Dr. Christian Czymara

EINFÜHRUNG IN DIE PANELREGRESSION

Day 1

DeZIM Summer School 2023

AGENDA

- Introduction & course structure
- Software, introduction to R
- Panel data management
- OLS assumptions and panel data

INTRODUCTION & COURSE STRUCTURE

LECTURER

- Fellow at *University of Tel Aviv* & lecturer at *Goethe University Frankfurt*
- Research interests: Immigration & integration, inter-group conflict, attitudes, mass media, political communication
- Methods: "Classical" quantitative methods for social research combined with computational methods and natural language processing
- More info: https://czymara.com/
- Contact me at czymara@tauex.tau.ac.il

GENERAL INFORMATION

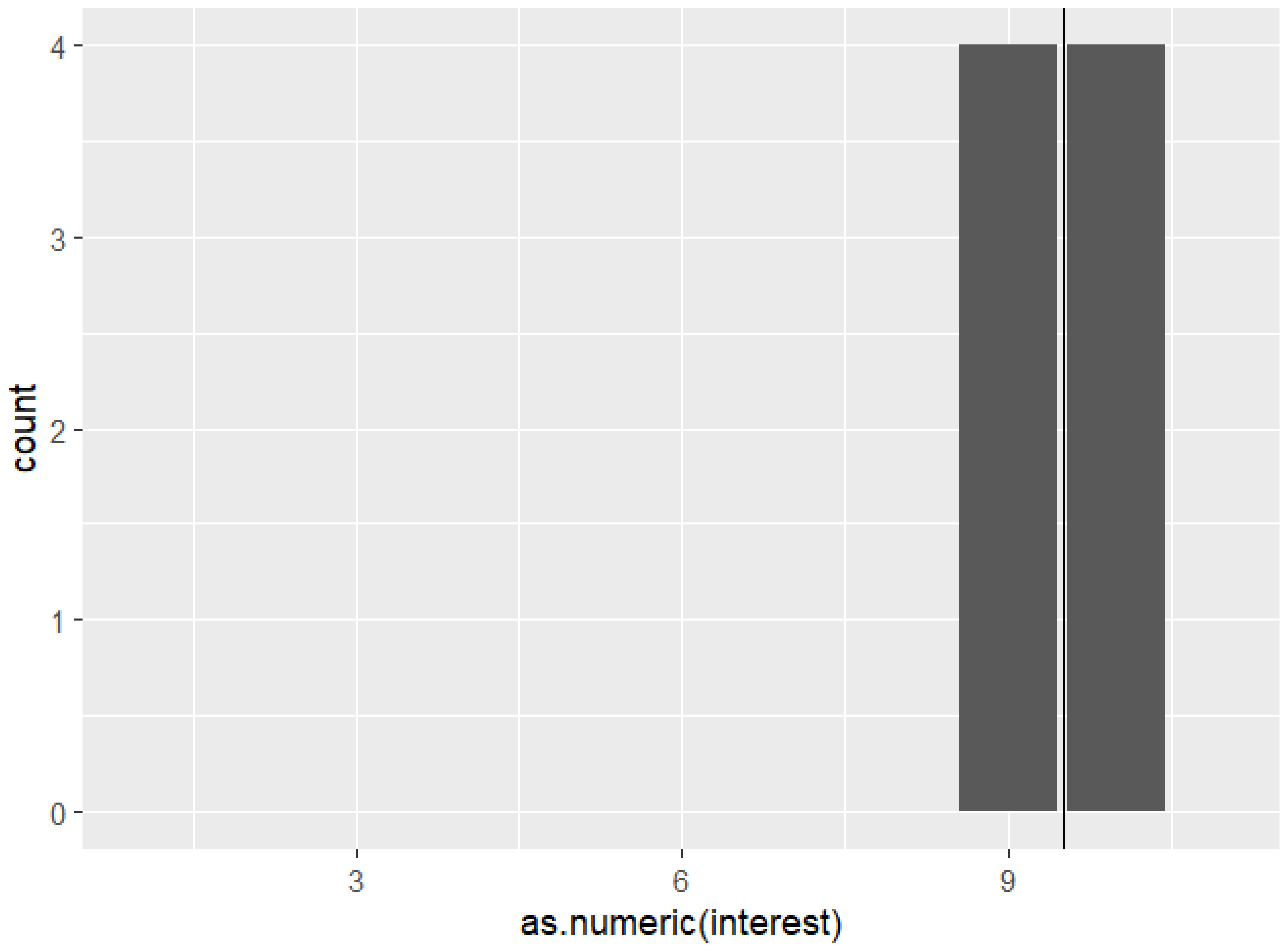
- 2 days, 09:00-17:00
  - First session: 09:00-12:30
  - Lunch break: 12:30-13:30
  - Second session: 13:30-17:00
- Each session consists of a lecture-style talk and a practical computer exercise (and a 15-minute break in between)
- Material available at: https://github.com/czymara/panelreg_DeZIM
- Slides in English, Kurs auf Deutsch ☺
- 100% für dozen
- Response rate: 50% (9/18)

YOU SHOULD HAVE…

- Interest in quantitative social research
- Good working knowledge of descriptive and inductive statistics (i.e., linear regression)
- Some knowledge of R or another statistics software / language

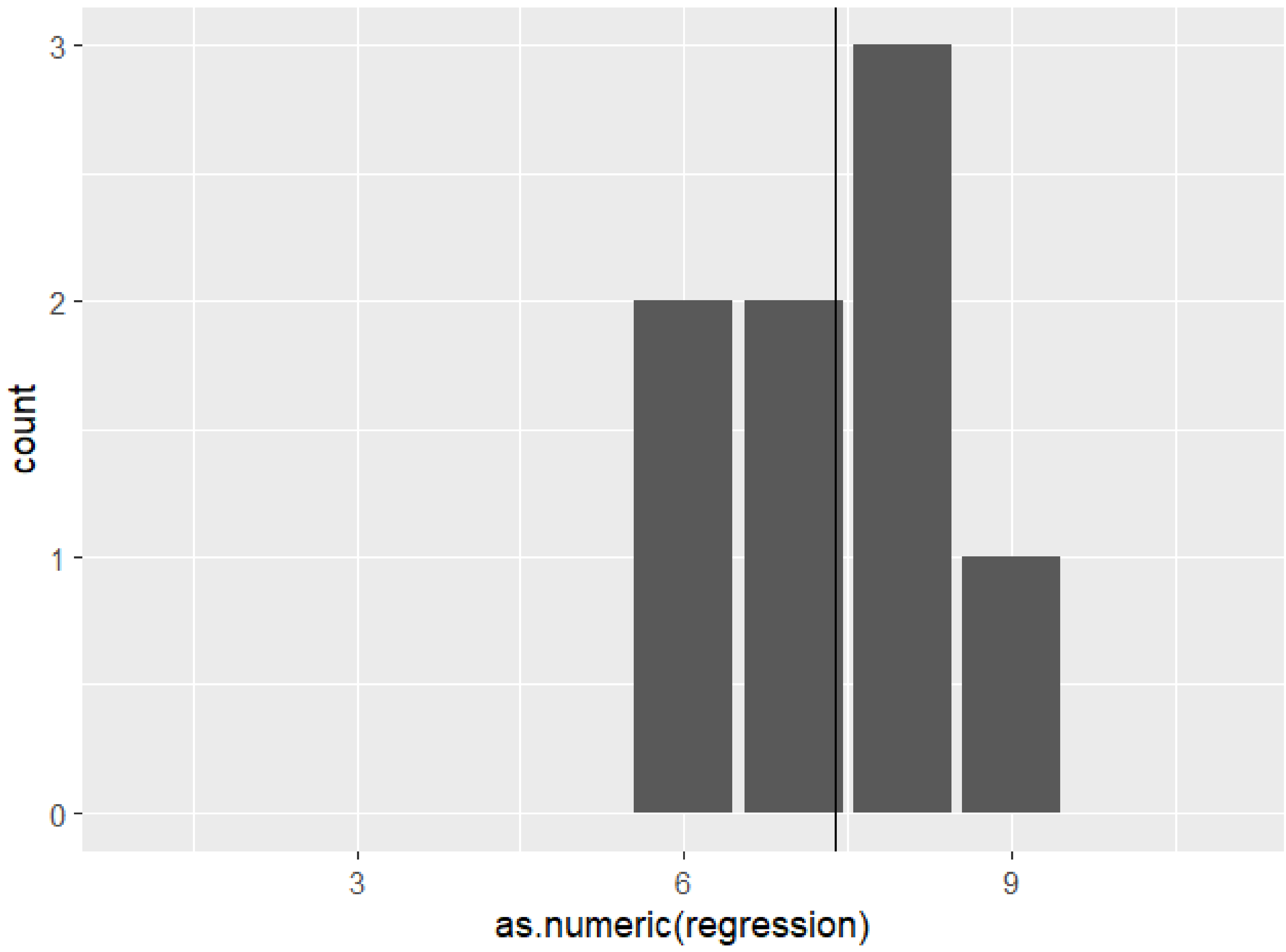
YOUR INTEREST

- Super motivated class: mean is 9.5/10



YOUR KNOWLEDGE

- Subjective knowledge also rather high (7.4 of 10)
- Four experts (>=8)
- Four medium to advanced (6 and 7)
- No one (who answered the question) is unfamiliar (below 5)



