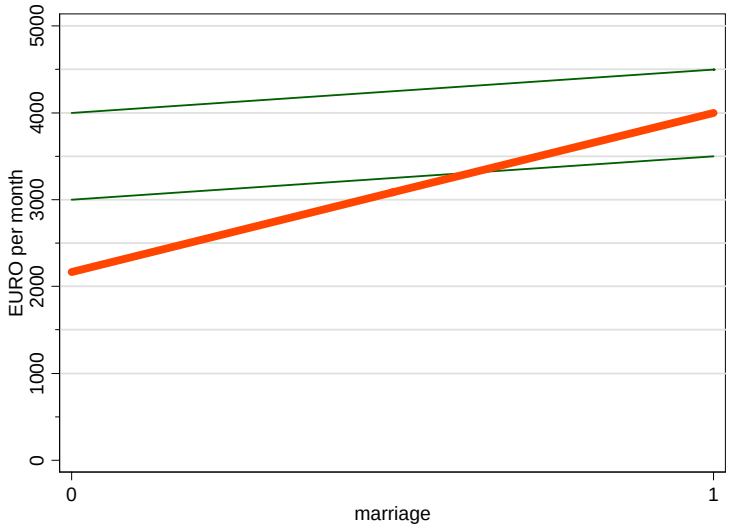
Another way to obtain the FE-estimator is to estimate a regression for every

individual (this is a kind of "random- coefficient model"). The mean of the slopes is the FE-estimator.

Our example: The slopes for the two high-wage men are +500. The regressions for the two low-wage men are not defined, because X does not vary (again we see that the FE-estimator does not use the "control group"). That means, the FE-estimator is +500, the true causal effect.

With `spagplot` (an Ado-file you can `net search`) it is easy to produce a plot with individual regression lines (the red curve is pooled-OLS). This is called a "spaghetti plot":

```
spagplot wage marr if id>2, id(id) ytitle("EURO per month") xlabel(0 1)
         ylabel(0(1000)5000, grid) ymtick(500(1000)4500, grid) note("")
```



Such a plot shows nicely, how much pooled-OLS is biased, because it uses also the between variation.

Restrictions

1. 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 (estimates are only more efficient)! *The "within logic" applies only with time-varying covariates*.
2. Further, there must be some variation in X. Otherwise, we cannot estimate its effect. This is a problem, if only a few observations show a change in X. For instance, estimating the effect of education on wages with panel data is difficult. Cross-sectional education effects are likely to be biased (ability bias). Panel methods would be ideal (cancelling out the unobserved fixed ability effect). But most workers show no change in education. The problem will show up in a huge S.E.

Random-Effects Estimation

Again we start from the error-components model (now with a constant)

$$y_{it} = \beta_0 + \beta_1 x_{it} + \nu_i + \epsilon_{it}.$$

Now it is assumed that the ν_i are random variables (i.i.d. random-effects) and that $Cov(x_{it}, \nu_i) = 0$. Then we obtain consistent estimates by using pooled-OLS. However, we have now serially correlated error terms u_{it}, and S.E.s are biased. Using a pooled-GLS estimator provides the random-effects (RE) estimator.

It can be shown that the RE-estimator is obtained by applying pooled-OLS to the data after the following transformation:

$$(y_{it} - \theta\overline{y}_i) = \beta_0(1-\theta) + \beta_1(x_{it} - \theta\overline{x}_i) + \{(1-\theta)\nu_i + (\epsilon_{it} - \theta\overline{\epsilon}_i)\},$$

where

$$\theta = 1 - \sqrt{\frac{\sigma_\epsilon^2}{T\sigma_\nu^2 + \sigma_\epsilon^2}}\,.$$

If $\theta = 1$ the RE-estimator is identical with the FE-estimator. If $\theta = 0$ the RE-estimator is identical with the pooled OLS-estimator. Normally θ will be between 0 and 1. The RE-estimator mixes within and between estimators! If $Cov(x_{it}, \nu_i) = 0$ this is ok, it even increases efficiency. But if $Cov(x_{it}, \nu_i) \neq 0$ the RE-estimator will be biased. The degree of the bias will depend on the magnitude of θ. If $\sigma_\nu^2 \gg \sigma_\epsilon^2$ then θ will be close to 1, and the bias of the RE-estimator will be low.

