Introduction to Panel Data Analysis

Oliver Lipps / Ursina Kuhn

Swiss Centre of Expertise in the Social Sciences (FORS)
c/o University of Lausanne

Lugano Summer School, 2016
Introduction panel data, data management
1 Introducing panel data (OL)
2 The SHP (UK)
3 Data Management with Stata (UK)

Regressions with panel data: basic
4 Regression refresher (UK)
5 Causality (OL)
6 Fixed effects models (OL)
7 Random Effects (random intercept) models (OL)
8 Nonlinear regression (UK)

Additional topics
9 Missing data (OL), 10 Random Slope models (OL)
11 Dynamic models (UK)
Organisation

**Morning** (8.30-12.30, break ca 10.30-10.45)
- Classroom
- Theory and application examples

Lunch Break 12.30-13.30

**Afternoon** (13.30-17.00, break ca. 15.30-15.45)
- Hands-on; aim: apply what we discussed in the morning
  - Data management and descriptive analysis with panel data
  - Regression models with panel data
Prepared data sets and exercises or work with your own data
Discussion of individual questions whenever possible
Purpose of this Summer School

To introduce basic methods of panel data analysis:

• Emphasis on **causal effects** (within variation) but also **descriptive** methods (OLS)
• **Less emphasis on complex methods** (dynamic models, instruments)
• Practical implementation with **Stata**, do-files (and data)
• Data preparation

• Presenting and **interpreting** results
• **Graphical** display of regression results
Surveys over time: repeated cross-sections vs. panels

• **Cross-Sectional Survey**: conducted at one or several points in time ("rounds") using **different respondents in each round**

• **Panel Survey**: conducted at several points in time ("waves") using **the same sample respondents over waves**
  → panel *data* mostly from prospective (panel) surveys
  → also: from retrospective ("biographical") survey
Panel Surveys: to distinguish

Length and sample size:
• **Time Series**: $N$ small (mostly=1), $T$ large ($T \to \infty$)
  → time series models (finance, macro-economics, demography, …)
• **Panel Surveys**: $N$ large, $T$ small ($N \to \infty$)
  → social science panel surveys (sociology, microeconomics, …)

Sample
• **General population**:
  - rotating: only few (pre-defined number) waves per individual (in CH: SILC, LFS)
  - indefinitely long (in CH: SHP)
• **Special population**:
  - e.g., age/birth cohorts (in CH e.g.: TREE, SHARE, COCON)
    representative for population of special age group / birth years
Panel surveys increasingly important

Changing focus in social sciences

→ Life course research: individual trajectories (e.g., growth curves, transitions into and out of states)

→ Identify “causal effects” (unbiased estimates) rather than correlations

→ Large investments in social science panel surveys, high data quality!

Analysis potential of panel data
- close to experimental design: before and after studies of treated
- Control of unobserved time-invariant individual characteristics (FE Models)
-> individual dynamics can only be measured with panel data!
Identification of age, time, and (birth) cohort effects

Fundamental relationship: \( a_{it} = t - c_i \) (eg 30 = 2014 - 1984)

- Effects from “formative” years (childhood, youth) -> cohort effect (e.g. taste in music)
- Time may affect behavior -> time effect (e.g. computer performance, economic cycle)
- Behavior may change over the life cycle -> age effect (e.g. health)

- In a cross-section, t is constant
  \( \rightarrow \) age and cohort collinear (only joint effect estimable)
- In a cohort study, cohort is constant
  \( \rightarrow \) age and time collinear (only joint effect estimable)
- In a panel, \( A_{it}, t, \) and \( c_i \) collinear.
  \( \rightarrow \) only two of the three effects can be estimated
  \( \rightarrow \) we can use \( (t,c_i), (A_{it},c_i), \) or \( (A_{it},t), \) but not all three
Problems of panel surveys

Fieldwork / data quality related

- **High costs** (panel care, tracking households, incentives):
  → increasing number of online panel surveys (randomly selected) e.g., LISS Panel, GiP, GESIS panel, ELIPSS, UK – GenPopWeb initiative

- **Initial nonresponse** (wave 1) and **attrition** (=drop-out of panel after wave 1):
  → increasing efforts (sampling frame in CH, incentives, tracking, questionnaire modularization, …)

- **Panel conditioning** effects (details largely unknown)

- ...

- Finally: you design a panel for the next generation …
2
Introducing the Swiss Household Panel (SHP)
Swiss Household Panel: overview

• Primary goal: observe social change and changing life conditions in Switzerland

• First wave in 1999, more than 5,000 households. Refreshment samples in 2004, more than 2,500 households, several new questions, and in 2013 (more than 4,500 households, full questionnaire from 2014 on (2013: biographical questionnaire)

• Run by FORS (Swiss Centre of Expertise in the Social Sciences ), c/o University of Lausanne

• Financed by Swiss National Science Foundation
SHP – sample and methods

• Representative of the Swiss residential population
• Each individual surveyed every year (Sept.-Jan.)
• All household members from 14 years on surveyed (proxy questionnaire if child or unable)
• Telephone interviews (central CATI), languages D/F/I
• Metadata: biography, interviewers, call data

Following rules:
• OSM followed if moving, from 2007 on all individuals
• All new household entrants surveyed
SHP – sample size (individuals) and attrition
SHP: Survey process and questionnaires

Grid Questionnaire: Inventory and characteristics of hh-members

| Persons 18+ years « reference person» | Persons 14+ years | Persons 13- years + « unable to respond» |

Household Questionnaire: housing, finances, family roles, …

Individual Questionnaires: work, income, health, politics, leisure, satisfaction of life …

Individual Proxy Questionnaires: school, work, income, health, …
SHP: Questionnaire Content

- **Social structure**: socio-demography, socio-economy, work, education, social origin, income, housing, religion
- **Life events**: marriages, births, deaths, deceases, accidents, conflicts with close persons, etc.
- **Politics**: attitudes, participation, party preference
- **Social participation**: culture, social network, leisure
- **Perception and values**: trust, confidence, gender
- **Satisfaction**: different satisfaction issues
- **Health**: physical and mental health self-evaluation, chronic problems
- **Psychological scales**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Social network</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religion</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social participation</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Politics</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leisure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological Scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
International Context

SHP is part of the **Cross National Equivalent File (CNEF)** = General population panel survey with data from:

- **USA** (PSID, data since 1980)
- **D** (SOEP, data since 1984)
- **UK** (BHPS, data since 1991, from 2009 Understanding Society)
- **Canada** (SLID, data since 1993)
- **CH** (SHP, data since 1999)
- **Australia** (HILDA, data since 2001)
- **Korea** (KLIPS, data since 1998)
- **Russia** (RLMS-HSE, data since 1995)

More countries will join (South Africa, Israel, Morocco …)

- **Subset** of variables (variables from original files can be added)
- Variables **ex-post harmonized**, names, categories

SHP – structure of the data

• 2 yearly files (currently available: 1999-2014 (+beta 2015))
  – household
  – Individual

• 5 unique files
  – master person (mp)
  – master household (mh)
  – social origin (so)
  – last job (lj)
  – activity (employment) calendar (ca)

• Complementary files
  – Interviewer data (2000, and yearly since 2003)
  – Call data (since 2005)
  – CNEF SHP data variables
  – Imputed income variables
Documentation (Website: D/E/F)

forscenter.ch/en/our-surveys/swiss-household-panel/ then link Documentation/FAQ:

- Questionnaires PDF
- User Guide PDF
- Variable by Domain (variable search by topic)
- List of Variables (if variable name is known)

...
SHP – data delivery

- **Data** ready about **1 year after end of fieldwork** – **downloadable** from SHP-server:
  
  forscenter.ch/en/our-surveys/swiss-household-panel/datasupport-2
  /telecharger-les-donnees/

  **Signed contract** with FORS

- Upon contract receipt, **login and password** sent by e-mail
- **Data free of charge**
- Users become **member of SHP scientific network and document all publications** based on SHP data
- **Data upon request:**
  - Imputed income
  - Call data
  - Interviewer matching ID
  - Context data (special contract); data is matched at FORS
3

Stata and panel data
Why Stata?

Capabilities

- Data management
- Broad range of statistics
  - Powerful for panel data!
  - Many commands ready for analysis
  - User-written extensions

Beginners and experienced users

- Beginners: analysis through menus (point and click)
- Advanced users: good programmable capacities
Starting with Stata

**Basics**
- Look at the data, check variables
- Descriptive statistics
- Regression analysis

  → **Handout Stata basics**

**Working with panel data**
- Merge
- Creating « long files »
- Working with the long file
- Add information from other household members

  → **Handout Stata SHP data management**
  (includes Syntax examples, exercises)
### 1. Merge: `_merge` variable

**Master file**

<table>
<thead>
<tr>
<th>idpers</th>
<th>p07c44</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

**User file**

<table>
<thead>
<tr>
<th>idpers</th>
<th>p08c44</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

**Master file**

<table>
<thead>
<tr>
<th>idpers</th>
<th>p07c44</th>
<th>p08c44</th>
<th><code>_merge</code></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

- **Merge variable**
  - 1 only in master file
  - 2 only in user file
  - 3 in both files
### Merge: identifier

#### Master file

<table>
<thead>
<tr>
<th>idpers</th>
<th>p07c44</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

#### Using file

<table>
<thead>
<tr>
<th>idpers</th>
<th>p08c44</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

#### Combined table

<table>
<thead>
<tr>
<th>idpers</th>
<th>p07c44</th>
<th>p08c44</th>
<th>_merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Merge files: identifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>filename</td>
<td>identifiers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual master file</td>
<td>shp_mp</td>
<td>idpers, idhous$$, idfath__, idmoth__</td>
<td></td>
</tr>
<tr>
<td>Individual annual files</td>
<td>shp$_$p_user</td>
<td>idpers, idint, idhous$$, idspou__, refper$$</td>
<td></td>
</tr>
<tr>
<td>Additional ind. files (Social origin, last job, calendar, biographic)</td>
<td>shp_so, shp_lj, shp_ca, shp0_*</td>
<td>idpers</td>
<td></td>
</tr>
<tr>
<td>Interviewer data</td>
<td>shp$_$v_user</td>
<td>idint</td>
<td></td>
</tr>
<tr>
<td>Household annual files</td>
<td>shp$_$h_user</td>
<td>idhous$$, refpers, idint, canton$$, (gdenr)</td>
<td></td>
</tr>
<tr>
<td>Biographic files</td>
<td></td>
<td>idpers</td>
<td></td>
</tr>
<tr>
<td>CNEF files</td>
<td>shpequiv_$$$$</td>
<td>x11101ll (=idpers)</td>
<td></td>
</tr>
</tbody>
</table>
The merge command