YOUR KNOWLEDGE

- 38% have already worked with panel data
- … No one has never heard of it
- Everybody (who answered the question) has already run linear regression, 87.5% logistic regression
- 55% usually work with R
- … But there are also people in the class who do not regularly use statistical software

YOUR EXPECTATIONS

- *“random and fixed effects”*
- *“random coefficient models, random slopes, dynamic models”*
- *“understand how time-series cross-sectional data such as the ESS or the ALLBUS can be analyzed using approaches such as the Mundlak model”*
- *“model specification and convergence issues in R”*
- *“application of weights to account for attrition” / “missing data”*

WHAT THIS COURSE WILL OFFER

- An introduction to the analysis of different types of longitudinal data
- … and why it may help to tackle the notoriously difficult issue of causality
- The means to conduct your own research
- Hands-on application of methods in exercises

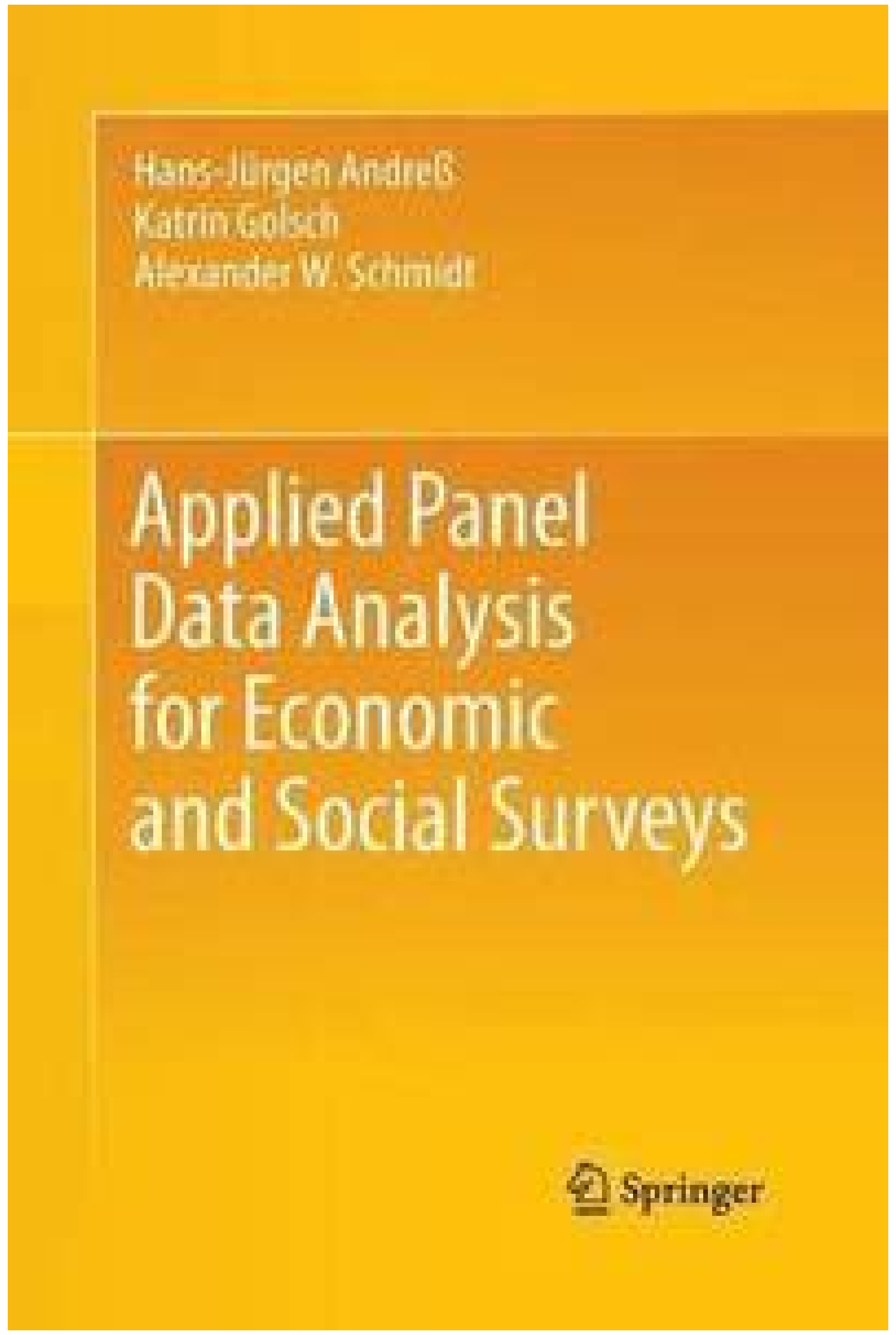
WHAT THIS COURSE WILL NOT OFFER

- Discussion of substantive theories
- In-depth understanding of mathematical foundation of methods
- Course is less suited as a general introduction into empirical research

QUESTIONS OR COMMENTS?

LITERATURE

- See GitHub for literature on individual sessions
- Literature on methods
- General textbook: Andreß, Golsch & Schmidt. *Applied panel data analysis for economic and social surveys*. Springer Science & Business Media, 2014



SOFTWARE

R

- You will need R for all tutorials
- To work with R, install on your computers
  - R: https://cloud.r-project.org/
  - RStudio: https://www.rstudio.com/products/rstudio/download/

GITHUB

- Material will be uploaded on GitHub
- Link: https://github.com/czymara/panelreg_DeZIM
- You can download files without having an account
- For advanced users: Feel free to make an account and use GitHub Desktop

PANEL DATA

PANEL DATA

- Panel data contain repeated observations of the same units

Table A.1. Inflows of foreign population into selected OECD countries and Russia

Cross-sectional dimension

Longitudinal dimension

Thousands

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Australia	203,9	219,4	202,2	206,4	236,0	244,8	233,9	223,7	218,5	224,2	186,6
Austria	94,4	91,7	96,9	109,9	125,6	135,2	154,3	198,7	158,7	139,3	131,7
Belgium	106,0	102,7	113,6	117,9	128,9	117,6	106,3	128,8	103,2	109,5	116,8
Canada	247,2	252,2	280,7	248,7	257,8	259,0	260,3	271,8	296,4	286,5	321,0
…											

https://www.oecd.org/els/mig/keystat.htm

PANEL DATA

- At least two repeated observations
- At least two units of analysis, otherwise we'd rather speak of time-series data
- Any units of analysis, e.g.
  - Individuals, Households, Countries, Parties, Organizations etc.
- Different time intervals between repeated measurements
  - Hourly, Daily, Weekly, Yearly etc.