Example

```
. xtreg wage marr, re theta

Random-effects GLS regression                   Number of obs      =        24
Group variable (i): 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

Random effects u_i ~Gaussian                    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)
------------------------------------------------------------------------------
```

The RE-estimator works quite well. The bias is only +3. This is because

$$\theta = 1 - \sqrt{\frac{68.8^2}{6 \cdot 706.5^2 + 68.8^2}} = 0.96.$$

The option `theta` includes the estimate in the output.

One could use a Hausman specification test on whether the RE-estimator is biased.

This test, however, is based on strong assumptions which are usually not met in finite samples. Often it does even not work (as with our data).

FE- or RE-Modelling?

1. For most research problems one would suspect that $Cov(x_{it}, v_i) \neq 0$. The RE-estimator will be biased. Therefore, one should use the FE-estimator to get unbiased estimates.
2. The RE-estimator, however, provides estimates for time-constant covariates. Many researchers want to report effects of sex, race, etc. Therefore, they choose the RE-estimator over the FE-estimator. In most applications, however, the assumption $Cov(x_{it}, v_i) = 0$ will be wrong, and the RE-estimator will be biased (though the magnitude of the bias could be low). *This is risking to throw away the big advantage of panel data only to be able to write a paper on "The determinants of Y".* To take full advantage of panel data the style of data analysis has to change: One should concentration on the effects of (a few) time-varying covariates only and use the FE-estimator consequently!
3. The RE-estimator is a special case of a parametric model for unobserved heterogeneity: We make distributional assumptions on the person-specific error term and conceive an estimation method that cancels the nuisance parameters. Generally, *such models do not succeed in solving the problem of unobserved heterogeneity*! In fact, they work only if there is "irrelevant" unobserved heterogeneity: $Cov(x_{it}, v_i) = 0$.

Further Remarks on Panel-Regression

1. Though it is not possible to include time-constant variables in a FE-regression, it is possible to include *interactions* with time-varying variables. E.g., one could include interactions of education and period-effects. The regression coefficients would show, how the return on education changed over periods (compared to the reference period).
2. *Unbalanced panels*, where T differs over individuals, are no problem for the FE-estimator.
3. With panel data there is always reason to suspect that the errors ϵ_{it} of a person i are correlated over time (*autocorrelation*). Stata provides `xtregar` to fit FE-models with AR(1) disturbances.
4. *Attrition* (individuals leave the panel in a systematic way) is seen as a big problem of panel data. However, an attrition process that is correlated with v_i does not bias FE-estimates! Only attrition that is correlated with ϵ_{it} does. In our example this would mean that attrition correlated with ability would not bias results.
5. Panel data are a special case of "*clustered samples*". Other special cases are sibling data, survey data sampled by complex sampling designs (`svy`-regression), and multi-level data (hierarchical regression). Similar models are used in all these literatures.

6. Especially prominent in the multi-level literature is the "random-coefficient model" (regression slopes are allowed to differ by individuals). The panel-analogon where "time" is the independent variable is often named "growth curve model". Growth curve models are especially useful, if you want to analyze how trajectories differ between groups. They come in three variants: the hierarchical linear model version, the MANOVA version, and the LISREL version. Essentially all these versions are RE-models. The better alternative is to use two-way FE-models (see below).
7. Our example investigates the "effect of an event". This is for didactical reasons. Nevertheless, panel analysis is especially appropriate for this kind of questions (including treatment effects). Questions on the effect of an event are prevalent in the social sciences However, panel models work also if X is metric!

Problems with the FE-Estimator

The FE-estimator still rests on the assumption that

$$Cov(x_{it}, \epsilon_{is}) = 0, \qquad \text{for all t and s.}$$

This is the assumption of (strict) exogeneity. If it is violated, we have an *endogeneity problem*: the independent variable and the idiosyncratic error term are correlated. Under endogeneity the FE-estimator will be biased: *endogeneity in this sense is a problem even with panel data*.