- Stata merge command

```
merge [type] [varlist] using filename [filename ...] [, options]
```

<table>
<thead>
<tr>
<th>varlist</th>
<th>identifier(s), e.g. idpers</th>
</tr>
</thead>
<tbody>
<tr>
<td>filename</td>
<td>data set to be merged</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>type</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1</td>
<td>each observation has a unique identifier in both data sets</td>
</tr>
<tr>
<td>1:m, m:1</td>
<td>in one data set several observations have the same identifier</td>
</tr>
</tbody>
</table>
### 1:1 merge individual files

#### 2 annual individual files

```stata
use shp08_p_user, clear
merge 1:1 idpers using shp00_p_user
```

<table>
<thead>
<tr>
<th>_merge</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,845</td>
<td>34.93</td>
<td>34.93</td>
</tr>
<tr>
<td>2</td>
<td>5,056</td>
<td>30.21</td>
<td>65.14</td>
</tr>
<tr>
<td>3</td>
<td>5,833</td>
<td>34.86</td>
<td>100.00</td>
</tr>
</tbody>
</table>

**Total** 16,734 100.00
annual individual file and individual master file

use `shp08_p_user`, clear // opens the file (master)
count // there are 10'889 cases
merge 1:1 `idpers` using `shp_mp` // identify & using file

```
<table>
<thead>
<tr>
<th>_merge</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>11,111</td>
<td>50.50</td>
<td>50.50</td>
</tr>
<tr>
<td>3</td>
<td>10,889</td>
<td>49.50</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>22,000</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```

drop if _merge==2 // if only ind. from 2008 wanted
drop _merge
m:1 merge

annual individual file and annual household file

use `shp08_p_user`, clear //master file
merge `m:1 idhous08` using `shp08_h_user` /*identifier & using file */

<table>
<thead>
<tr>
<th>_merge</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10,889</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>10,889</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>
More on merge

• Options of merge command
  – **keepusing** (varlist): selection of variables from using file
  – **keep**: selection of observations from master and/or using file
  – for more options: type  **help merge**

• Merge many files
  – loops (see handout)

• Create partner files (see handout)
Wide and long format

**Wide format**

<table>
<thead>
<tr>
<th>idpers</th>
<th>i04empyn</th>
<th>i05empyn</th>
<th>i06empyn</th>
<th>i07empyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>4101</td>
<td>103190</td>
<td>107730</td>
<td>113400</td>
<td>122470</td>
</tr>
<tr>
<td>42101</td>
<td>63180</td>
<td>69500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56102</td>
<td>35473</td>
<td></td>
<td>41400</td>
<td>45500</td>
</tr>
</tbody>
</table>

**Long format** (person-period-file)

<table>
<thead>
<tr>
<th>idpers</th>
<th>year</th>
<th>iempyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>4101</td>
<td>2004</td>
<td>103190</td>
</tr>
<tr>
<td>4101</td>
<td>2005</td>
<td>107730</td>
</tr>
<tr>
<td>4101</td>
<td>2006</td>
<td>113400</td>
</tr>
<tr>
<td>4101</td>
<td>2007</td>
<td>122470</td>
</tr>
<tr>
<td>42101</td>
<td>2004</td>
<td>63180</td>
</tr>
<tr>
<td>42101</td>
<td>2005</td>
<td>69500</td>
</tr>
<tr>
<td>56102</td>
<td>2004</td>
<td>35473</td>
</tr>
<tr>
<td>56102</td>
<td>2006</td>
<td>41400</td>
</tr>
<tr>
<td>56102</td>
<td>2007</td>
<td>45500</td>
</tr>
</tbody>
</table>
Use of long data format

• All panel applications: xt commands
  – descriptives
  – panel data models
    • fixed effects models, random effects, multilevel
    • discrete time event-history analysis

• declare panel structure
  panel identifier, time identifier
  xtset idpers wave
Convert wide form to long form

**reshape long command in stata**

```
reshape long varlist, i(idpers) j(wave)
```

But: stata does not automatically detect years in varname

```
reshape long i@wyn p@w32 age@ status@, ///
i(idpers) ///
j(wave "99" "00" "01" "02" "03" "04" ///
"05" "06" "07" "08" "09" "10"), atwl()
```
Create a long file with append

1. Modify dataset for each wave

<table>
<thead>
<tr>
<th>idpers</th>
<th>i99wyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50’000</td>
</tr>
<tr>
<td>2</td>
<td>24’800</td>
</tr>
<tr>
<td>3</td>
<td>108’000</td>
</tr>
</tbody>
</table>

2. Stack data sets: use temp1, clear forval y = 2/10 { append using temp`y' }

| idpers |  | iwyn |
|--------|--------|
| 1      | 1      | 50’000|
| 2      | 1      | 24’800|
| 3      | 1      | 108’000|

<table>
<thead>
<tr>
<th>idpers</th>
<th>i00wyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51’000</td>
</tr>
<tr>
<td>2</td>
<td>25’800</td>
</tr>
<tr>
<td>3</td>
<td>109’000</td>
</tr>
</tbody>
</table>

| idpers |  | iwyn |
|--------|--------|
| 1      | 2      | 51’000|
| 2      | 2      | 25’800|
| 3      | 2      | 109’000|

<table>
<thead>
<tr>
<th>idpers</th>
<th>i10wyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52’000</td>
</tr>
<tr>
<td>2</td>
<td>26’800</td>
</tr>
<tr>
<td>7</td>
<td>11’000</td>
</tr>
</tbody>
</table>

| idpers |  | iwyn |
|--------|--------|
| 1      | 10     | 58’000|
| 2      | 10     | 26’800|
| 7      | 10     | 11’0005|
Work with time lags

- If data in long format and defined as panel data (xtset)
- l. indicates time lag
- f. indicates time lead

- Example:
  social class of last job (see handout)
  life events
Missing data in the SHP

Missing data in the SHP: negative values
-1 does not know
-2 no answer
-3 inapplicable (question has not been asked)
-8/-4 other missings

Missing data in Stata: . .a .b .c .d etc.
- negative values are treated as real values
- missing data (. .a .b etc.) are defined as the highest possible values; . < .a < .b < .c < .d

→ recode to missing or analyses only positive values
  e.g. `sum i08empyn if i08empyn>=0`
→ care with operator >
  e.g. `count if i08empyn>100000` counts also missing values
→ write <. instead of !=.
Longitudinal data analysis with Stata

**xt commands:**

**descriptive statistics**
- `xtdescribe`
- `xtsum, xttab, xttrans`

**regression analysis**
- Random Intercept: `xtreg, xtgl, xtlogit, xtpoisson, xtcloglog`
- Random Slope: `mixed, melogit, ...`

**diagrams:** `xtline`
Descriptive analysis

• Get to know the data
• Usually: similar findings to complicated models
• Visualisation
• Accessible results to a wider public
• Assumptions more explicit than in complicated models
Example: variability of party preferences

Figure 1: Patterns of Change over Nine Yearly Observations: Frequencies (N = 1’994)

Kuhn (2009), Swiss Political Science Review 15(3): 463-494
Example: becoming unemployed

Satisfaction with life

Germany, 1984-2010

Switzerland, 2000-2010

### Example: Income mobility

<table>
<thead>
<tr>
<th></th>
<th>Switzerland</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low income 2009</td>
<td>Middle income 2009</td>
<td>High income 2009</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Low income 2005</td>
<td>56.2 %</td>
<td>40.8 %</td>
<td>3.1 %</td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>Middle income 2005</td>
<td>13.5 %</td>
<td>75.8 %</td>
<td>10.8 %</td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>High income 2005</td>
<td>4.4 %</td>
<td>34.4 %</td>
<td>61.1 %</td>
<td>100 %</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Germany</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low income 2009</td>
<td>Middle income 2009</td>
<td>High income 2009</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Low income 2005</td>
<td>61.7 %</td>
<td>36.4 %</td>
<td>1.9 %</td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>Middle income 2005</td>
<td>12.4 %</td>
<td>78.4 %</td>
<td>9.2 %</td>
<td>100 %</td>
<td></td>
</tr>
<tr>
<td>High income 2005</td>
<td>2.6 %</td>
<td>29.6 %</td>
<td>67.8 %</td>
<td>100 %</td>
<td></td>
</tr>
</tbody>
</table>

4
Linear regression
(Refresher course)
Aim and content

Refresher course on linear regression

- What is a regression?
- How to obtain regression coefficients?
- How to interpret regression coefficients?
- Inference from sample to population of interest (significance tests)
- Assumptions of linear regression
- Consequences when assumptions are violated
What is a regression?

A statistical method for studying the relationship between a single dependent variable and one or more independent variables.

- Y: dependent variable
- X: independent variable(s)

Simplest form: bivariate linear regression
- linear relationship between a dependent and one independent variable for a given set of observations

Examples
- Does the wage level affect the number of hours worked?
- Gender discrimination in wages?
- Do children increase happiness?
We start with a “scatter plot” of observations
Regression line: $\hat{y}_i = a + bx_i = 51375 + 693 \times x_i$
Regression line: $\hat{y}_i = a + b^*x_i = 51375 + 693^*x_i$

Estimated regression equation: $y_i = a + b^*x_i + e_i$
Components of (linear) regression equation

Estimated regression equation:

univariate: \[ y = a + bx + e \]
multivariate: \[ y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + e \]

- \( y \) dependent variable
- \( x \) independent variable(s) (predictor(s), regressor(s))
- \( a \) intercept (predicted value of \( Y \) if \( x=0 \))
- \( b \) regression coefficients (slope): measure of the effect of \( X \) on \( Y \)
  - multivariate regression: the portion of \( y \) explained by \( x \) that is not explained by the others \( x \)'s
- \( e \) part of \( y \) not explained by \( x \) (residual), due to
  - omitted variables, measurement errors, stochastic shock, disturbance

We assume a **linear** relationship between the conditional expectation value of \( Y \) and \( X \)
Scales of independent variables

• Continuous variables: linear

• Binary variables (Dummy variables) (0, 1)
  Example: female=1, male=0

• Ordinal or multivariate variables (n categories)
  Create n-1 dummy variables (base category)

Example: educational levels
  – 1 low educational level
    2 intermediate educational level
    3 high educational level
  – Include 2 dummy variables in regression model
Regression – graphical interpretation
### Example: gender wage gap

Sample: full-time employed, yearly salary between 20’000 and 200’000 CHF

<table>
<thead>
<tr>
<th></th>
<th>Bivariate</th>
<th>Multivariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>98790</td>
<td>45'369</td>
</tr>
<tr>
<td>Female</td>
<td>-17'737</td>
<td>-9'090</td>
</tr>
<tr>
<td>Education (Ref.: compulsory)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary</td>
<td></td>
<td>9'197</td>
</tr>
<tr>
<td>Tertiary</td>
<td></td>
<td>30'786</td>
</tr>
<tr>
<td>Supervision</td>
<td></td>
<td>17'128</td>
</tr>
<tr>
<td>Financial sector</td>
<td></td>
<td>15'592</td>
</tr>
<tr>
<td>Number of years in paid work</td>
<td></td>
<td>729</td>
</tr>
</tbody>
</table>
Test if $\beta \neq 0$

If $\beta=0$ (in population), there is **no relationship between x and y**

$\rightarrow$ **$H_0$: $\beta = 0$**

- **$H_0$: Distribution if $\beta=0$**
  $\rightarrow$ compare estimate with critical value
  $\rightarrow$ if $\text{abs}(b) > \text{abs(critical value)}$:
    - $b$ significant

Test: $\beta / \text{s.e.}(\beta)$ is t-distributed.

Rule of thumb: if $> 2$, then significant on 5% level.
Regression with Stata: cross-sectional regression

Example: life satisfaction, SHP data 2012