(SOME) IMPORTANT TERMS

- Balanced panel: Each unit is observed at each time-point (i.e., number of observations is the same for each unit)
- Unbalanced panel: A panel with missing values (i.e., number of observations per unit differs)
  - Panel attrition: Units of analysis drop out of the panel permanently
  - Non-response:
    - Temporary unit non-response
    - Panel attrition
    - Late entries (e.g. refreshment samples)
    - Rotating panels

OPPORTUNITIES

- Monitor social change (e. g. development of immigration rates)
- Examine change at the individual level instead of aggregate trends → May circumvent ecologic fallacy (inference on the individual level based on aggregate relationships)

PROBLEMS OF CROSS-SECTIONAL DATA

- Researchers normally want to make *causal* statements about the association of two variables
- Causal means that the correlation of x and y is not driven by another variable z (spurious correlation)
- The best way to establish this are experiments

→Randomly assigning individuals in treatment and control group

→All z are equally distributed between both groups

PROBLEMS OF CROSS-SECTIONAL DATA

- However, experiments often not feasible in social sciences
- Observational studies thus adjust for z by statistical *controlling* after data collection
- However, z is often not observed in the data at hand
- As a result, estimates based on cross-sectional data are often plagued by *omitted variable bias*
- This is the case if unobserved characteristics are correlated with the variables in the model (endogeneity)

SOLUTIONS OF LONGITUDINAL DATA

- With longitudinal data you can deal control *even for (some) unobserved characteristics*!
- Some ≙ all time constant characteristics
- This is because individuals act as "their own controls"
- This does not ensure causality
- But it at least comes closer
- Still, it is important to think about the *theoretical* model (e.g., using a DAG)

THE POWER OF PANEL DATA

"It is hard to overstate the gain in identifying power provided by the beautifully simple method of [Fixed Effects] estimation over standard cross-sectional estimators"

- Gangl 2010: 34

EXAMPLE

WHAT CAN WE LEARN FROM PANEL DATA?

- Describing and analyzing aggregate change

Table: Artificial Panel Data with Binary Indicator of Return Intentions

ID	Return 2020	Return 2021	Return 2022
1	0	0	1
2	1	0	0
…	…	…	…
999	1	1	1
1000	0	0	0
Sum	0.10	0.12	0.14

WHAT CAN WE LEARN FROM PANEL DATA?

- Describing and analyzing individual change

Table: Artificial Panel Data with Binary Indicator of Return Intentions

ID	Return 2020	Return 2021	Return 2022
1	0	0	1
2	1	0	0
…	…	…	…
999	1	1	1
1000	0	0	0
Sum	0.10	0.12	0.14

WHAT CAN WE LEARN FROM PANEL DATA?

- Describing and analyzing change

Table: Artificial Panel Data with Binary Indicator of Return Intentions *Version 1*

	Return intentions 2022	No return intentions 2022	Total
Return intentions 2020	4%	6%	10%
No return intentions 2020	10%	80%	90%
Total	14%	86%	100%

WHAT CAN WE LEARN FROM PANEL DATA?

- Describing and analyzing change

Table: Artificial Panel Data with Binary Indicator of Return Intentions *Version 2*

	Return intentions 2022	No return intentions 2022	Total
Return intentions 2020	9%	1%	10%
No return intentions 2020	5%	85%	90%
Total	14%	86%	100%

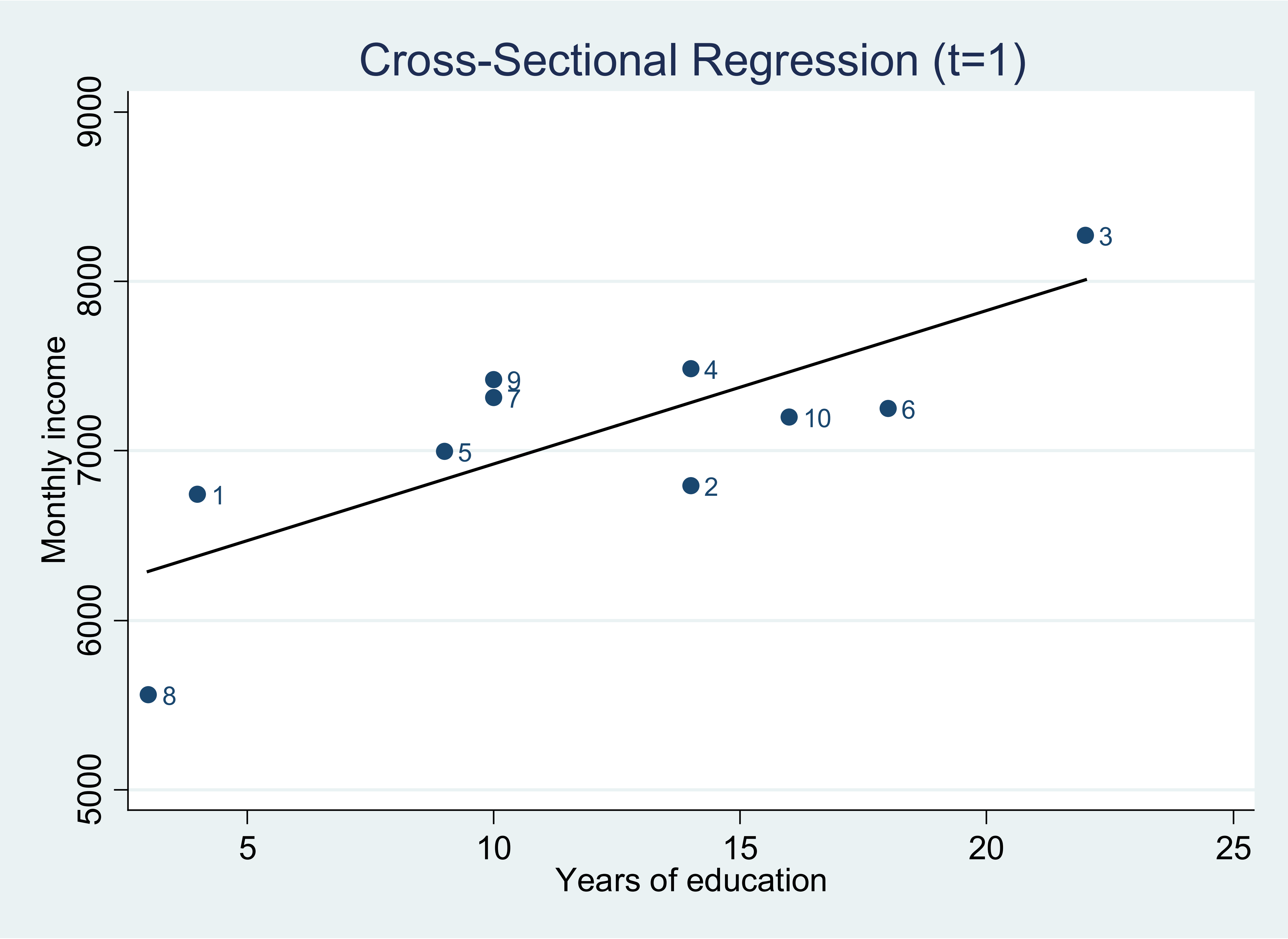
WHAT CAN WE LEARN FROM PANEL DATA?

- Separating age and cohort effects
- Age effect = maturation effect (age)
- Cohort effect = generational effect (time born)
- In Cross-sectional data, age and cohort are perfectly collinear ($birth = t - age$)
- With panel data, units of the same cohorts are observed at different ages

WHAT CAN WE LEARN FROM PANEL DATA?