Endogeneity could be produced by:

- After X changed there were systematic shocks (period effects)
- Omitted variables (unobserved heterogeneity)
- Random wage shocks trigger the change in X (simultaneity)
- Errors in reporting X (measurement error)

What can we do? The standard answer to endogeneity is to use IV-estimation (or structural equation modelling).

IV-estimation

The IV-estimator (more generally: 2SLS, GMM) uses at least one instrument and identifying assumptions to get the unbiased estimator (`xtivreg`). The identifying assumptions are that the instrument correlates high with X, but does not correlate with the error term. The latter assumption can never be tested! Thus, *all conclusions from IV-estimators rest on untestable assumptions*.

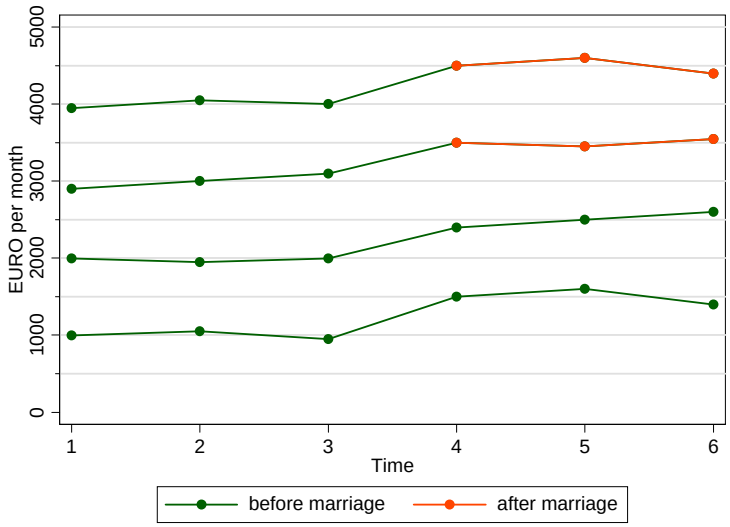
The same is true for another alternative, which is widely used in this context: structural equation modelling (LISREL-models for panel data). Results from these models also rest heavily on untestable assumptions.

Experience has shown that research fields, where these methods have been used abundantly, are full of contradictory studies. *These methods have produced a big mess in social research*! Therefore, do not use these methods!

Period Effects

It might be that after period 3 everybody gets a wage increase. Such a "systematic" shock would correlate with "marriage" and therefore introduce endogeneity. The FE-estimator would be biased. But there is an intuitive solution to remedy this. The problem with the FE-estimator is that it disregards the information contained in the control group. By introducing fixed-period-effects (two-way FE-model) only within variation that is above the period effect (time trend) is taken into regard.

We modify our data so that all men get +500 € at $T \geq 4$. Using the DID-estimator we would no longer conclude that marriage has a causal effect.



Nevertheless, the FE-estimator shows +500. This problem can, however, be solved by including dummies for the time periods (waves!). In fact, including period-dummies in a FE-model mimics the DID-estimator (see below). Thus, it seems to be a good idea to always include period-dummies in a FE-regression!

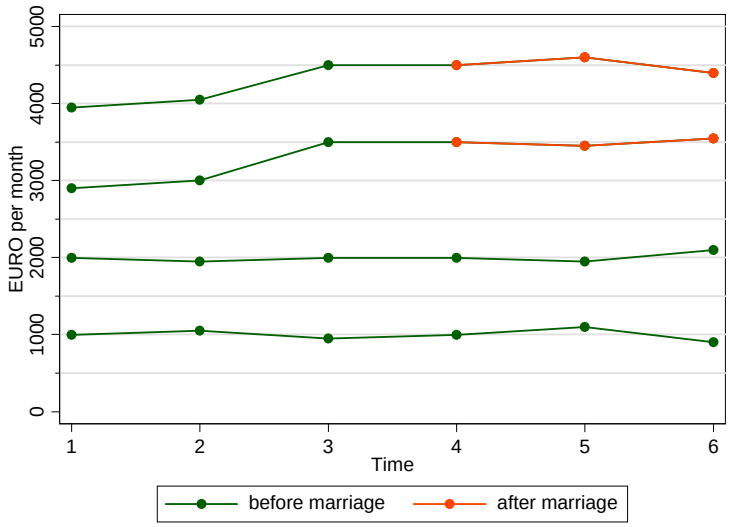
```
. tab time, gen(t)
. xtreg wage3 marr t2-t6, fe

Fixed-effects (within) regression               Number of obs      =        24
Group variable (i): id                          Number of groups   =         4
R-sq:  within  = 0.9498                         Obs per group: min =         6
                                                F(6,14)            =     44.13
corr(u_i, Xb)  = -0.0074                        Prob > F           =    0.0000

------------------------------------------------------------------------------
wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------+-----------------------------------------------------------------------
 marr |  -8.333333   62.28136    -0.13   0.895    -141.9136    125.2469
   t2 |         50   53.93724     0.93   0.370    -65.68388    165.6839
   t3 |         50   53.93724     0.93   0.370    -65.68388    165.6839
   t4 |   516.6667   62.28136     8.30   0.000     383.0864    650.2469
   t5 |   579.1667   62.28136     9.30   0.000     445.5864    712.7469
   t6 |   529.1667   62.28136     8.50   0.000     395.5864    662.7469
_cons |     2462.5   38.13939    64.57   0.000     2380.699    2544.301
------+-----------------------------------------------------------------------
```

