Lecture 4: Panel data analysis (I)

Prof. dr. Wolter Hassink
Utrecht University School of Economics
w.h.j.hassink@uu.nl

Contents:

Material:

Wooldridge:
Chapter 13: 13.1, 13.3, 13.4, 13.5

Motivation

- How to estimate the impact of the Covid-outbreak March 2020 on behaviour by individual persons?

- The Netherlands: Lockdown from March 16th 2020 onwards.

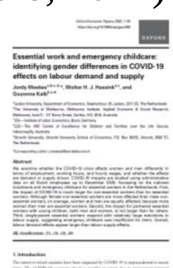
- What data do we need to investigate the impact of the lockdown?

- Requirement 1: We need information from before the lockdown and during the lockdown. Dimension: time

- Requirement 2: We need information from the same persons (or firms) in consecutive periods. Dimension: cross-sectional dimension.

- Methodological claim: empirical analyses that are not based on panel data are in general terms not very strong (= the results can easily be falsified).

- The two figures below give an impression about what happened in 2020 in the Netherlands. (source: Jordy Meekes, Wolter Hassink, Guyonne Kalb, *Oxford Economic Papers*, 2024)

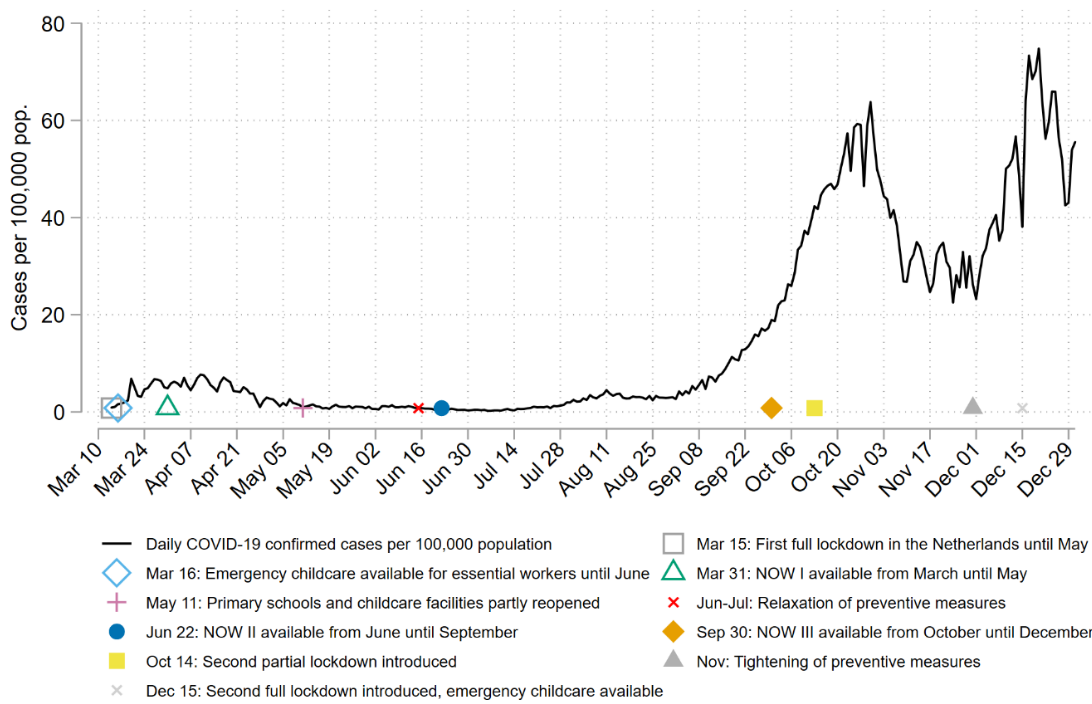


Essential work and emergency childcare: identifying gender differences in COVID-19 effects on labour demand and supply

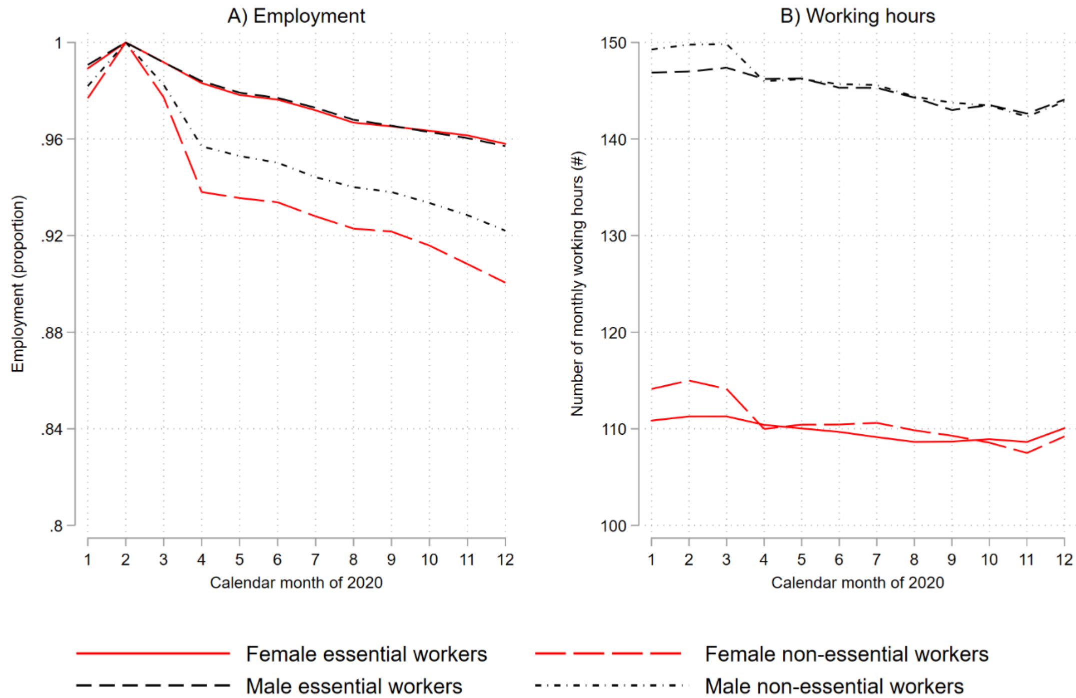
Abstract

1. Introduction

COVID-19 cases and timeline of policies in the Netherlands in 2020



Mean employment rate and monthly working hours in 2020 for employees who were employed in February 2020



Advantage of Panel Data

Advantage of (pseudo)panel over cross section data (I)