- Controlling for *omitted variable bias*
- Example: What are the returns to education (*"How much does additional education financially pay off?"*)
- y = income
- x = years of education
- Both measured at two time points (t)

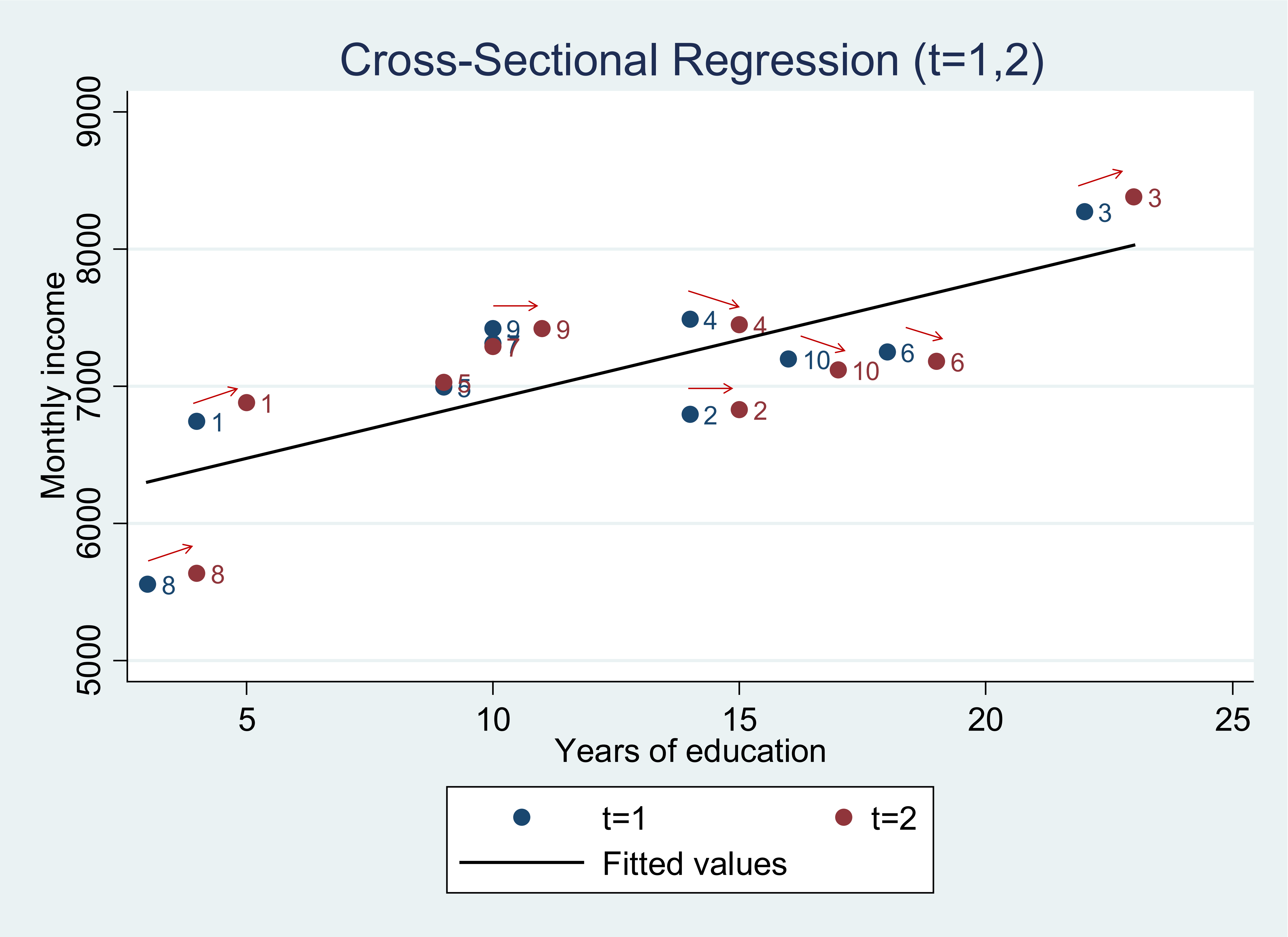
RETURNS TO EDUCATION



RETURNS TO EDUCATION

- $income_i = 6017 + 91 * educ_i + \varepsilon_i$
- Typical interpretation: "*If education increases by one unit (year) the income increases by 91 units (Euro.*"
- But is 91 the true effect of education?
- Only if we have not omitted relevant covariates
- A potential omitted variable in this example is skill

RETURNS TO EDUCATION



WHAT DID WE LEARN FROM PANEL DATA?

- On average, those with more years of education have higher income
- But…
  - Additional education does not pay off equally for everyone
  - More for those with lower levels of education
  - Not really for those with medium-high levels
- More broadly, observing change within units *What happens to y if x changes by one unit?*
- Any unobserved time-constant characteristics can be controlled by comparing within and not between units

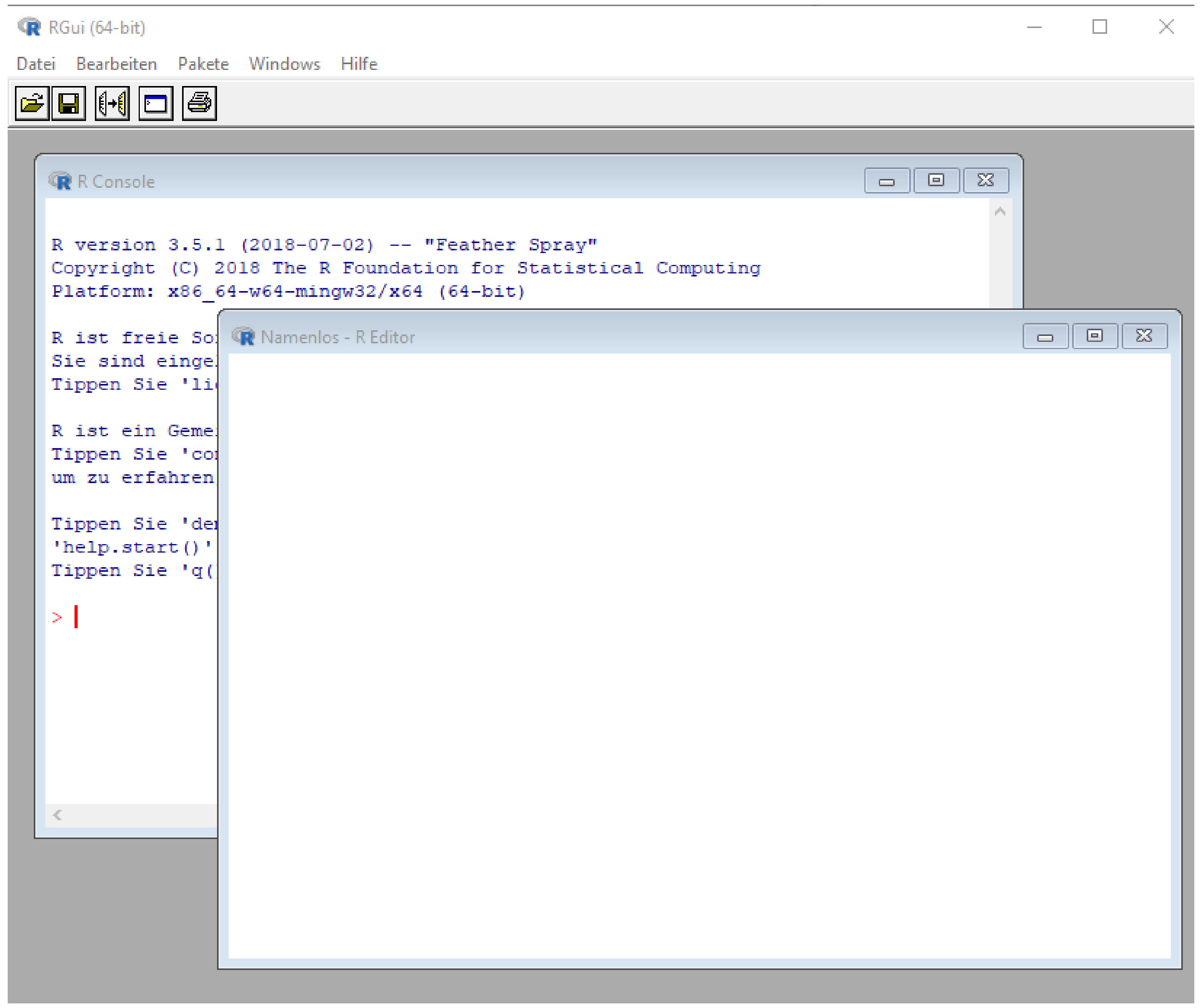
QUESTIONS OR COMMENTS?