Unobserved Heterogeneity

Remember, time-constant unobserved heterogeneity is no problem for the FE-estimator. But, time-varying unobserved heterogeneity is a problem for the FE-estimator. The hope, however, is that most omitted variables are time-constant (especially when T is not too large).

Simultaneity

We modify our data so that the high-ability men get +500 at $T \geq 3$. They marry in reaction to this wage increase. Thus, causality runs the other way around: wage increases trigger a marriage (simultaneity).



The FE-estimator now gives the wrong result:

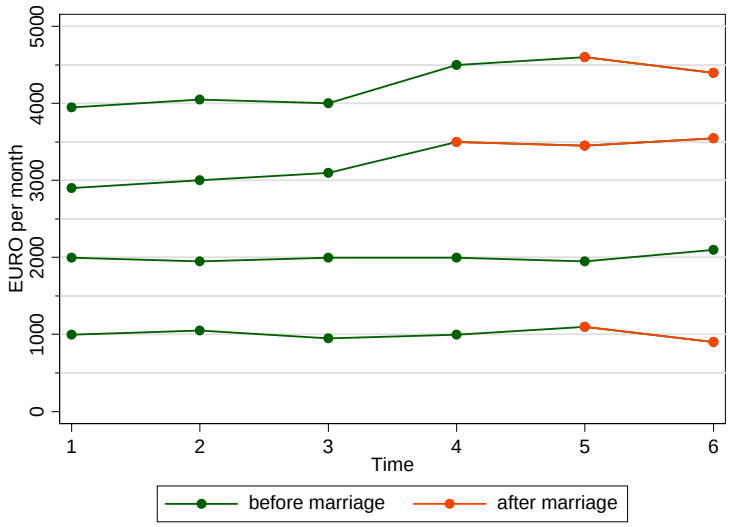
```
. xtreg wage2 marr, fe

Fixed-effects (within) regression       Number of obs      =        24
Group variable (i): id                  Number of groups   =         4
R-sq:  within  = 0.4528                 Obs per group: min =         6
                                        F(1,19)            =     15.72
corr(u_i, Xb)  = 0.5227                 Prob > F           =    0.0008

------------------------------------------------------------------------------
       wage2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        marr |         350   88.27456     3.96   0.001     165.2392    534.7608
       _cons |        2575   38.22401    67.37   0.000     2494.996    2655.004
-------------+----------------------------------------------------------------
```

An intuitive idea could come up, when looking at the data: The problem is due to the fact that the FE-estimator uses all information before marriage. The FE-estimator would be unbiased, if it would use only the wage information immediately before marriage. Therefore, the first-difference estimator will give the correct result. A properly modified DID-estimator would also do the job. Thus, by using the appropriate "within observation window" one could remedy the simultaneity problem. This "hand-made" solution will, however, be impractical with most real datasets (see however the remarks below).

Measurement Errors

Measurement errors in X generally produce endogeneity. If the measurement errors are uncorrelated with the true, unobserved values of X, then $\widehat{\beta}_1$ is biased downwards (attenuation bias).