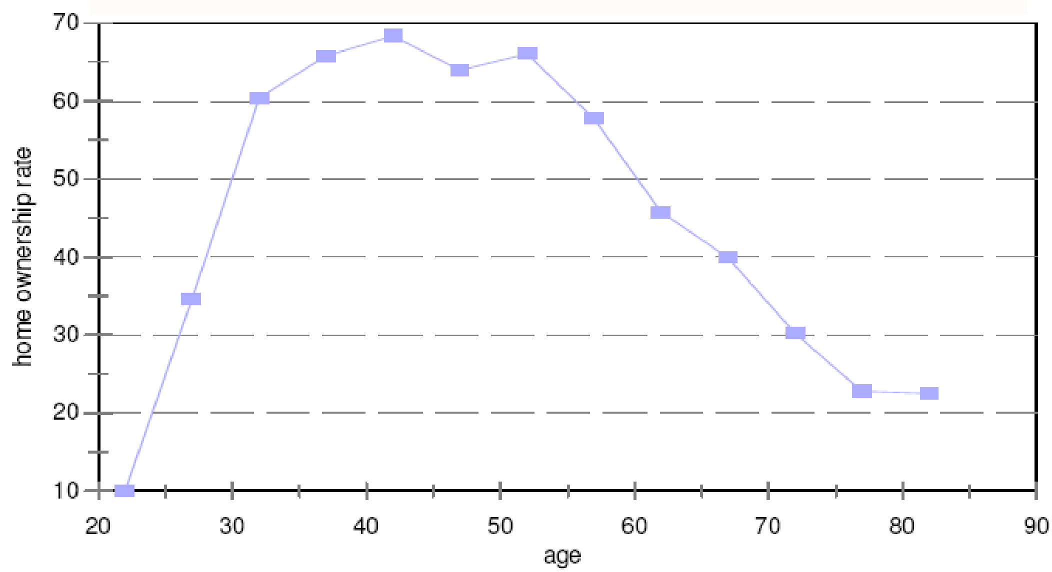
Aim: to motivate the use of panel-data models

- Panel data: the same individuals (households, companies) are followed over time.

Example 1:

- 'year-of-birth' cohorts are followed across time.
- The question is '*do households sell their house when they become old*'?
- The figure below cannot address this question because from one cross-section to another, it is not possible to disentangle cohort effects from age effects.

Figure 1: Home ownership rate by age in 1996 (Netherlands)



- The figure below is constructed by panel data.
- The figure indicates strong cohort effects! For each birth cohort (cohort 1913 to cohort 1968), in various years (t= 1,2,3,…,12) the average Dutch homeownership is given.
- From the cross section it looks like (on average) home ownership rate peaks at around 69% . However, this not necessarily the same for two different cohorts. E.g., compare the 1953 with the 1948 cohort.

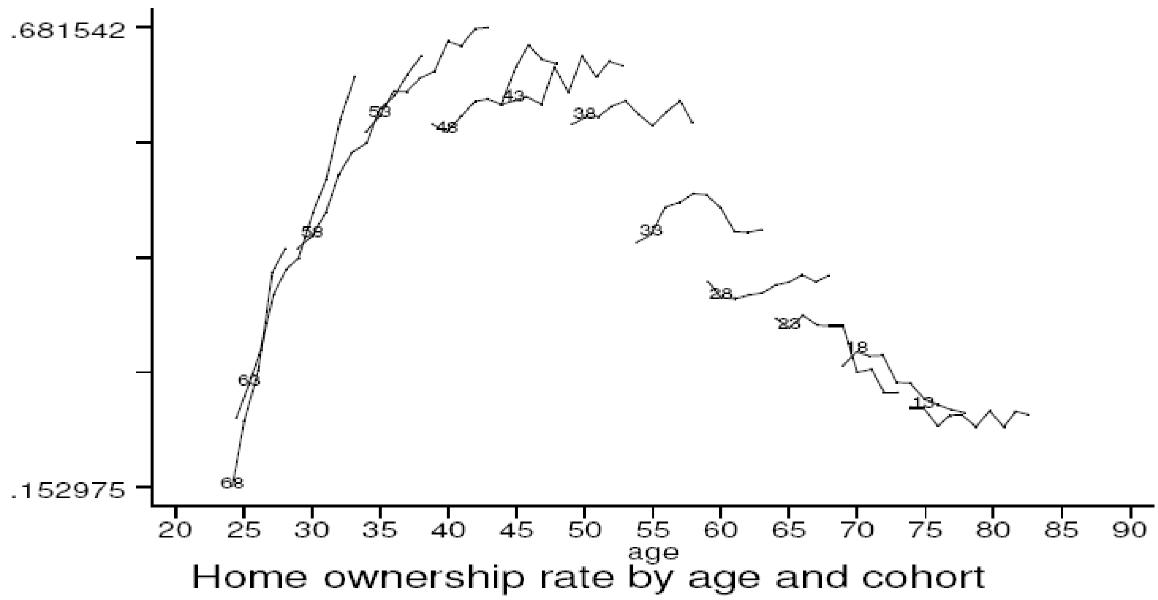


Figure 5a

Advantage of true panel data over (a time series of) cross sections (II)

Aim: to motivate the use of panel-data models

- Estimation of dynamic models (or transition models) is impossible in the case of a time series of cross sections (panel data).

  Example 2: Let's assume that a cross-section study suggests that female labor force participation is equal to 50%.
- There are two extreme possibilities that we cannot distinguish between cross-sections:
  - Possibility 1: 50% of the females are always employed (annual job turnover rate is 0%)
  - Possibility 2: In a homogenous population, there is a 50% turnover rate each year.
- We need panel data to solve this issue.

Advantage of true panel data (III)

Aim: to further motivate the use of panel-data models

- The primary reason for using panel data is to solve the statistical problem of omitted variables. See the figure below. For each of the N individuals (here four individuals) there is a separate scatter diagram.
- The slope of the solid line is the slope of the regression equation of OLS on all data of all individuals together.
- The slope of the dashed line is the slope of the regression equation in which it is corrected for the individual effect.
- In the figures below, the slope of the dashed lines is different from the slope of the solid line.