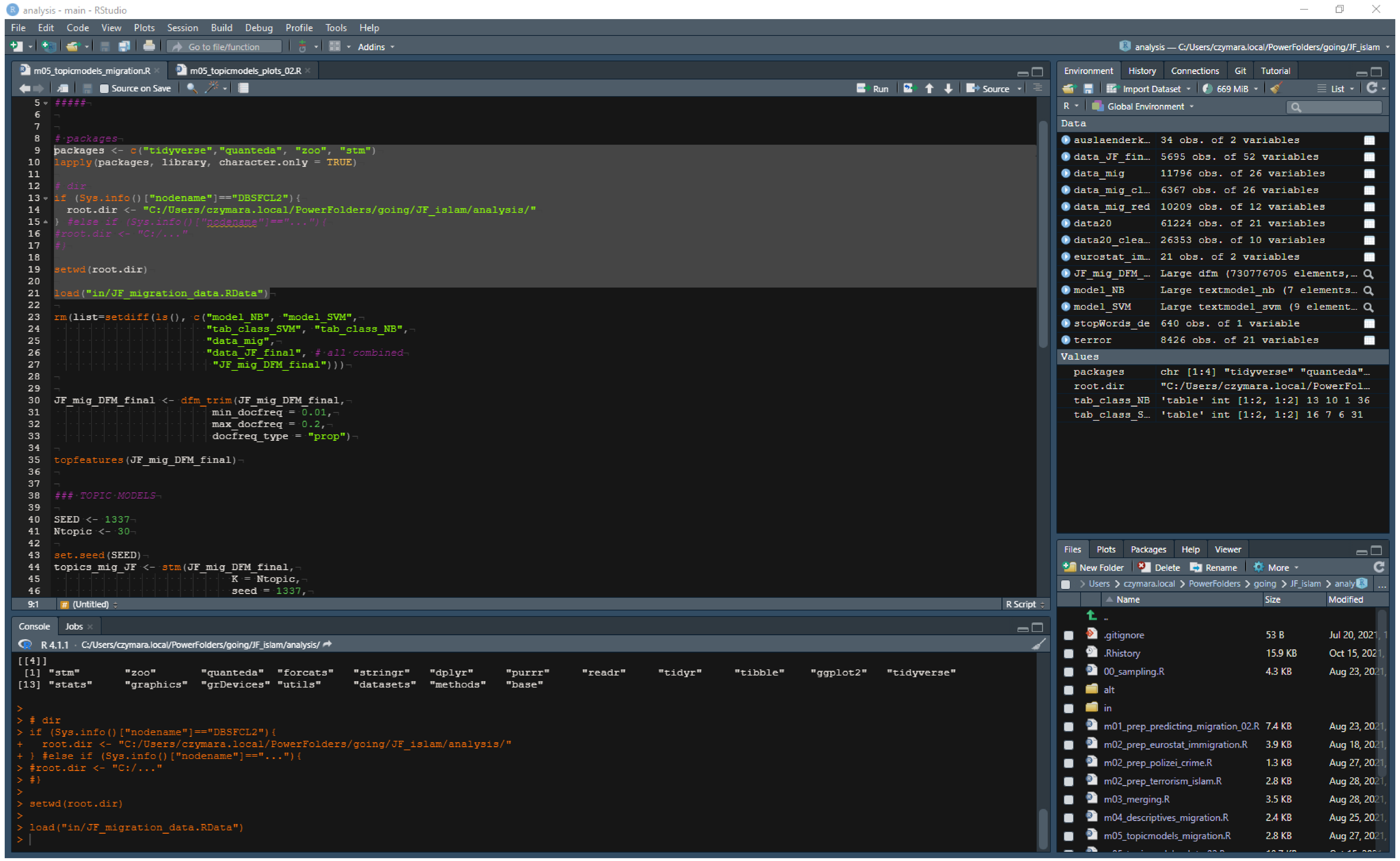
15 minutes break

R

- Why "R"? → *"R is an implementation of the S programming language"* (Wikipedia)
- R is a programming language for data analysis
- Rstudio is a so-called Integrated Development Environment (IDE), making your work a lot easier
  - Writing and running R Code
  - Overview of stored objects
  - Projects containing multiple files
  - Git connection
  - Etc.

R VS. RSTUDIO





R BENEFITS

- Free and open source
- Large and very helpful community
- Plethora of user-written packages on basically everything
- Very powerful tools for data manipulation and data visualization
- In addition to analyzing data, you can write programs, websites, books, and much more with R (and R Markdown)
- … and integrate with other languages

GOOGLE COLAB

- To understand the basics of R, we will work with this Google Colab

EXERCISE 1: R

Click link for exercise or see GitHub

THE NATURE OF PANEL DATA

PANEL DATA

- Panel data have a three-dimensional structure
  - Units ($i = 1, \ldots, n$): E. g. persons
  - Variables ($v = 1, \ldots, V$): E. g. poverty status
  - Time-points or *waves* ($t = 1, \ldots, T$): E. g. 2020

PANEL DATA

- How can you organize three-dimensional data space in a two-dimensional dataset?
- Cross-sectional dataset with n units and v variables:

ID	Var1	…	VarV
1	a	…	d
2	b	…	e
…	…	…	f
n	c	…	g

Units

Variables

PANEL DATA

- Two panel waves; each with n units and v variables:

ID	Var1	…	VarV
1	a	…	d
2	b	…	e
…	…	…	f
n	c	…	g

ID	Var1	…	VarV
1	a	…	d
2	b	…	e
…	…	…	f
n	c	…	g

PANEL DATA

- Time is a relevant information

ID	t	Var1	…	VarV
1	2011	a	…	d
2	2011	b	…	e
…	2011	…	…	f
n	2011	c	…	g

ID	t	Var1	…	VarV
1	2012	a	…	d
2	2012	b	…	e
…	2012	…	…	f
n	2012	c	…	g

THE PANEL DATA CUBE

- Time adds a third dimension

→ Panel data are cubic

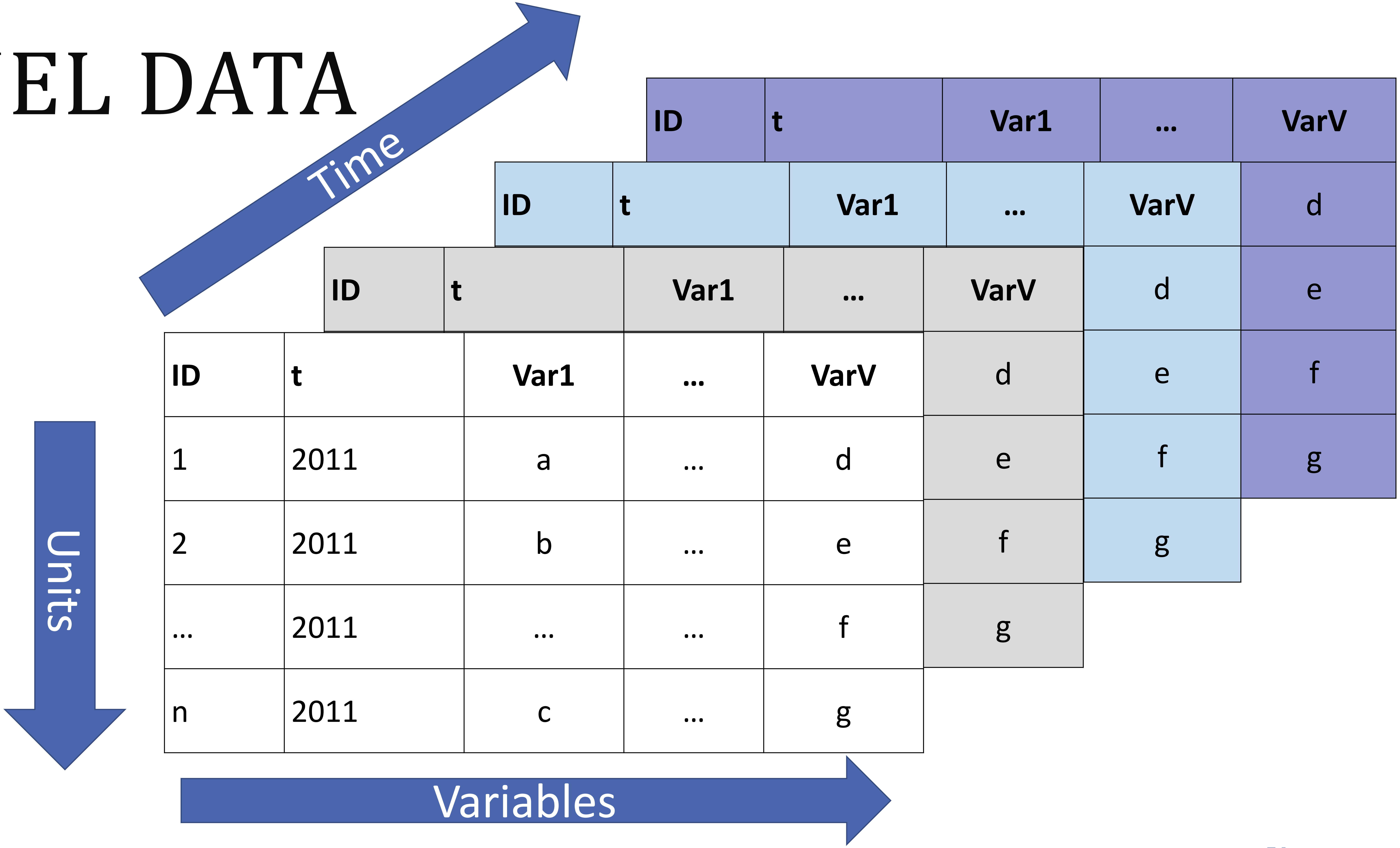
ID	t	Var1	…	VarV
1	2011	a	…	d
2	2011	b	…	e
…	2011	…	…	f
n	2011	c	…	g