We modify our data such that person 1 reports erroneously a marriage at $T = 5$ and person 4 "forgets" the first year of marriage. Everything else remains as above, i.e. the true marriage effect is $+500$.




```
. xtreg wage marr1, fe

Fixed-effects (within) regression   Number of obs      =        24
Group variable (i): id              Number of groups   =         4
R-sq:  within  = 0.4464             Obs per group: min =         6

------------------------------------------------------------------------------
wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+------------------------------------------------------------------------
marr1|         300   76.63996     3.91   0.001     139.5907    460.4093
cons |      2537.5   38.97958    65.10   0.000     2455.915    2619.085
-----+------------------------------------------------------------------------
```

The result is a biased FE-estimator of 300 €. Fortunately, the bias is downwards (conservative estimate). Unfortunately, with more X-variables the direction of the bias is unknown.

In fact, compared with pooled-OLS the bias due to measurement errors is amplified by using FD- or FE-estimators, because taking the difference of two unreliable measures generally produces an even more unreliable measure. On the other side pooled-OLS suffers from bias due to unobserved heterogeneity. Simulation studies show that generally the latter bias dominates. The suggestion is, therefore, to use within estimators: unobserved heterogeneity is a "more important" problem than measurement error.

Difference-in-Differences Estimator

Probably you have asked yourselves, why do we not use simply the DID-estimator? After all in the beginning we saw that the DID-estimator works.

The answer is that with our artificial data there is really no point to use the within estimator. DID would do the job. However, with real datasets the DID-estimator is not so straightforward. The problem is how to find an appropriate "control group". This is not a trivial problem.

First, I will show how one can obtain the DID-estimator by a regression. One has to construct a dummy indicating "treatment" (`treat`). In our case the men who married are in the treatment group, the others are in the control group. A second dummy indicates the periods after "treatment", i.e. marriage (`after`, note that `after` has to be defined also for the control group). In addition, one needs the multiplicative interaction term of these two dummies. Then run a regression with these three variables. The coefficient of the interaction term gives the DID-estimator:

```
. gen   treat   = id   >= 3
. gen   after   = time >= 4
. gen   aftertr = after*treat

. regr  wage after treat aftertr

------------------------------------------------
   wage |      Coef.   Std. Err.      t    P>|t|
--------+---------------------------------------
  after |   16.66667   318.5688     0.05   0.959
  treat |   2008.333   318.5688     6.30   0.000
aftertr |   483.3333   450.5244     1.07   0.296
  _cons |   1491.667   225.2622     6.62   0.000
------------------------------------------------
```

The DID-estimator says that a marriage increases the wage by 483 €. This is less than the 500 €, which we thought until now as correct. But it is the "true" effect of a marriage, because due to a "data construction error", there is a 17 € increase in the control group after $T = 3$ (reflected in the coefficient of `after`).

Thus, were the FE-estimates reported above wrong? Yes, the problem of the FE-estimator we saw at several places: it does not use the information contained in the control group. But meanwhile we know how to deal with this problem: we include period-dummies (this is a non-parametric variant of a growth curve model):

```
. xtreg wage marr t2-t6, fe

------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------
        marr |   483.3333    61.5604     7.85   0.000
          t2 |         50   53.31287     0.94   0.364
          t3 |         50   53.31287     0.94   0.364
          t4 |   45.83333    61.5604     0.74   0.469
          t5 |   70.83333    61.5604     1.15   0.269
          t6 |   33.33333    61.5604     0.54   0.597
       _cons |     2462.5   37.69789    65.32   0.000
-------------+----------------------------------------
```

The (two-way) FE-estimator provides the same answer as the DID-estimator! Thus it is more a matter of taste, whether you prefer the DID-estimator or the two-way FE-estimator. Two points in favor of the FE-estimator:

- With real data the FE-estimator is more straightforward to apply, because you do not need to construct a control group.
- As the example shows, the S.E. of the DID-estimator is much larger. This is because there is still "between variance" unaccounted for, i.e. within the two groups.