```
reg lifesat partner age agesq edulow eduhigh lnincome
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 6916</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>640.854082</td>
<td>6</td>
<td>106.809014</td>
<td>F(  6, 6909) = 58.64</td>
</tr>
<tr>
<td>Residual</td>
<td>12584.2476</td>
<td>6909</td>
<td>1.82142822</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>13225.1016</td>
<td>6915</td>
<td>1.91252374</td>
<td>R-squared = 0.0485</td>
</tr>
</tbody>
</table>

| lifesat | Coef.     | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|-----------|-----------|-------|------|---------------------|
| partner | .5842025  | .0404291  | 14.45 | 0.000 | .5049491 .6634559 |
| age     | -.0617406 | .0052466  | -11.77| 0.000 | -.0720254 -.0514557|
| agesq   | .0631447  | .0052222  | 12.09 | 0.000 | .0529076 .0733817 |
| edulow  | -.0318721 | .046435   | -0.69 | 0.492 | -.122899 .0591549 |
| eduhigh | .0765272  | .0460582  | 1.66  | 0.097 | -.0137611 .1668156 |
| lnincome| .2845323  | .0343396  | 8.29  | 0.000 | .2172162 .3518485 |
| _cons   | 5.697941  | .3903272  | 14.60 | 0.000 | 4.93278 6.463103 |

Root MSE = 1.3496
Inference: Variation of regression coefficient $b$

\[
V(b) = \sigma^2_\beta = \frac{\sum \sigma^2_\varepsilon}{\sum (x_i - \bar{x})^2}
\]

\[
\sigma^2_\varepsilon = \frac{\sum (\varepsilon_i)^2}{n - p}
\]

Variation of $b$ ($\sigma^2_\beta$): decreases if
- $n$ increases
- $x$ are more spread out
- squared residuals decrease

Distribution of $b$
- Student t-distribution
  - = normal distribution if $n$ large
Example: significance levels and sample size

Sample n=53

| Coef.  | st.e.  | t     | P>|t|   | [95% Conf. Interval] |
|--------|--------|-------|--------|---------------------|
years work.| 692.6  | 289.1 | 2.40   | 0.020   | 112.1    | 1273.0 |
_cons  |51375.9 |7340.4 | 7.00   | 0.000   | 36639.4 | 66112.5 |
R²: 0.101

Sample n=1787

| Coef.  | St.e.  | t     | P>|t|   | [95% Conf. Interval] |
|--------|--------|-------|--------|---------------------|
years work.| 931.4  | 50.6  | 18.40  | 0.000   | 832.1    | 1030.7 |
_cons  |51218.6 |1271.2 | 40.29  | 0.000   | 48725.5 | 53711.7 |
R²: 0.159

Note: Standard error has same scale as coefficient
Assumptions of OLS regression

General
- Continuous dependent variable
- Random sample

Coefficient estimation
- No perfect multicollinearity
- $E(e) = 0$ (artifact)
- **No endogeneity;** $Cov(x,e) = 0$

Inference
- No autocorrelation $Cov(e_i,e_k)=0$
- Constant variance (no heteroscedasticity)
- Preferentially: residuals normally distributed

\[
\text{Coefficients biased (inconsistent)} \quad \Rightarrow \quad \text{Standard errors of coefficients biased}
\]
Inference: assumptions

Assumptions on error terms

- Independence of error terms, no autocorrelation:
  \[ \text{Cov}(\varepsilon_i, \varepsilon_k) = 0 \text{ for all } i,k, i \neq k \]

- Constant error variance: \( \text{Var}(\varepsilon_i) = \sigma^2 \) for all \( i \);
  (Homoscedasticity)

Preferentially: \( e \) is normally distributed

Matrix of error terms

\[
\begin{array}{cccccccc}
& 1 & 2 & 3 & 4 & 5 & \cdots & n-1 & n \\
1 & \sigma^2 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
2 & 0 & \sigma^2 & 0 & 0 & 0 & \cdots & 0 & 0 \\
3 & 0 & 0 & \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\
4 & 0 & 0 & 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
n-1 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma^2 & 0 \\
n & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \sigma^2 \\
\end{array}
\]
Autocorrelation

Reason: Nested observations (e.g. households, schools, time, municipalities)
→ standard errors underestimated

What to do: OLS with adjust standard errors

<table>
<thead>
<tr>
<th>$i;k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
<th>$n-1$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_2^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_1^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i;k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
<th>$n-1$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i;k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
<th>$n-1$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_2^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_1^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i;k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>...</th>
<th>$n-1$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\sigma^2$</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$n-1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\sigma_1^2$</td>
<td>$\sigma_2^2$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$n$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$\sigma_2^2$</td>
<td>$\sigma_1^2$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Endogeneity

- Traditional meaning: Variable is determined within a model
- Here (econometric): Any situation where an explanatory variable is correlated with the residual Cov(x,e) = 0
- Reasons
  - Omitted variables
  - Measurement error
    - in explanatory variables: underestimated effects,
      - in dependent variable: larger variance of error term
  - Simultaneity
  - Nonlinearity in parameters (can be corrected)
- If a variable is endogenous: model cannot be interpreted as causal (bias)
x is correlated with an unobserved (omitted) variable if this omitted variable is correlated with y (conditional on x) -> all x’s are biased

\[ y = a + bx + e \]
\[ = a + bx + (cx + e') \] if x is correlated with an unobserved variable
\[ = a + (b+c)x + e' \] we estimate b+c instead of b (causal effect)

Example:
Causal model  civic engagement  \rightarrow  trust
Omitted variable
values, personality, childhood

civic engagement  \rightarrow  trust
Endogeneity: simultaneity, reverse causality

Causal model

\[ X \rightarrow Y \]

having a partner \quad life satisfaction

Simultaneity

\[ X \leftrightarrow Y \]

having a partner \quad life satisfaction

Consequence: Estimator biased
Detection and correction of endogeneity

• Difficult: caution for causal interpretation!

• Detection
  – Theory, literature (variable selection and interpretation) !!!!
  – Robustness checks

• Correction: instrumental variables, panel data (time ordering, within-models), structural equation modelling, discontinuity design….

• ! Overcontrol is common in social research based on regressions. Do not control for intervening mechanisms (“collider” variables)
Regression with panel data: Data structure

— Cross sectional data

<table>
<thead>
<tr>
<th>idpers</th>
<th>lifestat04</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

— Panel data

<table>
<thead>
<tr>
<th>idpers</th>
<th>year</th>
<th>lifesat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2004</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2005</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2006</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>2007</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>2004</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>2005</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>2004</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2006</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2007</td>
<td>5</td>
</tr>
</tbody>
</table>
OLS with pooled panel data: problems I

• Pooled data: long data format, different years in one file
• Problem: OLS assumption of independent observations violated (autocorrelation)
  → coefficients unbiased
  → but standard errors biased (underestimation)
• Possible measure: Correct for clustering in error terms
• But: OLS is not the best estimator for pooled data (not efficient)
Example: Partner -> Life satisfaction

- SHP 2000-2012
- 80’914 observations from 14’345 individuals

<table>
<thead>
<tr>
<th>Life satisfaction</th>
<th>OLS</th>
<th>OLS (correct for cluster)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partner</td>
<td>0.481***</td>
<td>0.481***</td>
</tr>
<tr>
<td>Age</td>
<td>-0.070***</td>
<td>-0.070***</td>
</tr>
<tr>
<td>Age squared</td>
<td>0.077***</td>
<td>0.077***</td>
</tr>
<tr>
<td>Education: low</td>
<td>-0.016</td>
<td>-0.016</td>
</tr>
<tr>
<td>Education: high</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>Income (ln)</td>
<td>0.230***</td>
<td>0.230***</td>
</tr>
<tr>
<td>Health: so so</td>
<td>1.248***</td>
<td>1.248***</td>
</tr>
<tr>
<td>Health: well</td>
<td>2.025***</td>
<td>2.025***</td>
</tr>
<tr>
<td>Health: very well</td>
<td>2.502***</td>
<td>2.502***</td>
</tr>
<tr>
<td>Constant</td>
<td>4.564</td>
<td>4.564</td>
</tr>
</tbody>
</table>

Stata: `reg lifesat partner age agesq …. lnincome, cluster(idpers)`
OLS with panel data: problems II and outlook

• OLS does not take advantage of panel structure

• Two different types of variation in panel data
  – Variation within individuals
  – Variation between individuals

• Control for unobservable variables (stable personal characteristics)
  – Within-models
  – Random Effect Models
    (multilevel / random intercept / hierarchical model / frailty for event history, mixed model)
5
Causality
Interpretation of model results

• Descriptive interpretation

versus

• Causal interpretation
  Idea: use only variance in treatment variable which is exogenous (exogenously manipulated by researcher)
Causality

**Def.:** Necessary (not sufficient) conditions for X to "cause" Y:
- X *precedes* Y (also anticipation)
- X *correlates* with Y
- *theoretical* explanation of mechanism between X and Y ("law")

Causality in social science *experiments*
- Random group receives "treatment" (manipulation): no omitted variable bias (self-selection into treatment)
- we have treatment and control group

**Problems:**
- Experiments usually not possible in social science (external validity): ethical or organizational problems
- What about effects of unchangeable variables (like sex)?
- Continuous variables?
How are causal effects analyzed?