ID	t	Var1	…	VarV
				d
				e
				f
				g

ID	t	Var1	…	VarV
				d
				e
				f
				g

ID	t	Var1	…	VarV
				d
				e
				f
				g

PANEL DATA


Time
ID t Var1 … VarV
ID t Var1 … VarV
ID t Var1 … VarV
ID t Var1 … VarV
1 2011 a … d
2 2011 b … e
… 2011 … … f
n 2011 c … g
d e f g
d e f g
d e f g
Units
Variables

WIDE OR LONG?

WIDE AND LONG FORMAT

- Three-dimensional panel data can be organized in a two-dimensional matrix in two ways
- Wide format
  - Rows ≙ units
  - Repeated measurements as separate variables
  - n rows and $t * v$ columns
- Long format
  - Rows ≙ single measurements
  - $n * t$ rows and v columns

WIDE FORMAT

ID	Gender	Poor_2012	Poor_2014	Poor_2016
1	0	0	0	1
2	1	1	0	0
…	…	…	…	…
999	1	1	1	1
1000	0	0	0	0

- Time dimension integrated in columns
- Variable names need to indicate time-point of measurement

LONG FORMAT

ID	Year	Poor
1	2012	0
1	2014	0
1	2016	1
…	…	…
1000	2012	0
1000	2014	0
1000	2016	0

- Time dimension integrated in rows
- Dataset needs a variables indicating time point at which information has been recorded

WIDE VS LONG FORMAT

ID	Poor_2012	Poor_2014	Poor_2016
1	0	0	1
2	1	0	0
…	…	…	…
999	1	1	1
1000	0	0	0

ID	Year	Poor
1	2012	0
1	2014	0
1	2016	1
…	…	
1000	2012	0
1000	2014	0
1000	2016	0

WIDE VS LONG FORMAT

- Most methods require long format
- Wide format better for analyzing associations of repeated measurements
- Wide format also demonstrates that measurements are not independent
- Hierarchical data structure; repeated measurements nested in units (e. g. person-years)

WIDE VS LONG IN R

- One way to wide and long transform data is provided by the `tidyr` package
  - From wide to long: `gather()`
  - From long to wide: `spread()`
- In the context of panel data, however, working with the `panelr` package is easier
  - First, declare the panel structure of the data using the `panel_data()` function, e.g.: `panel_data(pcspoverty, id = ID, wave = year)`
  - From wide to long: `long_panel()`
  - From long to wide: `widen_panel()`

LONG_PANEL()

- `long_panel(wide_data, prefix = "_", periods = c(2012, 2014, 2016), label_location = "end")`

ID	Poor_2012	Poor_2014	Poor_2016
1	0	0	1
2	1	0	0
…	…	…	…
999	1	1	1
1000	0	0	0

ID	year	Poor
1	2012	0
1	2014	0
1	2016	1
…	…	
1000	2012	0
1000	2014	0
1000	2016	0

WIDEN_PANEL()

- `widen_panel(long_data, separator = "_")`
- Both commands only work when information on the person and time identifiers was already provided with `panel_data()`

ID	year	Poor
1	2012	0
1	2014	0
1	2016	1
…	…	
1000	2012	0
1000	2014	0
1000	2016	0

ID	Poor_2012	Poor_2014	Poor_2016
1	0	0	1
2	1	0	0
…	…	…	…
999	1	1	1
1000	0	0	0

PREPARING PANEL DATA IN R

IMPORTING DIFFERENT FILE TYPES

- There are numerous ways to store data, each needs a different import function in R
- Stata's dta files: `read_dta()` (haven package)
- Excel xlsx files: `read_excel()` (readxl package)
- CSV files: `read.csv()` (base R)
- (Rdate files: `load()` (base R)
- And a lot more…

PANEL DATA MANAGEMENT

- Raw data typically provides units nested in time points
- Each new wave adds a new dataset

ID	t	Var1	…	VarV
1	2011	a	…	d
2	2011	b	…	e
…	2011	…	…	f
n	2011	c	…	g

ID	t	Var1	…	VarV
				d
				e
				f
				g

ID	t	Var1	…	VarV
				d
				e
				f
				g

ID	t	Var1	…	VarV
				d
				e
				f
				g

PANEL DATA MANAGEMENT

- Which period should be analyzed? (determine t)
- Which variables are relevant? (determine v)
- What is target population? (determine n)

→Identify which datasets provide necessary information

PANEL DATA MANAGEMENT

- Moreover, data from *one* wave may be provided in several files
- For example GSOEP: individual and household questionnaires

ID	HHID	t	Age	Gender
1	100	2011	36	0
2	101	2011	42	1
3	101	2011	40	0
4	102	2011	19	1

HHID	t	Income	Rent
100	2011	2200	900
101	2011	4100	1300
102	2011	1390	450

BINDING DATA

- Binding means combining rows (`rbind()`) or columns (`cbind()`) of two tables

ID	HHID	t	Age	Gender
1	100	2011	36	0
2	101	2011	42	1
3	101	2011	40	0

ID	HHID	t	Age	Gender
4	100	2011	8	1
5	101	2011	6	1

ID	HHID	t	Age	Gender
1	100	2011	36	0
2	101	2011	42	1
3	101	2011	40	0
4	100	2011	8	1
5	101	2011	6	1

BINDING ROWS

- Binding waves (in long format) means adding *rows* to an existing dataset → `rbind()`

ID	HHID	t	Age	Income
1	100	2011	36	2200
2	101	2011	42	3100
3	101	2011	40	1600

ID	HHID	t	Age	Income
1	100	2012	37	2400
2	101	2012	43	3100
3	101	2012	41	1900

ID	HHID	t	Age	Income
1	100	2011	36	2200
2	101	2011	42	3100
3	101	2011	40	1600
1	100	2012	37	2400
2	101	2012	43	3100
3	101	2012	41	1900

BINDING ROWS

ID	HHID	t	Age	Income
1	100	2011	36	2200
2	101	2011	42	3100
3	101	2011	40	1600

ID	HHID	t	Age	Income
1	100	2012	37	2400
2	101	2012	43	3100
3	101	2012	41	1900

ID	HHID	t	Age	Income
1	100	2011	36	2200
1	100	2012	37	2400
2	101	2011	42	3100
2	101	2012	43	3100
3	101	2011	40	1600
3	101	2012	41	1900

Sorted by ID (and t)

BINDING COLUMNS

- Binding variables means adding *columns* → `cbind()`

ID	HHID	t	Age
1	100	2011	36
2	101	2011	42
3	101	2011	40
1	100	2012	37
2	101	2012	43
3	101	2012	41

ID	HHID	t	Income
1	100	2011	2200
2	101	2011	3100
3	101	2011	1600
1	100	2012	2400
2	101	2012	3100
3	101	2012	1900

ID	HHID	t	Age	Income
1	100	2011	36	2200
2	101	2011	42	3100
3	101	2011	40	1600
1	100	2012	37	2400
2	101	2012	43	3100
3	101	2012	41	1900