But the DID-estimator also has an advantage: By appropriately constructing the control group, one could deal with endogeneity problems. Before using the DID-estimator one has to construct a control group. For every person marrying at $T = t$, find another one that did not marry up to t and afterwards. Ideally this "match" should be a "statistical twin" concerning time-varying characteristics, e.g. the wage career up to t (there is no need to match on time-constant variables, because the within logic is applied). This procedure (DID-matching estimator) would eliminate the simultaneity bias in the example from above! (Basically this is a combination of a matching estimator and a within estimator, the two most powerful methods for estimating causal effects from non-experimental data that are currently available.) Usually matching is done via the propensity score, but in the case of matching "life-courses" I would suggest optimal-matching (Levenshtein distance).

Dynamic Panel Models

All panel data models are dynamic, in so far as they exploit the longitudinal nature of panel data. However, there is a distinction in the literature between "static" and "dynamic" panel data models. Static models are those we discussed so far. Dynamic models include a lagged dependent variable on the right-hand side of the equation. A widely used modelling approach is:

$$y_{it} = \delta y_{i,t-1} + \beta_1 x_{it} + v_i + \epsilon_{it}.$$

Introducing a lagged dependent variable complicates estimation very much, because $y_{i,t-1}$ is correlated with the error term(s). Under random-effects this is due to the presence of v_i at all t. Under fixed-effects and within transformation $\Delta y_{i,t-1}$ is correlated with $\overline{\epsilon}_i$. Therefore, both the FE- and the RE-estimator will be biased. If $y_{i,t-1}$ and x_{it} are correlated (what generally is the case, because x_{it} and v_i are correlated) then estimates of both δ **and** β_1 will be biased. Therefore, IV-estimators have to be used (e.g., `xtabond`). As mentioned above, this may work in theory, but in practice we do not know.

Thus, with dynamic panel models there is no way to use "simple" estimation techniques like a FE-estimator. You always have to use IV-techniques or even LISREL. Therefore, *the main advantage of panel data - alleviating the problem of unobserved heterogeneity by "simple" methods - is lost, if one uses dynamic panel*

models!

So why bother with "dynamic" panel models at all?

- These are the "classic" methods for panel data analysis. Already in the 1940ies Paul Lazarsfeld analyzed turnover tables, a method that later was generalized to the "cross-lagged panel model". The latter was once believed to be a panacea for the analysis of cause and effect. Meanwhile several authors concluded that it is a "useless" method for identifying causal effects.
- You may have substantive interest in the estimate of δ, the "stability" effect. This, however, is not the case in "analyzing the effects of events" applications. I suppose there are not many applications where you really have a theoretical interest in the stability effect. (The fact that many variables tend to be very similar from period to period is no justification for dynamic models. This stability also captured in static models (by ν_i).)
- Including $y_{i,t-1}$ is another way of controlling for unobserved heterogeneity. This is true, but this way of controlling for unobserved heterogeneity is clearly inferior to within estimation (much more untestable assumptions are needed).
- Often it is argued that static models are biased in the presence of measurement error, regression toward the mean, etc. This is true as we discussed above (endogeneity). However, dynamic models also have problems with endogeneity. Both with static and dynamic models one has to use IV-estimation under endogeneity. Thus, it is not clear why these arguments should favor dynamic models.

Our example: The pooled-OLS-, RE-, and FE-estimator of the marriage effect in a model with lagged wage are 125, 124, and 495 respectively. All are biased. So I can see no sense in using a "dynamic" panel model.

Conditional Fixed-Effects Models

The FE-methodology applies to other regression models also. With non-linear regression models, however, it is not possible to cancel the v_i by demeaning the data. They enter the likelihood as "nuisance-parameters". However, if there exists a sufficient statistic allowing the fixed-effects to be conditioned out of the likelihood, the FE-model is estimable nevertheless (conditional likelihood). This is not the case for all non-linear regression models. It is the case for count data regression models (`xtpoisson`, `xtnbreg`) and for logistic regression.

For other regression models one could estimate unconditional FE-models by including person-dummies. Unconditional FE-estimates are, however, biased. The bias gets smaller the larger T becomes.