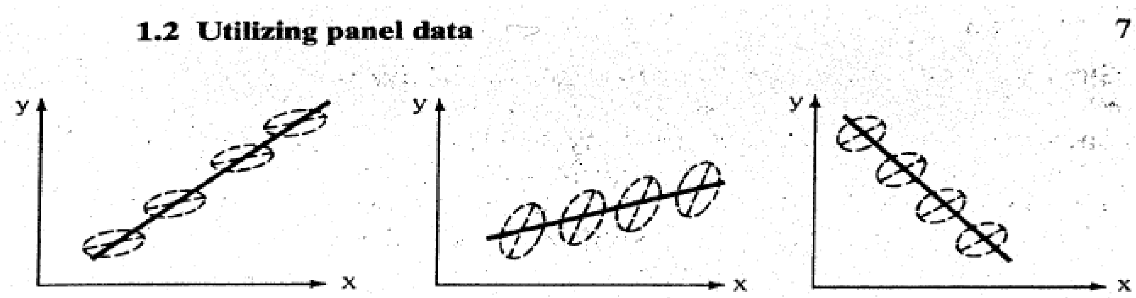


Fig. 1.1 Fig. 1.2 Fig. 1.3

Main features of panel models

Introduction to panel data models

Aim: to relate panel data to cross sections and time series. To introduce the assumption of no correlation between individual-specific effects and explanatory variables.

Note the different subscripts in each of the specifications.

Specification 1: Cross section:
In the first week we considered cross sections:
A random sample of N firms may have the following regression equation:

$$profits = \beta_0 + \beta_1 innovation + \beta_2 \, frmsize + u \qquad (1)$$

- Cross-sectional dimension: N.
- Important: the explanatory variables may be correlated.
- There is one intercept β_0. Equation (1) can be reformulated as by adding a subscript i for the i-th individual firm:

$$profits_i = \beta_0 + \beta_1 innovation_i + \beta_2 \, frmsize_i + u_i \qquad i = 1,...,N \qquad (1)$$

Specification 2: Time series
In the second week we considered the following static time-series model. It is based on a data set containing outcomes for one firm, which is observed over T periods.

$$profits_t = \beta_0 + \beta_1 innovation_t + \beta_2 \, frmsize_t + u_t \qquad t = 1,...,T \qquad (2)$$

- Again, there is one intercept β_0
- In equation (2), we add a subscript t for the t-th period:
- Time dimension: T.

Specification 3: Panel data

Equation (3) is a combination of equations (1) and (2)

$$profits_{it} = \beta_0 + \beta_1 innovation_{it} + \beta_2 \, frmsize_{it} + u_{it} \qquad (3)$$
$$i = 1,...,N; t = 1,...,T$$

- Again, the only intercept is β_0. It has a cross-sectional dimension *N* and a time dimension *T*.
- Subscript *i* refers to individual (firm) and subscript *t* denotes time.
- Important: the explanatory variables *innovation* and *frmsize* may be correlated.
- **Issue 1:** Can equation (3) be generalized by *N* intercepts a_i. Thus each firm (subscript *i*) has its own intercept?

$$profits_{it} = a_i + \beta_1 innovation_{it} + \beta_2 \, frmsize_{it} + u_{it} \qquad (4)$$
$$i = 1,...,N; t = 1,...,T$$

- **Issue 2:** Is there any correlation between these *N* intercepts a_i and each of the explanatory variables *innovation* and *frmsize*?

- **Issue 3:** Should variables that remain constant within individual firms be treated differently? E.g. in the following specification, firm size does not change across time. Thus, $frmsize_i$ has no subscript *t*, in case the size of the firms is constant in all of the *T* periods.

$$profits_{it} = a_i + \beta_1 innovation_{it} + \beta_2 \, frmsize_i + u_{it} \qquad (5)$$
$$i = 1,...,N; t = 1,...,T$$

- **Issue 4**: Are the explanatory variables of the regression equation strictly exogenous? This is an econometric issue that is required for unbiased estimators. It will be explained below further.

The individual-specific effect

The individual-specific effect

Aim: to formalize the use of the unobserved effect a_i

- Suppose the following 'true model'

$$y_{it} = a_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \qquad i = 1,\ldots,N; t = 1,\ldots,T$$

Where:

  - a_i is the individual-specific effect (a random variable)
  - u_{it} is the idiosyncratic (i.i.d.: identically and independently distributed) error term with expected value zero and constant variance.
  - *N*: cross-sectional dimension; *T*: time-dimension
  - There are *k* different explanatory variables.

- a_i captures all individual-specific variables that are not observed by the researcher; e.g. motivation (it is referred to as unobserved heterogeneity).
- It is possible that $E(a_i \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) \neq 0$ (e.g. in an equation where wage is the dependent variable, 'motivation' (subsumed in a_i) might be correlated with the RHS-variable 'experience').
- Suppose that instead one estimates the following model by OLS:

$$y_{it} = \underbrace{\beta_1 x_{1it} + \ldots + \beta_k x_{kit}}_{k \text{ explanatory variables}} + v_{it} \quad i = 1,\ldots,N; t = 1,\ldots,T \qquad (6)$$

- Where $v_{it} = a_i + u_{it}$ (=individual specific effect + idiosyncratic error term)
- In model (1), is it assumed that

$$E(v_{it} \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) = 0 \quad \text{so that}$$

$$E(a_i \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) = 0 \quad \text{and}$$

$$E(u_{it} \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}, a_i) = 0$$

- Violation of this assumption leads to a biased estimate of the regression parameters β .
- An application will be given later.

Econometric issue: what is strict exogeneity?

Strict exogeneity (see also lecture 2)

Aim: to formalize a linear panel-data model

- **Assumption TS.2 (strict exogeneity)**

For each *t*, the expected value of u_t, given **ALL** of the *k* explanatory variables **FOR ALL** *T* time periods, is equal to zero: $E(u_t \mid X) = 0$

- **Assumption TS.2' (contemporaneous exogeneity)**

For each *t*, the expected value of u_t, given **ALL** of the *k* explanatory variables in period *t*, is equal to zero: $E(u_t \mid x_{t1},...,x_{tk}) = E(u_t \mid x_t) = 0$

This assumption implies that the error term in period *t* is uncorrelated with all *k* regressors in the same period, *t*: $Corr(u_t, x_{jt}) = 0 \quad j=1,\ldots,k$

Violation of the strict exogeneity assumption while assumption contemporaneous exogeneity is satisfied

Aim: to reconsider the issue of strict exogeneity for panel data models. Models with a lagged dependent variable (y_{t-1}) or with a feedback mechanism are NOT strictly exogenous.