- **Multiple Regression**: attempt to control for all omitted variables
  - Problems: - omitted variables, unobserved heterogeneity
  - form of relationship must be specified

- **Propensity score matching**: attempt to compare members with same (or similar) scores on control variables
  - Problem: - omitted variables, unobserved heterogeneity
  - Advantage: - non-parametric

- **Instrumental variables**: use only variance of x that correlates with exogenous instrument z

- **Panel data**: before and after measurement
  - Problems: - little before/after variation (within individuals)
  - co-varying change variables (corr. with ε_{it}) must be controlled
  - Advantage: - co-varying change time-invariant variables (corr. with u_{it}) no longer a problem
Measuring causal effects (example: binary treat)

- **unbiased** Effects = $Y_{i,t_1}^{Treat} - Y_{i,t_1}^{NonTreat}$ counterfactual!

- Cross-sectional data: $Y_{i,t_1}^{Treat} - Y_{j,t_1}^{Treat}$ but: different persons i,j

- With **Panel data I**: within-estimation $Y_{i,t_1}^{Treat} - Y_{i,t_2}^{Treat}$

Causal effect but problem with time-variant effects (e.g., time, unmeasured within-changes with effects on Y)
Basic Approach

Between-estimation

- ok with experimental data
  - Due to randomization units differ only in the treatment
- But strong assumption of unit homogeneity causes bias
  - Problem: self-selection into treatment!
  - Unobserved unit heterogeneity

Within-estimation

- with control group often ok because the parallel trends assumption is much weaker
  - Unobserved unit heterogeneity will not bias within-estimation
  - Only differing time-trends in treatment and control group will bias within-estimation results
Hypothetical Data: does having a partner make happier?

```
list id time satlife partner,
separator(6)

+-------------------------------+
<table>
<thead>
<tr>
<th>id   time   satlife   partner</th>
</tr>
</thead>
</table>
1. |  1      1         2         0 |
2. |  1      2       2.1         0 |
3. |  1      3       1.9         0 |
4. |  1      4         2         0 |
5. |  1      5       2.2         0 |
6. |  1      6       1.8         0 |
7. |  2      1       4.2         0 |
8. |  2      2       3.9         0 |
9. |  2      3       4.1         0 |
10. |  2      4       4         0 |
11. |  2      5       3.9         0 |
12. |  2      6       4.1         0 |

+-------------------------------+  +-------------------------------+
<table>
<thead>
<tr>
<th>id   time   satlife   partner</th>
</tr>
</thead>
</table>
13. |  3      1       5.8         0 |
14. |  3      2       6         0 |
15. |  3      3       6.2         0 |
16. |  3      4       7         1 |
17. |  3      5       6.9         1 |
18. |  3      6       7.1         1 |
19. |  4      1       7.9         0 |
20. |  4      2       8.1         0 |
21. |  4      3       8         0 |
22. |  4      4       9         1 |
23. |  4      5       9.2         1 |
24. |  4      6       8.8         1 |
```
Those with and those without partner differ in characteristics other than partnership: no unit homogeneity.
Self-Selection: Treatment not under control

- Between-approach biases results
- Within-approach possible: we have before \((t=1,2,3)\) and after \((t=4,5,6)\) measurements
- Therefore unit heterogeneity no longer a problem
- We have in addition a control group: DiD

\[
\begin{align*}
\text{after-before (treat)} &= \text{sat}(t=4,5,6)-\text{sat}(t=1,2,3) \mid_{\text{treat}} \\
&= ((7-6)+(9-8))/2=1 \\
\text{after-before (control)} &= \text{sat}(t=4,5,6)-\text{sat}(t=1,2,3) \mid_{\text{control}} \\
&= ((2-2)+(4-4))/2=0
\end{align*}
\]

Average treatment effect: ATE: difference of averages of treatment and control group: ATE = 1
Can regression produce these results?

Cross-sectional regression at $t=4$

$$\beta_{t=4} = \frac{(9+7)}{2} - \frac{(4+2)}{2} = 8 - 3 = 5$$

massively biased! Compares average happiness between partnered and unpartnered people
Most critical assumption of a linear regression is the **exogeneity assumption**: \( \text{Cov}(x,e) = 0 \)

But: unobserved confounders (unobservables that affect both \( X \) and \( Y \))

\[ e_i \text{ (personality, attractiveness, intelligence, ...)} \]

\[ x_i \quad \text{Partner} \]

\[ y_i \quad \text{Happy} \]

\( \text{Cov}(x,e) \neq 0 \) (unobserved heterogeneity or **omitted variable bias**)

The happier self-select into partnership

Treatment and control group are not (initially) randomized
Pooled OLS is no solution

Bias still high: $\beta_{\text{pooled}} = 3.67$

Pooled OLS still relies on between-comparison

Pooled OLS: mean red points - mean green points
Towards panel models: error decomposition

Starting point: error decomposition \( e_{it} = \alpha_i + \varepsilon_{it} \)

\( \alpha_i \) person-specific time-constant error term («between»)
Assumption: person-specific random variable

\( \varepsilon_{it} \) time-varying error term (idiosyncratic error term) («within»)
Assumption: zero mean, homoscedasticity, no autocorrelation

\[ \begin{align*}
\alpha_i & \\
\varepsilon_{it} & \\
x_{it} & \\
y_{it} & \\
\end{align*} \]

Partner
Satisfaction

Personality, attractiveness, intelligence, ...
Sunshine, ...

5_13
Total Variance is equal to the Square of the Differences of all Observations from the Total Mean divided by the Sample Size (4) =

\[ \frac{(3-0)^2 + (2-0)^2 + (-1-0)^2 + (-4-0)^2}{4} = \frac{9 + 4 + 1 + 16}{4} = 7.5 \]
Illustration: between-variance

\[ \sum_{i=1}^{n} \sum_{t=1}^{T_i} (y_{it} - \bar{y})^2 = \sum_{i=1}^{n} \sum_{t=1}^{T_i} (y_{it} - \bar{y}_i)^2 + \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\bar{y}_i - \bar{y})^2 \]

\[ T_{yy} = W_{yy} + B_{yy} \]

variance of the individual means = “between”-variance \((\sigma_{\text{between}} = \sigma_{\alpha})\)

\[ (2.5^2 + (-2.5)^2)/2 = 6.25 \]

(remember: total variance = 7.5)
Illustration: within-variance

\[
\sum_{i=1}^{n} \sum_{t=1}^{T_i} (y_{it} - \bar{y})^2 = \sum_{i=1}^{n} \sum_{t=1}^{T_i} y_{it} - \bar{y}_i)^2 + \sum_{i=1}^{n} \sum_{t=1}^{T_i} (\bar{y}_i - \bar{y})^2
\]

\[
T_{yy} = W_{yy} + B_{yy}
\]

variance of measurements within individuals = “within”-Variance:

\[
((3-2.5)^2 + (2-2.5)^2 + (-1-(-2.5))^2 + (-4-(-2.5))^2)/4
\]

=1.25 (=18% of total variance)

variance of individual means = \( (2.5^2+(-2.5)^2)/2 = 6.25 \)

(=82% of total variance = \( \rho =ICC \) (intra-class-correlation))
Model: \( y_{it} = \beta x_{it} + \alpha_i + \varepsilon_{it} \)

POLLS is consistent only, if the regressor \( x_{it} \) is independent from both error components:

\[ E(\alpha_i \mid x_{it}) = 0 \] no person-specific time-constant unobserved heterogeneity («random effects» assumption)

\[ E(\varepsilon_{it} \mid x_{it}) = 0 \] no time-varying unobserved heterogeneity («strict exogeneity» assumption)

Error components model

Personality, attractiveness, intelligence, … Sunshine, …
6
Fixed Effects ("within") Models
Bias from omitted time-invariant variables

- Many *time-invariant* individual characteristics ($\alpha_i$) are **not** observed or not taken into account
  - e.g. enthusiasm, ability, willingness to take risks, attractiveness
- These may have an effect on **dependent variable**, and are **correlated** with **independent variables** like satlife – partner - attractiveness

Then regression coefficients will be **biased**!
Hypothetical Example:
Omitted time-invariant Variable Bias
BMI (Y) and Smoking (X) :
Continuous “Treatment”
Hypothetical Observations: BMI and number of cigarettes per day
Scatter plot with (naive) linear fit line from pooled OLS

BMI und Anzahl Zigaretten pro Tag

bmi

0 10 20 30

1 21 22 23 24 25 26 27 28 29 30 31 32

0 10 20 30 40

cigarettes
. * pooled regression:
. reg bmi cigarettes

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>25.4934645</td>
<td>1</td>
<td>25.4934645</td>
<td>F(1, 43) = 5.26</td>
</tr>
<tr>
<td>Residual</td>
<td>208.506536</td>
<td>43</td>
<td>4.8489892</td>
<td>Prob &gt; F = 0.0268</td>
</tr>
<tr>
<td>Total</td>
<td>234</td>
<td>44</td>
<td>5.31818182</td>
<td>R-squared = 0.1089</td>
</tr>
</tbody>
</table>

| bmi      | Coef.    | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|----------|----------|-----------|-------|-----|----------------------|
| cigarettes | .0966882 | .0421682 | 2.29  | 0.027 | .011648 .1817285 |
| _cons    | 25.06379 | .7722906 | 32.45 | 0.000 | 23.50632 26.62126 |

BMI higher by .1 if number of cigarettes is 1 higher
Between-individual effects (means of individuals)

BMI and number of cigarettes per day

BMI

mcigarettes
Between regression

* between-regression:
  * egen mbmi=mean(bmi), by(id)
  * egen mcigarettes=mean(cigarettes), by(id)
  * bysort id: gen n=_n
  * reg mbmi mcigarettes if n==1 // between-regression

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>22.134779</td>
<td>1</td>
<td>22.134779</td>
</tr>
<tr>
<td>Residual</td>
<td>41.198548</td>
<td>13</td>
<td>3.16911908</td>
</tr>
<tr>
<td>Total</td>
<td>63.333327</td>
<td>14</td>
<td>4.52380907</td>
</tr>
</tbody>
</table>

Number of obs = 15
F( 1, 13) = 6.98
Prob > F = 0.0203
R-squared = 0.3495
Adj R-squared = 0.2995
Root MSE = 1.7802

| mbmi  | Coef.  | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-------|--------|-----------|------|-----|----------------------|
| mcigarettes | .1709983 | .0647029 | 2.64 | 0.020 | .0312163 .3107803 |
| _cons  | 23.8319 | 1.166966 | 20.42 | 0.000 | 21.31082 26.35297 |

BMI higher by .17 if number of cigarettes is 1 higher
Omitted variable: Social class!