BINDING DATA

- A drawback of `rbind()` is that it will only work when both tables have the same number of columns
- … and `cbind()` only when both data sets have the same number of rows
- Hence, `rbind()` will only work when both data sets have the exact same variables (as in the example)
- … and `cbind()` is useful when you have the exact same observations in two datasets (hardly the case)

JOIN()

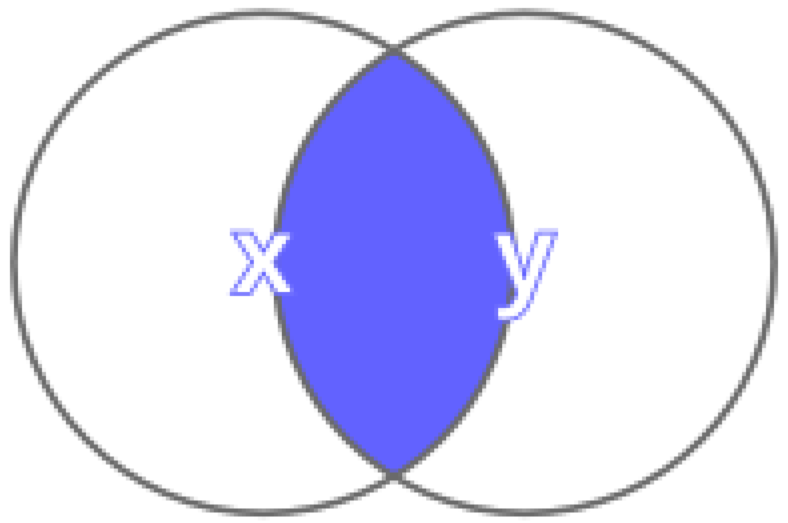
- The functions of the `join` family of the dplyr package combine two (or more) tables / data sets
- Let us call table 1 *master data*. It is the one to which we add other data (e. g.: individual-level GSOEP data)
- Table 2 should be added to data set 1, let us call it *using data* (e. g.: additional household-level GSOEP data)
- Finally, we need to know based on which column(s) we want to merge both data sets, let us call this the *key variable*
- The general syntax is: `join_type(masterData, usingData, by = keyVariable)`
- For example: `innerJoinDf <- inner_join(soep_ind, soep_hh, by = c("hid","welle"))`

DPLYER'S JOIN TYPES

- Inner Join (`inner_join()`): Combines observations of data 1 and 2 that are available in *both* data sets
- Left Join (`left_join()`): Adds data 2 to data 1
- Right Join (`right_join()`): Adds data 1 to data 2
- Full Join (`full_join()`): Combines observations of data 1 and 2 that are available in *either* data set
- Semi Join (`semi_join()`): Similar to `inner_join()`
- Anti Join (`anti_join()`): Only keeps observations of data 1 that are *not* available in data 2

INNER_JOIN()

inner_join(x, y)



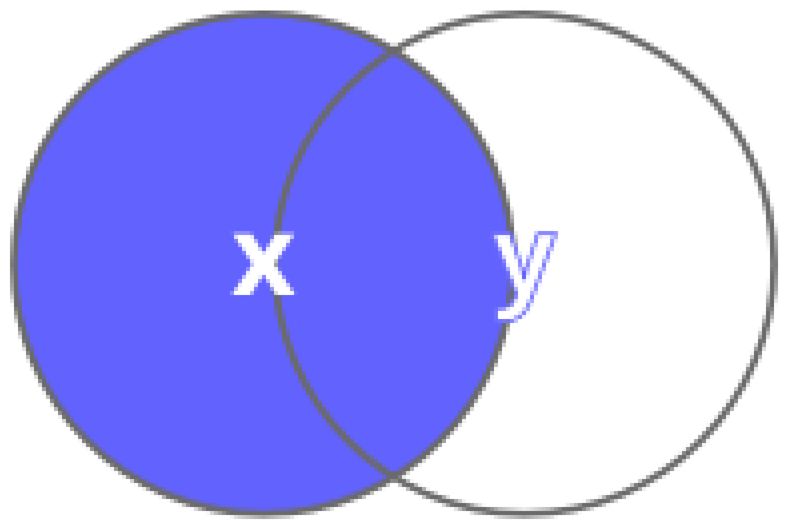
- Adds master data to using data based on key variable
- Only includes observations that exist in *both data*
- E. g.: `inner_join(master, using, by = "ID")`

ID	Age	Gender
1	36	0
2	42	1
3	23	0

ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender	Income	Rent
1	36	0	2200	900
2	42	1	4100	1300

left_join(x, y)



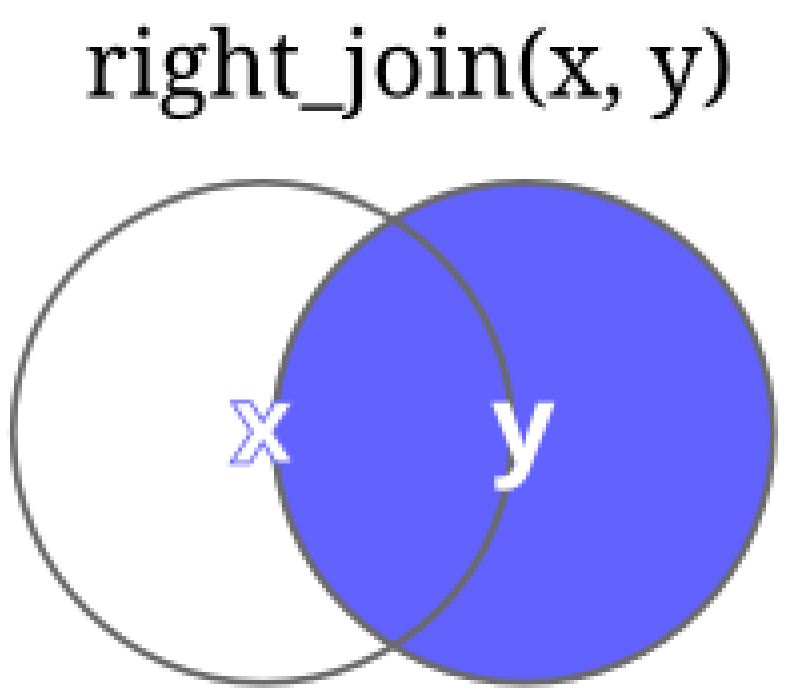
LEFT_JOIN()

- Adds using data to master data based on key variable
- Only includes observations that are included in the *master data*
- Generates *NA* if observation missing in using data
- E. g.: `left_join(master, using, by = "ID")`

ID	Age	Gender
1	36	0
2	42	1
3	23	0

ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender	Income	Rent
1	36	0	2200	900
2	42	1	4100	1300
3	23	0	*NA*	*NA*



RIGHT_JOIN()

- Adds master data to using data based on key variable
- Only includes observations that are included in the *using data*
- Generates NA if observation missing in master data
- E. g.: `right_join(master, using, by = "ID")`

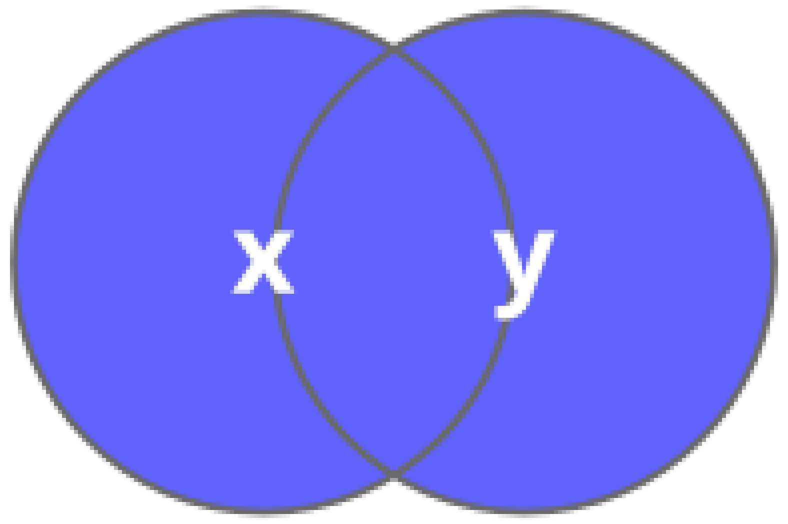
ID	Age	Gender
1	36	0
2	42	1
3	23	0

ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender	Income	Rent
1	36	0	2200	900
2	42	1	4100	1300
4	*NA*	*NA*	3600	1200