Example 3: Dynamic variable with lagged dependent variable

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \delta_0 z_t + u_t \qquad (7)$$

- u_t is assumed to be an idiosyncratic error term and contemporaneously exogenous: $E(u_t \mid y_{t-1}, z_t) = 0$, so that OLS yields consistent estimates.
- However, y_{t-1} is NOT a strictly exogenous variable:
    - The assumption of strict exogeneity implies that u_t is uncorrelated not only with $x_t = (y_{t-1}, z_t)'$, but also with $x_{t+1} = (y_t, z_{t+1})'$
    - According to equation (7), y_t and u_t are related to each other. In other words, $E(u_t \mid x_{t+1}) = E(u_t \mid y_t, z_{t+1}) \neq 0$,
- THUS THE LAGGED DEPENDENT VARIABLE y_{t-1} IS NOT STRICTLY EXOGENOUS IN EQUATION (7).

Example 4: Models with a feedback mechanism:

$$gGDP_t = \alpha_0 + \delta_0 r_t + u_t \tag{8}$$

- $gGDP_t$: GDP-growth rate
- r_t: Interest rate, which is assumed to be contemporaneously exogenous: $E(u_t \mid r_t) = 0$
- The independent variable r_t depends on the lagged value of the dependent variable: (feedback mechanism):

$$r_t = \gamma_0 + \gamma_1 (gGDP_{t-1} - 3) + v_t \tag{9}$$

- Equation (9) implies that r_{t+1} depends on $gGDP_t$, and consequently on u_t.
- $E(u_t \mid x_{t+1}) = E(u_t \mid r_{t+1}) \neq 0$
- THUS r_t IS NOT STRICTLY EXOGENOUS IN EQUATION (8)

Between and within variation

Between variation versus within variation?

Aim: to discuss the measurement of between variation and within variation.

Between variation: the cross-sectional variation (across individuals). For example:

	Profit (= dependent variable)	Innovation (=explanatory variable)
Firm A	500 thousand Euros	1 percent
Firm B	750 thousand Euros	3 percent

Between variation (across firms): as a result of the increased innovation (from 1 percent to 3 percent) the profits increase from 500 thousand Euros to 750 thousand Euros.

Within variation: the time-series variation (for a given individual). So the variation within individuals. For example:

	Profit (= dependent variable)	Innovation (=explanatory variable)
Time 1	500 thousand Euros	1 percent
Time 2	550 thousand Euros	3 percent

Within variation (for a given firm): as a result of the increased innovation (from 1 percent to 3 percent (thus by 2 percentage points) the profits increase from 500 thousand Euros to 550 thousand Euros from t=1 to t = 2.

Economists are usually interested in the within variation more than in the between variation.

To measure the within variation of x on y, we need to control for individual effects. Consequently, it allows for correlation between the individual effect and the explanatory variable.

Classification

Static model: a classification

Aim: to classify the different estimation techniques for static panel data models.

- Consider the following equation:

$$y_{it} = a_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \quad i = 1,\ldots,N; t = 1,\ldots,T \qquad (10)$$

- The assumptions in the table on the next slide are necessary to estimate equation (10). It leads to the following two questions:
- **Question 1:** Is there any nonzero correlation between a_i and all of the k RHS vars x_{it}? Is it nonzero in all T time periods?

$$E(a_i \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) = 0 \text{ or}$$

$$E(a_i \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) \neq 0 ?$$

- **Question 2:** Are all of the k variables x_{it} strictly exogenous (conditional on the unobserved individual effect a_i)?

$$E(u_{it} \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}, a_i) = 0$$

(thus, no lagged dependent variables, no feedback mechanism)

- In all of the estimators of this week we assume that the exogenous time-varying regressors are strictly exogenous. (conditional on the unobserved effect):

$$E(a_i \mid \underbrace{x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}}_{\text{all } k \text{ explanatory variables in all } T \text{ time periods}}) = 0$$

    - So again: no lagged dependent variables; no feedback mechanism.

- In explaining estimation procedures, we assume a **balanced panel**: for every cross-sectional unit, we have the same number of time periods *T*.

  Model: $y_{it} = a_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \quad i = 1,\ldots,N; t = 1,\ldots,T \quad (10)$

- Stata allows for estimation of **unbalanced panels**, in which not all units have *T* observations. Thus the *i*-th individual has T_i observations.

Table A: Estimation methods under different assumptions of strict exogeneity and on the correlation between the individual effect and RHS-variables:

	$E(a_i \mid x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}) \neq 0$ Correlation between a_i and all of the explanatory variables is allowed to be nonzero	$E(a_i \mid x_{1it},\ldots,x_{1iT},\ldots,x_{ki1},\ldots x_{kiT}) = 0$ A zero correlation between a_i and all of the explanatory variables is assumed.
All x_{it} strictly exogenous	1. First differences 2. LSDV procedure 3. Within estimation	4. Random effects
Some x_{it} not strictly exogenous	Instrumental Variables (IV)	5. Pooled OLS (no lagged dependent variables) 6. Instrumental variables (IV) (lagged dep. vars. Included)

$$y_{it} = a_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + u_{it} \quad i = 1,\ldots,N; t = 1,\ldots,T$$

Table B: Estimating the effect of time-invariant variables z_i

	No estimation of γ: no effect of z_i on y_{it}	Estimation of γ: effect of z_i on y_{it}
All x_{it} strictly exogenous	1. First differences 2. LSDV procedure 3. Within estimation	4. Random effects
Some x_{it} not strictly exogenous	Instrumental Variables (IV)	5. Pooled OLS (no lagged dependent variables) 6. Instrumental variables (IV) (lagged dep. vars. Included)

$$y_{it} = a_i + \beta_1 x_{1it} + ... + \beta_k x_{kit} + \gamma_1 z_{1i} + ... + \gamma_l z_{li} + u_{it} \quad i = 1,..., N; t = 1,..., T$$

- The variables *z* do not change across time but they are different across individuals
- E.g. z_i: gender and ethnicity in wage equation
- z_i picks up the between variation (between individuals).
  - Between individuals: cross-sectional perspective
- x_{it} picks up the within variation (within individuals)
  - Within individuals: time-series perspective for a given individual
- Economists are usually more interested in within variation.

The first-difference estimator

An estimator for a non-zero correlation between a_i and the explanatory variables

Aim: to introduce an estimation technique for models with a correlation between a_i and the explanatory variables.

- The regression equation is: $y_{it} = a_i + \beta_1 x_{1it} + ... + \beta_k x_{kit} + u_{it}$
- Recall that in regression equations the explanatory variables x_{it} are allowed to be correlated. E.g. education and experience are correlated in a wage equation.
- This is also the case in the fixed effects model. Equation (10) allows for correlation between the explanatory variable a_i and the other explanatory variables $x_{1it}, ... x_{kit}$.