BMI und Anzahl Zigaretten pro Tag

Lines have negative slopes within social class.
Panel Data is even better:
We can take a look within individuals
Panel Data: Observations are clustered in Individuals!

BMI und Anzahl Zigaretten pro Tag

Lines have negative slopes within individuals
Individually de-meaned values: OLS curves through Origin (0,0)

BMI und Anzahl Zigaretten pro Tag: entmittelte Werte

Within (FE) regression line has negative slope
How can we calculate a within-regression coefficient?
Error components in panel data models

• We separate the error components:
  \[ e_{it} = \alpha_i + \varepsilon_{it} \]
  \( \alpha_i \) = person-specific unobserved heterogeneity (level) = „fixed effects“ (e.g., social origin, ability)
  \( \varepsilon_{it} \) = „residual“ (e.g., sunshine)

Model:

\[ bmi_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it} \]

• Remember: Pooled OLS assumes that \( x \) is not correlated with both error components \( \alpha_i \) and \( \varepsilon_{it} \) (omitted variable bias)
Fixed effects regression

• We can eliminate the fixed effects $\alpha_i$ by estimating them as *person specific* dummies

• -> remains **only within-variation**

• Corresponds to “de-meaning” for each individual:

$$bmi_{it} = \beta_1 x_{it} + \alpha_i + \varepsilon_{it} \quad (1)$$

individual mean:

$$\bar{bmi}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{\varepsilon}_i \quad (2)$$

• subtract

$$bmi_{it} - \bar{bmi}_i = \beta_1 (x_{it} - \bar{x}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

-> **Fixed (all time invariant) effects $\alpha_i$ disappear**, i.e. person-constant unobserved heterogeneity is eliminated
Because $\alpha_i$ is not in the differenced equation, $E(\alpha_i \mid x_{it}) = 0$ is no longer required for consistency.

De-meaning identifies the causal effect under weaker assumptions.
OLS of individually de-meaned Data

We de-mean and regress the Data:

\[
\text{egen mbmi}=\text{mean}(\text{bmi}), \text{ by(id)} \\
\text{egen mcigarettes}=\text{mean}(\text{cigarettes}), \text{ by(id)} \\
\text{gen wbmi}=\text{bmi}-\text{mbmi} \\
\text{gen wcigarettes}=\text{cigarettes}-\text{mcigarettes} \\
. \text{reg wbmi wcigarettes}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>34.0828</td>
<td>1</td>
<td>34.0828</td>
<td>F( 1, 43) = 147.78</td>
</tr>
<tr>
<td>Residual</td>
<td>9.917154</td>
<td>43</td>
<td>.230631488</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>44</td>
<td>1</td>
<td>Adj R-squared = 0.7694</td>
</tr>
</tbody>
</table>

| wbmi | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|---|------|----------------------|
| wcigarettes | -.2733918 | .0224893 | -12.16 | 0.000 | -.3187459 to -.2280377 |
| _cons | -2.86e-07 | .0715901 | -0.00 | 1.000 | -.1443755 to .1443749 |

**BMI decreases by .27 with each additional cigarette**

wrong standard error
Direct modeling of fixed Effects in Stata

.xtreg bmi time, fe (calculates correct df; this causes higher Std. Err.)

Fixed-effects (within) regression
Group variable: id

R-sq: within = 0.7746
  between = 0.3495
  overall = 0.1089

F(1,29) = 99.67
Prob > F = 0.0000

corr(u_i, Xb) = -0.8080

|         | Coef.    | Std. Err. |     t   | P>|t| | [95% Conf. Interval] |
|---------|----------|-----------|---------|------|---------------------|
| bmi     |          |           |         |      |                     |
| cigarettes | -0.2733918 | 0.027385  | -9.98   | 0.000 | -0.3294003  -0.2173833 |
| _cons   | 31.1989  | 0.4622757 | 67.49   | 0.000 | 30.25344   32.14436  |

| sigma_u | 3.6906387 |
| sigma_e | 0.58478272 |
| rho     | 0.97550841 | (fraction of variance due to u_i) |

F test that all u_i=0: F(14, 29) = 41.48
Prob > F = 0.0000
Alternative: OLS with individual dummies controlled

```
.xi i.id, noomit
.reg bmi cigarettes _I*, noconst
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>32224.0828</td>
<td>16</td>
<td>2014.00518</td>
<td>F( 16, 29) = 5889.41</td>
</tr>
<tr>
<td>Residual</td>
<td>9.917154</td>
<td>29</td>
<td>.341970827</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>32234</td>
<td>45</td>
<td>716.311111</td>
<td>R-squared = 0.9997</td>
</tr>
</tbody>
</table>

| bmi | Coef.     | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|-----|-----------|-----------|-------|-------|---------------------|
| cigarettes | -.2733918 | .027385   | -9.98 | 0.000 | -.3294003 to -.2173833 |
| _Iid_1 | 23.70029  | .3643332  | 65.05 | 0.000 | 22.95515 to 24.44544  |
| _Iid_2 | 29.06725  | .4347228  | 66.86 | 0.000 | 28.17814 to 29.95636  |
| _Iid_3 | 29.43421  | .5317197  | 55.36 | 0.000 | 28.34672 to 30.5217   |
| _Iid_4 | 31.80117  | .6434009  | 49.43 | 0.000 | 30.48527 to 33.11707  |
| _Iid_5 | 34.16813  | .7633481  | 44.76 | 0.000 | 32.60691 to 35.72935  |
| _Iid_6 | 37.53509  | .8882188  | 42.26 | 0.000 | 35.71848 to 39.3517   |

useful for small N, the $u_i$ are estimated (only approximate)
Summary: Fixed Effects Estimation

- Solves problem of **time-invariant** unobserved heterogeneity
- Causal interpretation of coefficients

But:

- If number of groups large, **many extra parameters**
- Enough **within-variance needed** in data
- Estimation of **person-constant covariates** (like sex) not possible, dropped from the model. But: possibility to use **interactions** with time-changing variables (like sex*nrchildren: include main effect nrchildren)
- Measurement errors (change!) may cause problems
- Assumption that most important omitted variables are time-invariant is quite strong
Fixed Effects Model example from literature:
Does civic engagement increase generalised trust?


- Data: Swiss Household Panel 2004-2008
- Variables:
  - Y: Belief that most people can be trusted (scale 0 – 10)
  - X: Number of memberships in voluntary associations (0 - 9)
  - Control: Education, health, employment, having a partner
- Cross-sectional interpretation: compare trust of members/non-members with more or less membership
- Longitudinal interpretation: does trust change once individuals join or quit organisations?
Civic engagement and trust example: between and within regression

<table>
<thead>
<tr>
<th></th>
<th>Between</th>
<th>Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membership count</td>
<td><strong>0.413</strong></td>
<td>0.037</td>
</tr>
<tr>
<td>Education</td>
<td><strong>0.089</strong></td>
<td>0.008</td>
</tr>
<tr>
<td>Partner</td>
<td><strong>-0.262</strong></td>
<td>0.015</td>
</tr>
<tr>
<td>Health: not well (ref)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>so, so / average</td>
<td>0.293</td>
<td>0.023</td>
</tr>
<tr>
<td>well</td>
<td><strong>1.109</strong></td>
<td>0.064</td>
</tr>
<tr>
<td>very well</td>
<td><strong>1.278</strong></td>
<td>0.069</td>
</tr>
<tr>
<td>Employed</td>
<td>0.005</td>
<td><strong>-0.206</strong></td>
</tr>
<tr>
<td>Intercept</td>
<td><strong>3.676</strong></td>
<td><strong>5.263</strong></td>
</tr>
</tbody>
</table>

n = 13’534 observations, 4’436 individuals; Controlled for year of measurement
Source: Van Ingen and Bekkers (2013)
Civic engagement and trust: assumptions / limitations of FE model

• All transitions are assumed to have the **same effect** (general assumption of linear regression)

  – Effects of joining and quitting an organisation are symmetric
  – Effect of maintaining 4 memberships (4-4) and staying uninvolved (0-0) are equal

In addition:

• Other life events may impact both membership and trust (third variable)
Civic engagement and generalised trust:  
First-difference model excursus

- Distinction between 4 groups to test for asymmetry:
  - Non-participants: remain uninvolved (Reference groups)
  - Join organisation (entry): from 0 to at least 1 membership
  - Quit organisation (exit): from at least 1 membership to 0
  - Participants: remain involved

- Compares change in trust of four groups at two time points
  \[ y_t - y_{t-1} = a + b_1 \ast entry + b_2 \ast exit + b_3 \ast stay\ involved + \cdots + e \]

- Advantage: theoretically less restrictive
- Disadvantage: captures only short-term effects
  
  More observations lost due to gaps
Graphical interpretation of First-difference model
Remember: Graphical interpretation of FE model
Civic engagement and generalised trust: (FD) results

<table>
<thead>
<tr>
<th></th>
<th>Between (n=13'534)</th>
<th>FE (n=13’534)</th>
<th>First difference (n=8327)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Membership count</td>
<td>0.413**</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>Remain uninvolved (n=1485) (ref)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entry / start (n=810)</td>
<td></td>
<td></td>
<td>0.178*</td>
</tr>
<tr>
<td>Exit / quit (n=791)</td>
<td></td>
<td></td>
<td>0.038</td>
</tr>
<tr>
<td>Remain involved (n=5333)</td>
<td></td>
<td>0.109*</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>0.089**</td>
<td>0.008</td>
<td>0.007</td>
</tr>
<tr>
<td>Partner</td>
<td>-0.262**</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>Health: not well (reference)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>so, so / average</td>
<td>0.293</td>
<td>0.023</td>
<td>0.310*</td>
</tr>
<tr>
<td>well</td>
<td>1.109**</td>
<td>0.064</td>
<td>0.173</td>
</tr>
<tr>
<td>very well</td>
<td>1.278**</td>
<td>0.069</td>
<td>0.138</td>
</tr>
<tr>
<td>Employed</td>
<td>0.005</td>
<td>-0.206**</td>
<td>0.133**</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.676</td>
<td>5.263**</td>
<td>0.083**</td>
</tr>
</tbody>
</table>
What if we have a treatment which does not account for general developments (e.g., trend)?

With Panel data II: “difference-in-difference” (DID):

\[
(Y_{i,t_2}^{Treat} - Y_{i,t_1}^{Treat}) - (Y_{j,t_2}^{NonTreat} - Y_{j,t_1}^{NonTreat})
\]

In reality, treatment effect is positive!

Assumption: parallel trends
DiD

Comparison of groups at different time points (a version of FE-model)

i.e., we calculate treatment effect and control for time

‘DID’ – estimator in case of a simple treatment:

\[
(\text{after}_{\text{treat}} - \text{before}_{\text{treat}}) - (\text{after}_{\text{control}} - \text{before}_{\text{control}})
\]

FE/within \hspace{2cm} trend

We can also include time dummies or (linear) trend
Fixed Effects Individual Slopes (FEIS) models

\[ y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{1i} + \alpha_{2i} t + \varepsilon_{it} \]

Individual level and linear slope controlled: only difference around individual trend.

-> weaker assumption than standard FE: part of \( \varepsilon_{it} \) which is due to individual trend (\( \alpha_{2i} t \)) needs not be independent of \( x_{it} \)

-> In FEIS model time-varying unobserved heterogeneity that is due to individual-specific trends is no longer a problem

Use ado xtf.feis.ado in Stata
7
Random Effects Models
Motivation: multilevel (RE) models

If data have different levels with
- observations are not independent of levels and
- There true social interactions

Examples:
**Schools – classes – students**: first applications
**Networks**: people are influenced by their peers
**Spatial context**: from environment (e.g., poor people are less happy if they live in a rich environment) – US: “neighborhood-effects”
**Interviewer - effects**: respondents clustered in interviewers and:
**Panel-surveys**: waves clustered in respondents (households)
Within versus cross-sectional research questions

“Within” – “causal” effects of time-variant variables:
→ modeling intrapersonal change (FE models)

Cross-sectional – association with time-invariant effects:
→ OLS with robust standard errors
In unbalanced panels:
→ RE models

Interpretation (e.g. presence of children):
within: effect of additional children
between: differences between people with a different number of children
Starting point RE: “null” (“Variance Components” (VC)) model

\[ y_{it} = \alpha_{0i} + \varepsilon_{it} \]  
(note: no intercept \( \alpha \) in VC model)

where:

\( \alpha_{0i} = \) individual specific random variable (N(0, \( \sigma_{u0} \)) assumed); "between"

\( \varepsilon_{it} = \) deviation from individual specific mean (N(0, \( \sigma_{\varepsilon} \)) assumed); "within"

the VC model allows for variance decomposition:

\[ \rho = \text{correlation between different time points} \ t \ \text{within an individual} \ i : \]

\[ \rho = \frac{\sigma^2}{\sigma^2_{u0} + \sigma^2_{\varepsilon}} \]  
(= ICC = intra-class-correlation = autocorrelation in Panels)

(note: \( \rho \) significant \( \rightarrow \) multilevel model necessary)
Idea RE: weighted within and between

\[ W-G \text{ regression} \quad (\frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + T \sigma^2_u}) \quad \text{pooled OLS regression} \quad B-G \text{ regression} \]

\[ \psi = \infty \]

RE - Regression is equivalent to pooled OLS after the Transformation:

\[ (y_{it} - \theta \bar{y}_i) = \beta_0 (1 - \theta) + \beta_1 (x_{it} - \theta \bar{x}_i) + (u_i (1 - \theta) + (\varepsilon_{it} - \theta \bar{\varepsilon}_i)) \]

with \[ \theta = 1 - \sqrt{\frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + T \sigma^2_u}}, \quad 0 < \theta < 1 \]

\( \sigma^2_u \) large (bias from omitted time - invariant variables may cause trouble)

\[ \rightarrow \theta \text{ close to 1} \]

\[ \rightarrow \text{RE close to FE} \]

RE allows estimation of time - invariant variables \( u_i \)

RE biased because \( u_i \) remains in error term (if \( \text{cov}(x, u_i) \neq 0) \)
Example $\Theta$ based on Satisfaction / Partner data