Fixed–Effects Logit (Conditional Logit)

This is a conventional logistic regression model with person-specific fixed-effects v_i:

$$P(y_{it} = 1) = \frac{\exp(\beta_1 x_{it} + v_i)}{1 + \exp(\beta_1 x_{it} + v_i)}.$$

Estimation is done by conditional likelihood. The sufficient statistic is $\sum_{t=1}^{T} y_{it}$ the number of 1s. It is intuitively clear that with increasing v_i the number of 1s on the dependent variable should also increase.

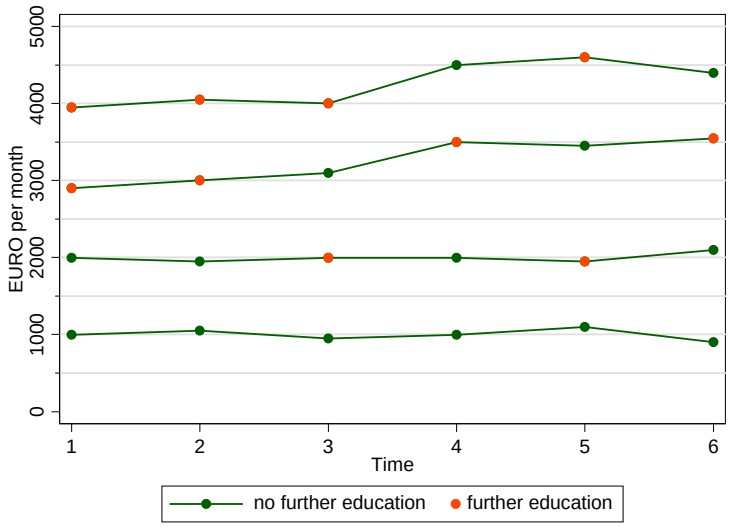
Again, we profit from the big advantage of the FE-methodology: *Estimates of β_1 are unbiased even in the presence of unobserved heterogeneity* (if it is time-constant). It is not possible to estimate the effects of time-constant covariates.

Finally, persons who have only 0s or 1s on the dependent variable are dropped, because they provide no information for the likelihood. This can dramatically reduce your dataset! *Thus, to use a FE-logit you need data with sufficient variance both on X and Y*, i.e. generally you will need panel data with many waves!

Remark: If you have multi-episode event histories in discrete-time, you can analyze these data with the FE-logit. Thus, this model provides FE-methodology for event-history analysis!

Example

We investigate whether a wage increase increases the probability of further education. These are the (artificial) data:



We use the same wage data as above. High-wage people self-select in further education (because their job is so demanding). At $T = 3$ the high-wage people get a wage increase. This reduces their probability of participating in further education a little.

Pooled-logit regression says that with increasing wage the probability of further education (feduc) increases. This is due to the fact that pooled-logit uses the between variation.

```
. logit feduc wage

Logit estimates                           Number of obs   =         24
```

```
                                                LR chi2(1)      =       6.26
                                                Prob > chi2     =     0.0124
Log likelihood = -13.170762                     Pseudo R2       =     0.1920

------------------------------------------------------------------------------
       feduc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |   .0009387   .0004238     2.21   0.027      .000108    .0017693
       _cons |  -2.905459   1.293444    -2.25   0.025    -5.440562   -.3703554
------------------------------------------------------------------------------
```

A FE-logit provides the correct answer: A wage increase reduces the probability of further education.

```
. xtlogit feduc wage, fe

note: multiple positive outcomes within groups encountered.
note: 1 group (6 obs) dropped due to all positive or
      all negative outcomes.

Conditional fixed-effects logistic regression   Number of obs      =  18
Group variable (i): id                          Number of groups   =   3
                                                Obs per group: min =   6
                                                               avg = 6.0
                                                               max =   6

                                               LR chi2(1)        = 1.18
Log likelihood  = -7.5317207                   Prob > chi2       = 0.2764

------------------------------------------------------------------------------
       feduc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        wage |  -.0024463   .0023613    -1.04   0.300    -.0070744    .0021817
------------------------------------------------------------------------------
```

The first note indicates that at least one person has more than once $Y = 1$. This is good, because it helps in identifying the fixed-effects. The second note says that one person has been dropped, because it had in all waves $Y = 0$.

Note that for these effects we can not calculate probability effects, because we would have to plug in values for the ν_i. We have to use the sign or odds interpretation.