- There are three methods to estimate parameter β consistently:
    - The first-differences estimator (method 1 from Table A; this week).
    - The Least Squares Dummy Variable estimator (LSDV-method). (method 2 from Table A; next week)
    - The within estimator (method 3 from Table A; next week)

Estimation method 1: first-differences estimator

Aim: to introduce the first-difference estimator.

- Let's assume there is only one explanatory variable *x*
- It is possible to consistently estimate the β parameter by taking first differences

  $y_{it} = \beta x_{it} + a_i + u_{it}$ (the regression equation in period *t*) (8a)

  $y_{it-1} = \beta x_{it-1} + a_i + u_{it-1}$ (the regression equation in period *t*-1) (8b)
- Difference of (8a) and (8b):

  $$y_{it} - y_{it-1} = \beta x_{it} - \beta x_{it-1} + a_i - a_i + u_{it} - u_{it-1}$$

  or

  $$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it} \qquad i = 1,...,N; t = 2,...,T \qquad (11)$$
- It means that all variables of equation (11) have the same first-differences transformation Δ
- By taking first differences, the individual effect a_i is removed from the model.
- We calculate the OLS estimator for equation (11): the so-called first-difference estimator of the regression parameter β
- Denote the first-difference estimator by $\hat{\beta}_{fdif}$
- Note that the first period *t*=1 is used for $\Delta y_{i2}, \Delta x_{i2}, \Delta u_{i2}$,
- Why is $\hat{\beta}_{fdif}$ a consistent (or unbiased) estimator? Let's assume for simplicity we have a bivariate model: $y_{it} = x_{it}\beta + a_i + u_{it}$ and $\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}$
- Consistency requires that the error term is uncorrelated with the explanatory variable:

  $Corr(\Delta u_{it}, \Delta x_{it}) = Corr(u_{it} - u_{it-1}, x_{it} - x_{it-1}) = 0$

  It means that u_{it} is uncorrelated with $x_{it-1}, x_{it}, x_{it+1}$. In other words, strict exogeneity is needed (no lagged dependent variable, no feedback effects) for consistency (unbiasedness) of the first-difference estimator.

- Note that contemporaneous exogeneity is too weak to prove consistency of the first-difference estimator $\hat{\beta}_{fdif}$, because it does not exclude consistency between u_{it} and x_{it-1}.
- Furthermore, it is assumed that the error term u_{it} follows a random walk, i.e.:
  - $u_{it} = u_{it-1} + e_{it}$
  - $E(e_{it}) = 0$; $Var(e_{it}) = \sigma_e^2$ (expected value of zero; constant variance)
  - e_{it} is independent over time and across individuals

Then one can show that (here less important):
  - $\hat{e}_{it} = \Delta y_{it} - \Delta x_{it}\hat{\beta}_{fdif}$
  - $\hat{\sigma}_e^2 = \dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2}{N(T-1)-k}$
  - Consistent estimates of $Var(\hat{\beta}_{within})$ (using $\hat{\sigma}^2$)
- In other words, OLS estimation of equation (11) gives the correct standard errors of $\hat{\beta}_{fdif}$.
- Suppose that there is autocorrelation and heteroskedasticity in e_{it}. In that case $Var(\hat{\beta}_{fdif})$ is incorrect, because $\hat{\sigma}_e^2$ is incorrect: robust Newey-West standard errors are necessary.
  - Note that it is a different estimator for $Var(\hat{\beta}_{fdif})$ than the robust standard error (that only corrects for heteroskedasticity)
  - Stata: cluster option: corrects for heteroskedasticity and autocorrelation.

Some useful Stata commands

Aim: to introduce specific Stata commands

- Use the tsset command at the beginning of the do-file:
  - tsset i t
    - 'i' is the name of the variable that refers to the individual
    - 't' is the name of the variable that refers to time
  - alternative: xtset i t

- summary statistics that take into account of both dimensions (individuals and time):
  - xtsum y x

- First differences
  - reg d.y d.x
  - reg d.y d.x, robust
  - reg d.y d.x, cluster(number)
  - Some students make the following mistake:
    ~~xtreg d.y d.x~~