```
. xtreg satlife partner, re theta

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

R-sq:  within = 0.8982  between = 0.8351  overall = 0.4065

omega = 0.9613836

-------------------------------------------------------------
satlife |  Coef.  Std. Err.     z  P>|z|   [95% Conf. Interval]
-------------+--------------------------------------------------
partner | 1.00596   .0886305   11.35  0.000    0.8322479    1.179673
    _cons | 4.99851   .8120631    6.16  0.000    3.406896    6.590124

sigma_u | 1.4131587
sigma_e | 0.13377144
    rho | 0.99111886  (fraction of variance due to u_i)

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

- Estimate (1.006) is close to that from FE model (1.000) because $\Theta$ close to 1
- About 90% of variance is explained by partnership status change
Decision if FE or RE appropriate: Hausman test

Test if FE or RE model (**basic assumption: FE unbiased**)

Test \( H_0: E(u_i | x_{it}) = cov(u_i, x_{it}) = 0 \)

\[ cov(u_i, x_{it}) = 0 \rightarrow \text{FE and RE unbiased, FE is inefficient} \rightarrow \text{RE} \]

\[ cov(u_i, x_{it}) \neq 0 \rightarrow \text{FE is unbiased and RE is biased} \rightarrow \text{FE} \]

If \( H_0 \) is true (between-coeff.=within-coeff.), no differences between FE and RE equivalently:

Hausman compares estimation coefficients \( \hat{\beta}_{FE} \) and \( \hat{\beta}_{RE} \)

if \( H_0, \hat{\beta}_{FE} = \hat{\beta}_{RE} \) and \( \hat{\beta}_{RE} \) more efficient \((var(\hat{\beta}_{FE}) > var(\hat{\beta}_{RE}))\)

if \( H_1, \hat{\beta}_{FE} \) unbiased but \( \hat{\beta}_{RE} \) not

**Note:**
- \( H_0 \) almost always rejected (sample size high enough even with small differences)
- Test is only formal and does not replace research question driven check for model appropriateness
Hausman Test: RE or FE estimate?

\texttt{xtreg satlife partner, re}
\texttt{estimates store randeff}
\texttt{xtreg satlife partner, fe}
\texttt{estimates store fixdeff}
\texttt{hausman fixdeff randeff, sigmamore}

---- Coefficients ----
|      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B)) |
|    fixdeff      randeff       Difference          S.E.        |
-------------+----------------------------------------------------------------
partner |    .9999999      1.00596       -.0059605        .0024201   |

\begin{itemize}
  \item \texttt{b} = consistent under \texttt{Ho} and \texttt{Ha}; obtained from \texttt{xtreg}
  \item \texttt{B} = inconsistent under \texttt{Ha}, efficient under \texttt{Ho}; obtained from \texttt{xtreg}
\end{itemize}

Test: Ho: difference in coefficients not systematic

\[
\text{chi2}(1) = (b-B)' \left[ \left( V_b - V_B \right)^{-1} \right] (b-B)
\]
\[
= 6.07
\]
Prob>\text{chi2} = 0.0138
### FE versus RE models

**Regression equation:** \( y_{it} = \alpha + \beta x_{it} + u_i + \varepsilon_{it} \)

<table>
<thead>
<tr>
<th>Fixed effects models</th>
<th>Random effect models</th>
</tr>
</thead>
<tbody>
<tr>
<td>• OLS estimated</td>
<td>• Cannot be estimated with OLS</td>
</tr>
<tr>
<td>• only variance within- individuals used</td>
<td>• Uses both within- and between- individuals variance</td>
</tr>
<tr>
<td>• Controls for unobserved heterogeneity (consistent also if ( \text{Cov}(u,x)\neq0 ))</td>
<td>• assumes exogeneity: ( \text{Cov}(u,x)=0 ) (no effects from unobserved variables)</td>
</tr>
<tr>
<td>• Effects of time-invariant characteristics cannot be estimated (e.g., gender, cohort)</td>
<td>• Effects from time-invariant and time-varying covariates</td>
</tr>
</tbody>
</table>

→ If research interest is **longitudinal or causal**

→ If research interest is 1) **cross-sectional** or 2) **on variance on different levels**

- formal test for RE against FE model: Hausman test (test of unbiasedness)
FE versus RE models: substantive questions

- Within estimators cannot estimate the effects of time-constant variables
  - sex, nationality, social origin, birth cohort, etc.
- Panel data do not help to identify the causal effect of time-constant variables
- The "within logic" applies only with time-varying variables (Something must “happen”)

Only then a before-after comparison is possible: Analyzing the effects of events

- Such questions are the main strength of panel data and the within methodology
  - [Event variables can not only be categorical, but also metric]

- If one has substantive interest in the effect of a time constant regressor, one could estimate group-specific FE models (e.g., for men and women separately).
## Fixed und random effect example, Hybrid model

<table>
<thead>
<tr>
<th>DepVar: Wage</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Z-values)</strong></td>
<td>FE</td>
<td>RE</td>
<td>RE-FE</td>
<td>RE</td>
<td>RE-FE</td>
</tr>
<tr>
<td>Occupational status (SEI/10)</td>
<td>.037 (6.40)</td>
<td>.046 (8.67)</td>
<td>.009 (4.06)</td>
<td>.046 (8.56)</td>
<td>.009 (4.06)</td>
</tr>
<tr>
<td>Union</td>
<td>.083 (3.93)</td>
<td>.121 (6.22)</td>
<td>.038 (4.95)</td>
<td>.124 (6.39)</td>
<td>.041 (4.95)</td>
</tr>
<tr>
<td>Schooling (years) (time invariant)</td>
<td></td>
<td></td>
<td></td>
<td>0.64 (7.30)</td>
<td></td>
</tr>
<tr>
<td>Black (time invariant)</td>
<td></td>
<td></td>
<td>-.140 (2.89)</td>
<td></td>
<td>-.130 (2.77)</td>
</tr>
<tr>
<td>SEI time mean (time invariant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union time mean (time invariant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hausman chi-square</td>
<td>45.3</td>
<td>46.8</td>
<td>24.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8

Non linear regression
Non-linear regression: motivation

- Linear regression: requires continuous dependent variable e.g. BMI, income, satisfaction on scale from 0-10 (?)
- Most variables in social science are not continuous but discrete
  - Opinions: agree vs. disagree
  - Poverty status
  - Party voted for
  - Number of visits to the doctor
  - Having a partner
- We need appropriate regression models!
## Non-linear models