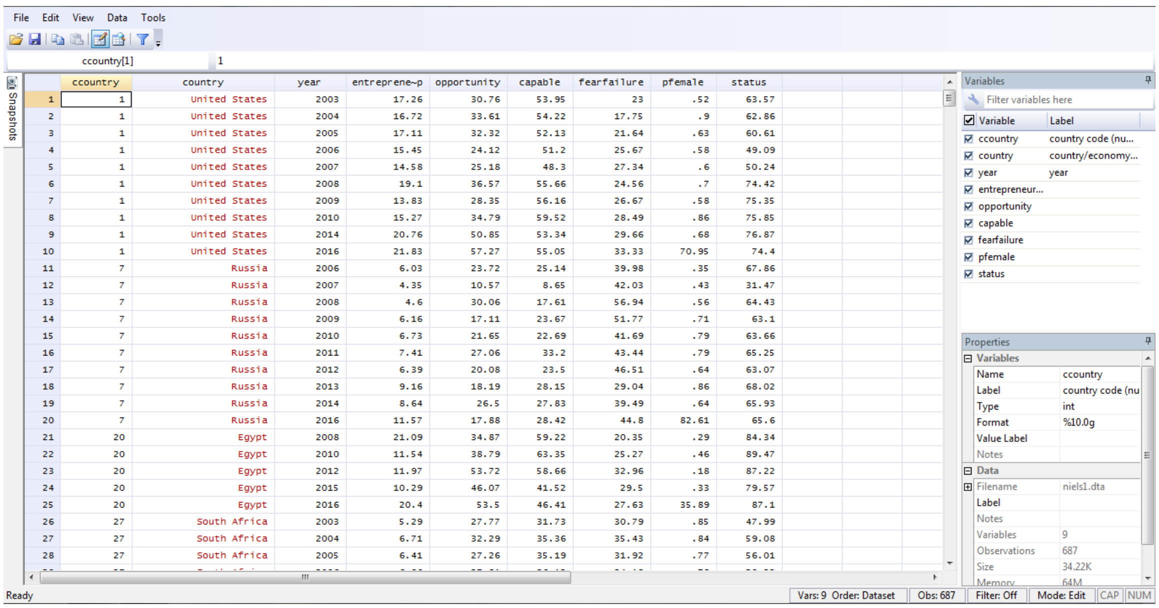
Example 5: xtsum and first differences

niels1.dta

```
. use "niels1.dta", clear
. tabstat year,statistics(n min max) by(country)

Summary for variables: year
     by categories of: country (country/economy        )

         country |         N       min       max
-----------------+------------------------------
         Algeria |         4      2009      2013
          Angola |         4      2008      2014
       Argentina |        13      2003      2016
       Australia |         9      2003      2016
         Austria |         3      2005      2012
      Bangladesh |         1      2011      2011
        Barbados |         4      2011      2015
         Belgium |        13      2003      2015
          Belize |         2      2014      2016
         Bolivia |         3      2008      2014
Bosnia and Herze |         7      2008      2014
        Botswana |         4      2012      2015
          Brazil |        12      2003      2015
        Bulgaria |         2      2015      2016
    Burkina Faso |         2      2015      2016
...
  United Kingdom |        14      2003      2016
   United States |        10      2003      2016
         Uruguay |        11      2006      2016
         Vanuatu |         1      2010      2010
       Venezuela |         4      2003      2011
         Vietnam |         3      2013      2015
           Yemen |         1      2009      2009
          Zambia |         3      2010      2013
-----------------+------------------------------
           Total |       687      2003      2016
------------------------------------------------
```

```
. sort ccountry year

. xtset ccountry year
       panel variable:  ccountry (unbalanced)
        time variable:  year, 2003 to 2016, but with gaps
                delta:  1 unit
```

Structure of the dataset:



	ccountry	country	year	entreprene~p	opportunity	capable	fearfailure	pfemale	status
1	1	United States	2003	17.26	30.76	53.95	23	.52	63.57
2	1	United States	2004	16.72	33.61	54.22	17.75	.9	62.86
3	1	United States	2005	17.11	32.32	52.13	21.64	.63	60.61
4	1	United States	2006	15.45	24.12	51.2	25.67	.58	49.09
5	1	United States	2007	14.58	25.18	48.3	27.34	.6	50.24
6	1	United States	2008	19.1	36.57	55.66	24.56	.7	74.42
7	1	United States	2009	13.83	28.35	56.16	26.67	.58	75.35
8	1	United States	2010	15.27	34.79	59.52	28.49	.86	75.85
9	1	United States	2014	20.76	50.85	53.34	29.66	.68	76.87
10	1	United States	2016	21.83	57.27	55.05	33.33	70.95	74.4
11	7	Russia	2006	6.03	23.72	25.14	39.98	.35	67.86
12	7	Russia	2007	4.35	10.57	8.65	42.03	.43	31.47
13	7	Russia	2008	4.6	30.06	17.61	56.94	.56	64.43
14	7	Russia	2009	6.16	17.11	23.67	51.77	.71	63.1
15	7	Russia	2010	6.73	21.65	22.69	41.69	.79	63.66
16	7	Russia	2011	7.41	27.06	33.2	43.44	.79	65.25
17	7	Russia	2012	6.39	20.08	23.5	46.51	.64	63.07
18	7	Russia	2013	9.16	18.19	28.15	29.04	.86	68.02
19	7	Russia	2014	8.64	26.5	27.83	39.49	.64	65.93
20	7	Russia	2016	11.57	17.88	28.42	44.8	82.61	65.6
21	20	Egypt	2008	21.09	34.87	59.22	20.35	.29	84.34
22	20	Egypt	2010	11.54	38.79	63.35	25.27	.46	89.47
23	20	Egypt	2012	11.97	53.72	58.66	32.96	.18	87.22
24	20	Egypt	2015	10.29	46.07	41.52	29.5	.33	79.57
25	20	Egypt	2016	20.4	53.5	46.41	27.63	35.89	87.1
26	27	South Africa	2003	5.29	27.77	31.73	30.79	.85	47.99
27	27	South Africa	2004	6.71	32.29	35.36	35.43	.84	59.08
28	27	South Africa	2005	6.41	27.26	35.19	31.92	.77	56.01

```
. tab year

   year     |      Freq.     Percent        Cum.
------------+-----------------------------------
       2003 |         31        4.51        4.51
       2004 |         34        4.95        9.46
       2005 |         35        5.09       14.56
       2006 |         42        6.11       20.67
       2007 |         41        5.97       26.64
       2008 |         43        6.26       32.90
       2009 |         54        7.86       40.76
       2010 |         58        8.44       49.20
       2011 |         47        6.84       56.04
       2012 |         58        8.44       64.48
       2013 |         66        9.61       74.09
       2014 |         63        9.17       83.26
       2015 |         54        7.86       91.12
       2016 |         61        8.88      100.00
------------+-----------------------------------
      Total |        687      100.00
```

```
. describe

Contains data from niels1.dta
  obs:           687
 vars:             9                          29 Aug 2017 13:54
 size:        34,350
------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------
ccountry        int     %10.0g                country code (numeric)
country         str22   %22s                  country/economy
year            int     %10.0g                year
entrepreneurs~p float   %9.0g
opportunity     float   %9.0g
capable         float   %9.0g
fearfailure     float   %9.0g
pfemale         float   %9.0g
status          float   %9.0g
------------------------------------------------------------------------------------------
Sorted by:  ccountry  year
     Note:  dataset has changed since last saved
```

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    ccountry |       687    241.3901    276.1463          1        995
     country |         0
        year |       687    2010.323    3.863612       2003       2016
entreprene~p |       687    18.75245    11.20719       3.27      75.29
 opportunity |       687    40.05189    16.22148       2.85      85.54
-------------+--------------------------------------------------------
     capable |       687     49.5062    15.31129       8.65      87.93
 fearfailure |       687    34.49453    9.001162      10.43      75.42
     pfemale |       687    6.443654    19.40803        .05     123.81
      status |       687    69.54263    10.72374      31.47        100

. xtsum

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
ccountry overall |  241.3901   276.1463          1        995 |     N =     687
         between |             312.7506          1        995 |     n =     106
         within  |                    0   241.3901   241.3901 | T-bar = 6.48113
                 |                                            |
country  overall |         .          .          .          . |     N =       0
         between |                    .          .          . |     n =       0
         within  |                    .          .          . |     T =       .
                 |                                            |
year     overall |  2010.323   3.863612       2003       2016 |     N =     687
         between |             2.223647       2004       2016 |     n =     106
         within  |             3.438053    2002.18   2018.523 | T-bar = 6.48113
                 |                                            |
entrep~p overall |  18.75245   11.20719       3.27      75.29 |     N =     687
         between |             12.70803       6.09      75.29 |     n =     106
         within  |             4.312329   5.345778   37.99411 | T-bar = 6.48113
                 |                                            |
opport~y overall |  40.05189   16.22148       2.85      85.54 |     N =     687
         between |             15.44536   8.769167      84.13 |     n =     106
         within  |             7.770043    14.8242   82.21989 | T-bar = 6.48113
                 |                                            |
capable  overall |   49.5062   15.31129       8.65      87.93 |     N =     687
         between |             15.51558   13.12167   86.21667 |     n =     106
         within  |             5.001509   26.30159   69.98335 | T-bar = 6.48113
                 |                                            |
fearfa~e overall |  34.49453   9.001162      10.43      75.42 |     N =     687
         between |             9.658087   14.94333      72.01 |     n =     106
         within  |             5.303271   14.42453   66.01453 | T-bar = 6.48113
                 |                                            |
pfemale  overall |  6.443654   19.40803        .05     123.81 |     N =     687
         between |              10.9578   .1366667       46.6 |     n =     106
         within  |             18.12183  -39.28635   104.6797 | T-bar = 6.48113
                 |                                            |
status   overall |  69.54263   10.72374      31.47        100 |     N =     687
         between |             10.26139   47.97333        100 |     n =     106
         within  |             5.853495   37.07571   89.32264 | T-bar = 6.48113
```

First-differences

```
. reg d.entrepreneurship d.opportunity d.capable d.fearfailure d.pfemale d.status
      Source |       SS       df       MS              Number of obs =     483
-------------+------------------------------           F(  5,   477) =    6.13
       Model |  580.703558     5  116.140712           Prob > F      =  0.0000
    Residual |  9033.89875   477  18.9389911           R-squared     =  0.0604
-------------+------------------------------           Adj R-squared =  0.0505
       Total |   9614.6023   482  19.9473077           Root MSE      =  4.3519
------------------------------------------------------------------------------
D.           |
entreprene~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 opportunity |
         D1. |   .0085752   .0285149     0.30   0.764    -.0474551    .0646056
             |
     capable |
         D1. |   .1838622   .0405581     4.53   0.000     .1041675    .2635569
             |
 fearfailure |
         D1. |  -.0413421   .0306715    -1.35   0.178      -.10161    .0189257
             |
     pfemale |
         D1. |   .0066679   .0100145     0.67   0.506    -.0130101    .0263459
             |
      status |
         D1. |   .0005231   .0338882     0.02   0.988    -.0660656    .0671117
             |
       _cons |   .2839998   .2079547     1.37   0.173    -.1246207    .6926203
------------------------------------------------------------------------------
. predict vhat, resid
(204 missing values generated)

. reg vhat l.vhat d.opportunity d.capable d.fearfailure d.pfemale d.status
      Source |       SS       df       MS              Number of obs =     371
-------------+------------------------------           F(  6,   364) =   13.54
       Model |  948.903762     6  158.150627           Prob > F      =  0.0000
    Residual |   4250.3595   364  11.6768118           R-squared     =  0.1825
-------------+------------------------------           Adj R-squared =  0.1690
       Total |  5199.26327   370  14.0520629           Root MSE      =  3.4171
------------------------------------------------------------------------------
        vhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        vhat |
         L1. |  -.3907954   .0437721    -8.93   0.000    -.4768734   -.3047174
             |
 opportunity |
         D1. |   .0065333   .0250719     0.26   0.795    -.0427707    .0558372
             |
     capable |
         D1. |   .0185389    .039742     0.47   0.641    -.0596139    .0966917
             |
 fearfailure |
         D1. |   .0211696   .0299744     0.71   0.480    -.0377752    .0801144
             |
     pfemale |
         D1. |  -.0109633   .0088292    -1.24   0.215     -.028326    .0063995
             |
      status |
         D1. |  -.0002112   .0319314    -0.01   0.995    -.0630045     .062582
             |
       _cons |   .0687234   .1865105     0.37   0.713    -.2980501    .4354968
------------------------------------------------------------------------------
```

```
. reg d.entrepreneurship d.opportunity d.capable d.fearfailure d.pfemale
d.status, cluster(ccountry)

Linear regression                                      Number of obs =     483
                                                       F(  5,    83) =    2.79
                                                       Prob > F      =  0.0221
                                                       R-squared     =  0.0604
                                                       Root MSE      =  4.3519

                              (Std. Err. adjusted for 84 clusters in ccountry)
------------------------------------------------------------------------------
D.           |               Robust
entreprene~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 opportunity |
         D1. |   .0085752   .0310053     0.28   0.783    -.0530932    .0702436
             |
     capable |
         D1. |   .1838622   .0521886     3.52   0.001     .0800613    .2876631
             |
 fearfailure |
         D1. |  -.0413421   .0297418    -1.39   0.168    -.1004975    .0178132
             |
     pfemale |
         D1. |   .0066679   .0080202     0.83   0.408     -.009284    .0226197
             |
      status |
         D1. |   .0005231   .0408331     0.01   0.990    -.0806923    .0817384
             |
       _cons |   .2839998   .1632255     1.74   0.086    -.0406493    .6086488
------------------------------------------------------------------------------
```

Estimation procedure:

- We started with the first-difference estimator $\hat{\beta}_{fdif}$
- Next, we checked for autocorrelation, using the Breusch Godfrey test.
- The parameter on the lagged residual was statistically different from zero (t-value: -8.93).
- We re-estimated the model with the first-difference estimator, using clustered standard errors.

Pooled OLS

An estimator for a zero correlation between a_i and the explanatory variables

Aim: to introduce the estimator that assumes there is no correlation between a_i and the explanatory variables

- Now we assume that there is no correlation between a_i and x_{it}.
- Consider the following methods:
  - Random effects (method 4 from Table A; next week)
  - Pooled OLS (method 5 from Table A)

Estimation method 5: Pooled OLS

Aim: to introduce the OLS-estimator for a panel data specification.

- Now we assume that there is no correlation between a_i and all of the *k* explanatory variables x_{it}.
- Equation (12) can be estimated with OLS. This is referred to as pooled OLS:
- $y_{it} = \beta_1 x_{1it} + ... + \beta_k x_{kit} + v_{it}$ (12)
  - For which $v_{it} = a_i + u_{it}$ where u_{it} is i.i.d.
  - As a result of the identical distribution: $Var(v_{it}) = Var(v_{it-1})$
  - In a pooled regression $v_{it} = a_i + u_{it}$
  - So, in equation (12), it is assumed that:

$$E(a_i \mid x_{1it},...,x_{kit}) = 0 \quad \text{and}$$
$$E(u_{it} \mid x_{1it},...,x_{1iT},...,x_{ki1},...x_{kiT}, a_i) = 0$$

  - a_i is uncorrelated with all of the explanatory variables.
  - u_{it} is uncorrelated with all of the explanatory variables and a_i

- Autocorrelation between the error terms v_{it} and v_{it-1} of equation (12) is:
- $$Corr(v_{it}, v_{it-1}) = \frac{Cov(v_{it}, v_{it-1})}{\sqrt{Var(v_{it})}\sqrt{Var(v_{it-1})}} = \frac{Cov(v_{it}, v_{it-1})}{Var(v_{it})}$$

- We can show that the numerator:

$$Cov(v_{it}, v_{it-1}) = Cov(a_i + u_{it}, a_i + u_{it-1}) =$$
$$= \underbrace{Cov(a_i, a_i)}_{=\sigma_a^2} + \underbrace{Cov(a_i, u_{it-1})}_{=0} + \underbrace{Cov(u_{it}, a_i)}_{=0} + \underbrace{Cov(u_{it}, u_{it-1})}_{=0}$$
$$= Var(a_i) + 0 + 0 + 0$$

$$= \sigma_a^2$$

- And the denominator:
  $Var(v_{it}) = Var(a_i + u_{it})$
  $$= \underbrace{Var(a_i)}_{=\sigma_a^2} + \underbrace{Var(u_{it})}_{=\sigma_u^2} + 2\underbrace{Cov(a_i, u_{it})}_{=0} =$$
  $$= \sigma_a^2 + \sigma_u^2$$
- So that: $\dfrac{Cov(v_{it}, v_{it-1})}{Var(v_{it})} = \dfrac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}$

- Conclusion: In pooled OLS there is always autocorrelation.
- The estimation procedure for pooled OLS: OLS on (12), with Newey-West robust standard errors, which is also referred to as clustered standard errors. It corrects for both heteroskedasticity and autocorrelation. Stata command Pooled OLS: reg y x, cluster(i)

Example 5 (continued): pooled OLS