Dependent variable is not continuous: non-linear regression

<table>
<thead>
<tr>
<th>Binary variables (dummy variables, 0 or 1) (e.g. yes-no, event – no event)</th>
<th>Logistic Regression, Probit Regression, and many more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial (unordered variables) (e.g. vote choice, occupation)</td>
<td>Multinominal logistic Regression</td>
</tr>
<tr>
<td></td>
<td>Multinominal probit Regression</td>
</tr>
<tr>
<td>Ordinal (e.g. satisfaction)</td>
<td>Ordinal Regression</td>
</tr>
<tr>
<td>Count variable (e.g. doctor visits)</td>
<td>Poisson Regression</td>
</tr>
<tr>
<td></td>
<td>Negative Binomial Regression</td>
</tr>
</tbody>
</table>
Linear probability model for binary variables

\[ E(Y) = P(Y_i = 1) \]

Linear probability model: \[ P(Y_i = 1 \mid X_i) = \alpha + \beta_1 x_1 + \ldots \]

Estimation with OLS regression

\[
\begin{align*}
y_i &= \begin{cases} 
1 & \text{if } y_i^* \geq 0 \\
0 & \text{if } y_i^* < 0
\end{cases}
\end{align*}
\]
Advantages and problems of linear probability model

- **Advantages**
  - Estimation with OLS regression
  - Direct interpretation of coefficients
  - Less biased if \( P(Y=1) \) not too close to 0 or 1

- **Problems: violation of regression assumptions**
  - Predicted probabilities may be negative or greater than one
  - Relationship between response probability and \( x \) may not be linear, especially for \( P(Y=1) \) close to 0 and \( P(Y=1) \) close to 1
  - The variance of \( y \) for binary variables is \( P(Y=1) \times P(Y=0) \)
    → residual variance depends on \( x \)
    → heteroskedasticity
  - Residuals can take only two values for fixed \( x \)
    → residuals are not normally distributed
- $P(Y=1)$ between 0 and 1
- Non-linear
- Symmetry
- Often used function: Cumulative logistic distribution (Logit model)

$$\Pr(y = 1) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots)}}$$

- Other functions also used (e.g. probit). For practical purposes, these models provide very similar predicted probabilities
Generalised linear model

- A latent (unobserved) continuous variable $y^*$ which underlies the observed data
  
  $$y^* = a + b_1 x_1 + \ldots + e^*$$

- Assume $y_i^*$ is generated by a linear regression structure
- Link function between $y$ and $y^*$: $E(y) = f(y^*) = f(a + b_1 x_1 + \ldots + e^*)$
  
  e.g. logit, probit, poisson, negative binomial, identity

- Because $y^*$ is not observed
  - $e_i^*$ are not observed, variance of $e_i^*$ has to be assumed

  Logit model: $e_i^*$ standard logistic (with $\text{var} = \pi^2/3 \approx 3.29$)
Maximum likelihood estimation (MLE)

- Usually, non-linear models are estimated by maximum likelihood
- Principle for MLE: Which set of parameters has the highest likelihood to generate the data actually observed \((x_i, y_i)\)?
- Advantages
  - Extremely flexible and can easily handle both linear and non-linear models (Linear model: MLE = OLS estimator)
  - Desirable asymptotic properties: consistency, efficiency, normality, (consistent if missing at random MAR)
- Disadvantages
  - Requires assumptions on distribution of residuals
  - Desired properties hold only if model correctly specified
  - Best suited for large samples
- Often, there is no closed form (algebraic) solution. Coefficients have to be estimated through iteration methods
Example: logistic regression

.logit svp female age1830 age4660 age60plus satdem

Logistic regression Number of obs = 6,224
LR chi2(5) = 124.97
Prob > chi2 = 0.0000

Log likelihood = -2193.4524 Pseudo R2 = 0.0277

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>z</th>
<th>P&gt;z</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>svp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>-.4469116</td>
<td>.0798622</td>
<td>-5.60</td>
<td>0.000</td>
<td>-.6034387 -.2903845</td>
</tr>
<tr>
<td>age1830</td>
<td>-.0496199</td>
<td>.1343308</td>
<td>-0.37</td>
<td>0.712</td>
<td>-.3129034 .2136636</td>
</tr>
<tr>
<td>age4660</td>
<td>-.0781095</td>
<td>.1185982</td>
<td>-0.66</td>
<td>0.510</td>
<td>-.3105577 .1543388</td>
</tr>
<tr>
<td>age60plus</td>
<td>.0865025</td>
<td>.1152573</td>
<td>0.75</td>
<td>0.453</td>
<td>-.1393976 .3124026</td>
</tr>
<tr>
<td>satdem</td>
<td>-.1987384</td>
<td>.0199542</td>
<td>-9.96</td>
<td>0.000</td>
<td>-.2378478 -.1596289</td>
</tr>
<tr>
<td>_cons</td>
<td>-.5746696</td>
<td>.1567911</td>
<td>-3.67</td>
<td>0.000</td>
<td>-.8819746 -.2673647</td>
</tr>
</tbody>
</table>
Interpretation of non-linear models

- $y^*$ has no units, scale of $y^*$ changes if additional $x_i$ are included
- Because of the non-linearity, effects depend on values of $x$ and cannot be interpreted directly (→ not constant)
- Coefficients cannot be compared across different models
- Interpretation of coefficients
  - Qualitative interpretations (direction and significance level)
  - Odds ratio (problematic)
  - Predicted probabilities
Excursus: Odds ratios (OR)

OR often misunderstood as relative risk

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>Odds (\frac{P}{1-P})</th>
<th>OR (\frac{Odds\ Group 1}{Odds\ Group 2})</th>
<th>RR (\frac{P\ Group 1}{P\ Group 2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Group 1</td>
<td>0.10</td>
<td>0.11</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>Group 2</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Group 1</td>
<td>0.40</td>
<td>0.67</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>Group 2</td>
<td>0.20</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Group 1</td>
<td>0.80</td>
<td>4.00</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>Group 2</td>
<td>0.40</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>Group 1</td>
<td>0.60</td>
<td>1.50</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>Group 2</td>
<td>0.20</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Group 1</td>
<td>0.40</td>
<td>0.67</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>Group 2</td>
<td>0.10</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>

Ref.: Best and Wolf 2012, Kölner Zeitschrift für Soziologie 8-11
Compute predicted probabilities

- Remember: predicted probabilities depend on values of \( x \) and parameter estimates (and unobserved heterogeneity)
- Predicted probabilities are estimates → confidence intervals
- Discrete change: predict probabilities for different values of \( x \)
- Marginal effect or partial effect: The slope of \( \Pr(y=1) \) at \( x \).
- Two methods
  - Adjusted predictions: Specify values for each of the independent variables, compute probability for individual who has those values. Usually: \( x \) at the mean; Alternative: representative values
  - Average effects: Compute predicted probability for each individual at observed values of \( x \). Average probability over all individuals (average marginal effect, average adjusted predictions)
Predicted probabilities: example

Logit model

\[
Pr(y = 1) = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2)}} = \frac{\exp(\text{logit})}{1 + \exp(\text{logit})}
\]

Hypothetical regression result: \( y^* = 3 - 2x_1 + 0.5x_2 \)

Example Individual i with \( x_1 = 2 \) and \( x_2 = 1 \),

\[
P(Y_i = 1 \mid X_i) = \frac{1}{1 + e^{-(3 + (-2)2 + 0.5)}}
\]

\[
= \frac{1}{1 + e^{0.5}} = 0.62
\]
Example: Average adjusted predictions (AAP)

logit svp i.female i.agegr satdem
margins female agegr

<table>
<thead>
<tr>
<th>Margin</th>
<th>Std. Err.</th>
<th>z</th>
<th>P&gt;z</th>
<th>[95% Conf.Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.1434225</td>
<td>0.0066169</td>
<td>21.68</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>0.0974338</td>
<td>0.0049952</td>
<td>19.51</td>
<td>0.000</td>
</tr>
<tr>
<td>agegr</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-30</td>
<td>0.1131354</td>
<td>0.0094684</td>
<td>11.95</td>
<td>0.000</td>
</tr>
<tr>
<td>31-45</td>
<td>0.1180999</td>
<td>0.0095385</td>
<td>12.38</td>
<td>0.000</td>
</tr>
<tr>
<td>46-60</td>
<td>0.1103665</td>
<td>0.0069987</td>
<td>15.77</td>
<td>0.000</td>
</tr>
<tr>
<td>over 60</td>
<td>0.127197</td>
<td>0.0072908</td>
<td>17.45</td>
<td>0.000</td>
</tr>
</tbody>
</table>
1. **Average marginal effects**

   \[ \text{margins}, \ dydx(\text{satdem}) \]

   |         | Std. Err. |    Z |    P>|z| |  [95% Conf. Interval] |
   |---------|-----------|-----|--------|-------------------------|
   | satdem  | -0.0201908| 0.0020354| -9.92  | 0.000 | -0.0241801, -0.0162015 |

2. **Average adjusted predictions**

   Average adjusted predictions
Model performance

- Linear regression: $R^2$, adjusted $R^2$
- Non linear regression
  - variance of the residual not observed
  - many so-called pseudo-$R^2$ (0,1) (Stata: fitstat)

\[
R^2 = \frac{\text{explained var. in } y}{\text{total var. in } y}
\]

Log-Lik Intercept Only: $-2255.939$  
Log-Lik Full Model: $-2193.452$
D(6216): $4386.905$  
LR(5): $124.973$
Prob > LR: $0.000$
McFadden's R2: $0.028$  
McFadden's Adj R2: $0.024$
Maximum Likelihood R2: $0.020$  
Cragg & Uhler's R2: $0.039$
McKelvey and Zavoina's R2: $0.052$  
Efron's R2: $0.020$
Variance of y*: $3.469$  
Variance of error: $3.290$
Count R2: $0.882$  
Adj Count R2: $0.000$
AIC: $0.707$  
AIC*n: $4402.905$
BIC: $-49917.116$  
BIC': $-81.292$
The likelihood ratio test

- To test hypotheses involving several predictors (multiple constraints) (e.g. Test $\beta_2 = \beta_3 = 0$)
- Compare log-likelihoods of constrained and unconstrained model, e.g.
  - $M_u: \pi = (F(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3))$
  - $M_c: \pi = (F(\alpha + \beta_1 x_1))$
- Generally: $L_c \leq L_u$
  - Constraints valid: $L_u - L_c = 0$
  - Constraints invalid: $L_u - L_c > 0$
- Test statistic: $LR = 2(L_u - L_c) \sim \chi^2(q)$; ($q$: number of constraints, d.o.f.)
- Prerequisites of LR test
  - Models are based on the same sample
  - Models are nested
LR test example: test for joint significance

- Vote intention model (support SVP vs. supporting another party)
  \[ P(Y = 1|X) = F(\alpha + \beta_1(\text{female}) + \beta_2(\text{age})) + \beta_3(\text{lnincome}) + \beta_4(\text{contra EU}) + \beta_5(\text{for nuclear energy}) + \beta_6(\text{satisfaction democracy}) \]