```
. reg entrepreneurship status

      Source |       SS       df       MS              Number of obs =     687
-------------+------------------------------           F(  1,   685) =   74.47
       Model |  8448.43236     1  8448.43236           Prob > F      =  0.0000
    Residual |  77713.8823   685  113.450923           R-squared     =  0.0981
-------------+------------------------------           Adj R-squared =  0.0967
       Total |  86162.3147   686  125.601042           Root MSE      =  10.651

------------------------------------------------------------------------------
entrepreneurship |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      status |   .3272499   .0379224     8.63   0.000     .2527919     .401708
       _cons |  -4.005375   2.668347    -1.50   0.134    -9.244497    1.233746
------------------------------------------------------------------------------

. reg entrepreneurship opportunity capable fearfailure pfemale status

      Source |       SS       df       MS              Number of obs =     687
-------------+------------------------------           F(  5,   681) =  112.58
       Model |  38989.9631     5  7797.99262           Prob > F      =  0.0000
    Residual |  47172.3516   681  69.2692388           R-squared     =  0.4525
-------------+------------------------------           Adj R-squared =  0.4485
       Total |  86162.3147   686  125.601042           Root MSE      =  8.3228

------------------------------------------------------------------------------
entrepreneurship |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 opportunity |    .136798   .0261802     5.23   0.000     .0853944    .1882015
     capable |   .3941343   .0286309    13.77   0.000     .3379188    .4503497
 fearfailure |   .1020699    .039778     2.57   0.011     .0239677    .1801721
     pfemale |   .0167714   .0164528     1.02   0.308    -.015533     .0490758
      status |   .0545589   .0331565     1.65   0.100    -.0105423    .1196601
       _cons |  -13.66175   2.719823    -5.02   0.000      -19.002   -8.321506
------------------------------------------------------------------------------

. predict uhat, resid

. reg uhat l.uhat opportunity capable fearfailure pfemale status

      Source |       SS       df       MS              Number of obs =     483
-------------+------------------------------           F(  6,   476) =  131.42
       Model |  15829.8654     6  2638.31091           Prob > F      =  0.0000
    Residual |  9555.84622   476  20.0753072           R-squared     =  0.6236
-------------+------------------------------           Adj R-squared =  0.6188
       Total |  25385.7117   482  52.6674516           Root MSE      =  4.4805

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        uhat |
         L1. |   .7642528   .0273884    27.90   0.000     .7104357    .8180699
             |
 opportunity |   .0001926   .0167499     0.01   0.991    -.0327203    .0331055
     capable |  -.0089727   .0187121    -0.48   0.632    -.0457413     .027796
 fearfailure |  -.0412445   .0275817    -1.50   0.135    -.0954414    .0129524
     pfemale |   .0008219   .0103745     0.08   0.937    -.0195636    .0212074
      status |  -.0243557   .0218146    -1.12   0.265    -.0672206    .0185091
       _cons |   3.616266   1.851214     1.95   0.051    -.0212952    7.253827
------------------------------------------------------------------------------
```

```
. reg entrepreneurship opportunity capable fearfailure pfemale status,
cluster(ccountry)

Linear regression                                      Number of obs =     687
                                                       F(  5,   105) =   14.91
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4525
                                                       Root MSE      =  8.3228

                             (Std. Err. adjusted for 106 clusters in ccountry)
------------------------------------------------------------------------------
             |               Robust
entreprene~p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 opportunity |    .136798   .0378957     3.61   0.000     .0616578    .2119381
     capable |   .3941343   .0576298     6.84   0.000      .279865    .5084035
 fearfailure |   .1020699   .0913241     1.12   0.266    -.0790089    .2831487
     pfemale |   .0167714   .0142713     1.18   0.243    -.0115259    .0450686
      status |   .0545589   .0619831     0.88   0.381    -.0683422      .17746
       _cons |  -13.66175   6.119695    -2.23   0.028    -25.79598   -1.527528
------------------------------------------------------------------------------
```

Estimation procedure:

- We started with the pooled OLS estimator.
- Next, we checked for autocorrelation, using the Breusch Godfrey test.
- The parameter on the lagged residual was statistically different from zero (t-value: 27.90).
- We re-estimated the model with the pooled OLS estimator, using clustered standard errors.

Winding up

Issues:

- Within variation versus between variation.
- Within effects versus between effects.
- Estimator for within effects: the first-difference estimator.
- Advantage: we correct for unobserved effects, which are allowed to correlate with the explanatory variables.
- Estimator for between effects: Pooled OLS estimator; we need to compute clustered standard errors.