- Do demographic variables (age, sex) matter? \( H_0: (\beta_1, \beta_2) = 0 \)

<table>
<thead>
<tr>
<th></th>
<th>Unconstrained</th>
<th>Constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.143*</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>-0.296***</td>
<td>-0.301***</td>
</tr>
<tr>
<td>Contra EU</td>
<td>0.834***</td>
<td>0.819***</td>
</tr>
<tr>
<td>Pro nuclear energy</td>
<td>0.241***</td>
<td>0.185**</td>
</tr>
<tr>
<td>Satisfaction democracy</td>
<td>-0.213***</td>
<td>-0.216***</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-3762.1</td>
<td>-3772.2</td>
</tr>
</tbody>
</table>

- LR test:
  \[ LR = 2( L_u - L_c ) \approx 2( -3762.1 - (-3772.2)) = 20.08 \]

- Chi-squared distribution with 2 degrees of freedom: \( p=0.0000 \)
  - we reject \( H_0 \) (demographic variables seem to affect the probability to vote for SVP)
Difficulties of nonlinear models: frequent mistakes

- Interpretation of coefficients (Logits, OR)
- Comparison of estimates across models and samples (estimates reflect also unobserved heterogeneity)
  - Be cautious with interpretation
  - Use different measures to show effects (predicted probabilities)
  - Correction proposed by Karlson et al. (2012)
    References: Mood 2010, Best and Wolf 2012, Karlson et al. 2012, Stata: ado kha
- Interaction effect: cannot be interpreted as in linear models
- Model performance
Multinomial dependent variables

- More than two response categories (m categories)
- Unordered → Multinomial regression
e.g. Voting preference (different parties), type of education, compare each pair of response categories
  - estimate probability for each category,
    (1 reference category, m-1 equations)
- Ordered → Ordinal regression
e.g. Opinions (strongly agree, agree, neither, disagree, strongly disagree), health status
  - latent variable with m-1 thresholds
  - estimate cumulative probability (prob. y ≤ mi)
    (one equation with dummies for m-1 thresholds)
Example: multinomial regression

Voting: FDP/CVP, SP/Greens, other parties; Base category: vote SVP

<table>
<thead>
<tr>
<th></th>
<th>FDP &amp; CVP</th>
<th>SP &amp; Greens</th>
<th>Other party</th>
<th>No party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.377***</td>
<td>0.406***</td>
<td>0.274*</td>
<td>0.583***</td>
</tr>
<tr>
<td>Age 18 - 30</td>
<td>-0.162</td>
<td>0.070</td>
<td>-0.167</td>
<td>-0.220</td>
</tr>
<tr>
<td>Age 46 - 60</td>
<td>-0.092</td>
<td>0.031</td>
<td>-0.129</td>
<td>-0.075</td>
</tr>
<tr>
<td>Age 60+</td>
<td>-0.052</td>
<td>-0.413**</td>
<td>-0.584**</td>
<td>-0.274*</td>
</tr>
<tr>
<td>Education: intermed</td>
<td>0.179</td>
<td>0.300*</td>
<td>0.206</td>
<td>-0.135</td>
</tr>
<tr>
<td>Education: high</td>
<td>0.780***</td>
<td>0.982***</td>
<td>1.114***</td>
<td>0.287</td>
</tr>
<tr>
<td>Income (ln)</td>
<td>0.317***</td>
<td>0.229**</td>
<td>0.397***</td>
<td>0.113</td>
</tr>
<tr>
<td>Against EU-integration</td>
<td>-1.704***</td>
<td>-2.790***</td>
<td>-1.559***</td>
<td>-1.665***</td>
</tr>
<tr>
<td>Against Foreigners</td>
<td>-0.746***</td>
<td>-1.608***</td>
<td>-1.254***</td>
<td>-0.862***</td>
</tr>
<tr>
<td>Pro nuclear energy</td>
<td>-0.064</td>
<td>-1.542***</td>
<td>-0.967***</td>
<td>-0.570***</td>
</tr>
<tr>
<td>Satisfaction democracy</td>
<td>0.252***</td>
<td>0.193***</td>
<td>0.183***</td>
<td>0.044</td>
</tr>
<tr>
<td>_cons</td>
<td>-3.138***</td>
<td>-0.694</td>
<td>-4.259***</td>
<td>1.364</td>
</tr>
</tbody>
</table>

8-21
Fixed effects models

\[ Y_{it} = \beta_1 x_{it} + \beta_1 x_{it} + \alpha_i + \epsilon_{it} \]

- \( \alpha_i \): unobservable stable individual characteristics (as variable, not residual)
- only variance within individuals taken into account
- Control for unobserved heterogeneity (consistent also if \( \text{Cov}(\alpha, x) \neq 0 \))
- Effects of time-invariant characteristics cannot be estimated (e.g., sex, cohorts)

Random effect models

\[ Y_{it} = \alpha + \beta_1 x_{it} + \beta_1 x_{it} + \alpha_i + \epsilon_{it} \]

- assumes \( \alpha_i \sim N(0, \sigma_\alpha) \)
- \( \alpha_i \): unobservable stable individual characteristics, part of residual
- Multilevel model with random intercepts
- Controls for unobserved heterogeneity (but consistent only if \( \text{Cov}(\alpha, x) = 0 \))
- Effects of time-invariant and time-varying covariates
Fixed effects for non-linear models

- Linear model: by differencing out (or including dummy variables), the $u_i$ disappears from (FE) equation
- Non-linear model: there is no equivalent FE model
  - Incidental parameter problem -> inconsistent estimates
- Instead: Conditional ML estimation (similar to FE)
  - Technical trick to eliminate individual-specific intercepts (number of 1 for each individual as sufficient condition)
  - (Also called Chamberlain fixed-effects model)
  - Only possible for logit and poisson
    (here possibly: Logistic Fixed Effects Estimation for two time periods.doc)
- Drawback: Only subsample of individuals with change in $y_{it}$
  -> information loss
  -> potential bias from excluding stable individuals (external validity)
- Linear probability model (FE) used as alternative
Cross-sectional analysis: ordered logistic estimation, ordered probit model

No Fixed Effects estimator, but different Strategies proposed

- Dichotomise variables and estimate fixed effects logit (choose one cut point)
- Estimate logistic model with every possible dichotomizing cutoff point and then combine the results (Das and van Soest 1999)
- Estimate logistic model with every possible dichotomizing jointly (Beatschmann, Staub and Winkelmann 2011)
- Dichotomise every individual separately (Ferrer-i-Carbonell and Frijters 2004), most frequently at the mean
Random effects for non-linear models

RE model equivalent to linear regression

\[ y_{it} = F(\alpha_0 + \alpha_i + x_{1it}\beta_1 + x_{2it}\beta_2 + \ldots + \alpha_i + \epsilon_{it}) \]

with \( \alpha_i \sim N(0, \sigma_\alpha) \), \( \text{Cov}(\alpha_i, x_i) = 0 \)

But in contrast to linear models

- Predicted probabilities depend on values of \( u_i \): we have to assume a value for \( u_i \) to estimate predicted probabilities
- Measures for variance decomposition questionable
  - only variance of unobserved heterogeneity estimated, within variance is fixed (usually at 1)
    - \( (\rho) \) not meaningful
  - Alternative: How much can the unexplained variance between individuals be reduced relative to the empty model?
Stata commands for non-linear panel models

Stata built-in commands

- Random intercept models: xt prefix
  - xtlogit, xtprobit, xtpoisson, xttobit, xtcloglog, xtnbreg, xtologit
- Random slope models: meqrlogit, meqrpoisson

Other software necessary for multinomial panel models (run from Stata)

- gllamm add-on (Rabe-Hesketh and Skrondal) to Stata
  (very powerful, freely available (but Stata necessary), slow, become familiar with syntax)
- runmlwin: command to run mlwin software from within Stata
  (Mlwin software needs to be purchased)
<table>
<thead>
<tr>
<th></th>
<th>logit</th>
<th>random</th>
<th>fe</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>0.116**</td>
<td>0.214***</td>
<td>.0</td>
</tr>
<tr>
<td>18 - 30 years</td>
<td>-0.133**</td>
<td>-0.033</td>
<td>0.620***</td>
</tr>
<tr>
<td>46 - 60 years</td>
<td>-0.119**</td>
<td>-0.162**</td>
<td>-0.306***</td>
</tr>
<tr>
<td>over 60 years</td>
<td>-0.097*</td>
<td>-0.250***</td>
<td>-0.601***</td>
</tr>
<tr>
<td>med education</td>
<td>-0.229***</td>
<td>-0.497***</td>
<td>-0.318*</td>
</tr>
<tr>
<td>high education</td>
<td>-0.633***</td>
<td>-1.350***</td>
<td>-0.537***</td>
</tr>
<tr>
<td>hh income, log</td>
<td>-0.128***</td>
<td>-0.159***</td>
<td>-0.013</td>
</tr>
<tr>
<td>stay outside EU</td>
<td>0.749***</td>
<td>0.622***</td>
<td>0.049</td>
</tr>
<tr>
<td>prefer Swiss</td>
<td>0.471***</td>
<td>0.402***</td>
<td>-0.010</td>
</tr>
<tr>
<td>nuc_pro</td>
<td>0.225***</td>
<td>0.076</td>
<td>-0.073</td>
</tr>
<tr>
<td>satisf. democracy</td>
<td>-0.160***</td>
<td>-0.153***</td>
<td>-0.043***</td>
</tr>
<tr>
<td>_cons</td>
<td>1.190***</td>
<td>1.818***</td>
<td>2.034***</td>
</tr>
<tr>
<td>Lnsig2u _cons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>55177</td>
<td>55177</td>
<td>26420</td>
</tr>
</tbody>
</table>
RE logistic model for event history analysis

- Example for discrete event history analysis
- Dependent variable
  - 0 event has not occurred
  - 1 event has occurred (since last observation)
- Independent variable
  - Time until event occurrence
  - Any other variable
- Estimate logistic model or random effect logistic model
- Example: change of vote intention (between parties)

<table>
<thead>
<tr>
<th>Ind.</th>
<th>Wave</th>
<th>Vote intention, Party</th>
<th>Change between parties</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>A</td>
<td>.</td>
<td>24</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>B</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>B</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>A</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>No party</td>
<td>0</td>
<td>28</td>
</tr>
</tbody>
</table>

Stata: `xtlogit change age age^2`
Example: discrete event history analysis