FULL_JOIN()

full_join(x, y)



- Adds master data to using data based on key variable
- Includes all observations that exist in *either data*
- E. g.: `full_join(master, using, by = "ID")`

ID	Age	Gender
1	36	0
2	42	1
3	23	0

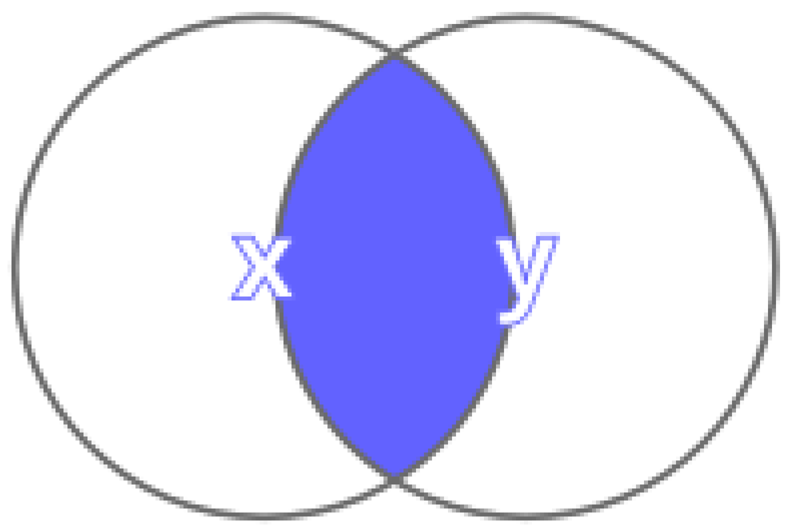
ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender	Income	Rent
1	36	0	2200	900
2	42	1	4100	1300
3	23	0	*NA*	*NA*
4	*NA*	*NA*	3600	1200

SEMI_JOIN()

- Adds master data to using data based on key variable
- Only includes observations that exist in *both data*
- … but only keeps variables that exist in the master data
- E. g.: `semi_join(master, using, by = "ID")`

semi_join(x, y)



(never duplicate rows of x)

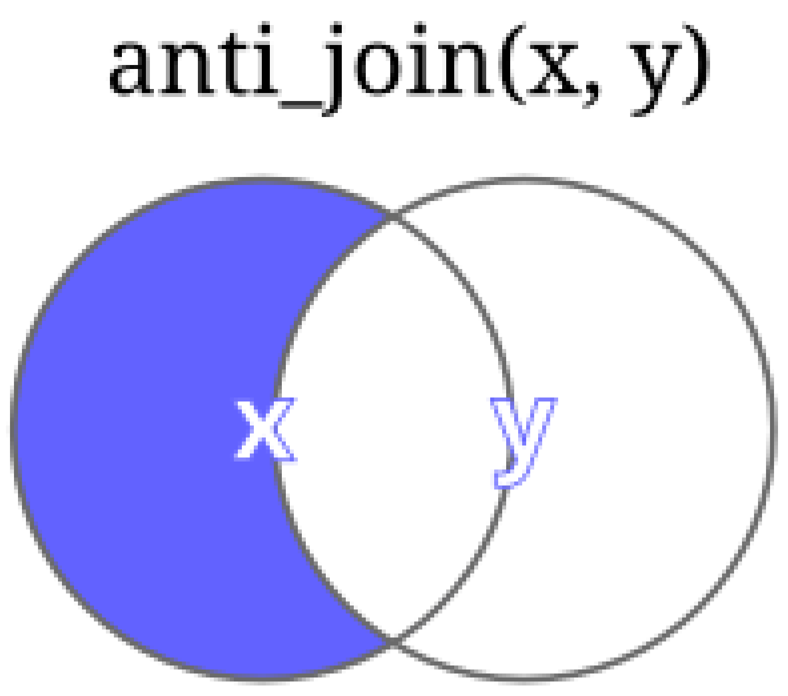
ID	Age	Gender
1	36	0
2	42	1
3	23	0

ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender
1	36	0
2	42	1

ANTI_JOIN()

- Keeps observations of the master data that do not match the using data
- Generates NA if missing in master data
- E. g.: `anti_join(master, using, by = "ID")`



ID	Age	Gender
1	36	0
2	42	1
3	23	0

ID	Income	Rent
1	2200	900
2	4100	1300
4	3600	1200

ID	Age	Gender
3	23	0

JOINING CLUSTERED DATA

- The logic of each join function also applies when we have several observations per key variable value (e. g.: multiple interviews per individual)
- In this case, each person-year in data 1 will get the (time constant) person value of the respective person in data 2

ID	Year	income
1	2021	980
1	2022	1000
2	2021	2600
2	2022	2600
3	2021	2300
3	2022	2400

ID	Birth year
1	1980
2	2002
3	1967

ID	Year	income	Birth year
1	2021	980	1980
1	2022	1000	1980
2	2021	2600	2002
2	2022	2600	2002
3	2021	2300	1967
3	2022	2400	1967

JOINING CLUSTERED DATA

- The same logic also applies for multiple members per household
- In this case, each respondent of the household in data 1 will get the household’s value in data 2

ID	HHID	age
1	100	34
2	100	57
3	101	35
4	102	64
5	102	24
6	102	36

HHID	rent
100	900
101	1300
102	1700

ID	HHID	age	rent
1	100	34	900
2	100	57	900
3	101	35	1300
4	102	64	1700
5	102	24	1700
6	102	36	1700

MULTIPLE DATA SETS OR MULTIPLE KEY VARIABLES

- More than two data sets can also easily be combined stepwise:

→
```
left_join(data1, data2, by = "id") %>%
    left_join(., data3, by = "id") %>%
    left_join(., data4, by = "id")
```

- With panel data, we will often have to combine data sets based on multiple key variables because we have variation by *person* and by *wave* (so person ID and year):

```
left_join(data1, data2, by=c("id", "year"),
match="all")
```

- Of couse, don't forget to assign these operations to an object

MULTIPLE KEY VARIABLES

- What if you want to add household-level panel data to individual-level panel data (Like the GSOEP)?

ID	HHID	t	Age	Gender
1	100	2011	36	0
1	100	2012	37	0
2	101	2011	40	1
2	101	2012	41	1
3	101	2011	37	0
3	101	2012	38	0

HHID	t	Income	Rent
100	2011	4500	1400
100	2012	4800	1400
101	2011	2200	800
101	2012	2000	820

MULTIPLE KEY VARIABLES

ID	HHID	t	Age	Gender
1	100	2011	36	0
1	100	2012	37	0
2	101	2011	40	1
2	101	2012	41	1
3	101	2011	37	0
3	101	2012	38	0

HHID	t	Income	Rent
100	2011	4500	1400
100	2012	4800	1400
101	2011	2200	800
101	2012	2000	820

ID	HHID	t	Age	Gender	Income	Rent
1	100	2011	36	0	4500	1400
1	100	2012	37	0	4800	1400
2	101	2011	40	1	2200	800
2	101	2012	41	1	2000	820
3	101	2011	37	0	2200	800
3	101	2012	38	0	2000	820

→Combination of HHID and t uniquely identifies observations

SUMMING UP

- Simple combination of data sets can be achieved with `rbind()` or `cbind()`
- However, in many instances this is not sufficient (e.g., missing data in one data set, clustered data, …)
- The `join()` family, which merges data based on key variables, helps in these cases
- This is especially relevant in the case of panel data, where we have multiple observations per unit
- I.e.: each observation (person-year) can only be identified by the (time-constant) person ID and the wave *simultaneously*
- *ALWAYS CHECK YOUR DATA AFTER COMBINING*

LAG AND LEAD VARIABLES

- Lag and lead variable relevant in *long* format
- A *lagged variable* takes at t the value of $t-1$
- In R (using dplyr):
- `data %<>%`
  - `group_by(id) %>%`
  - `mutate(var_lag = lag(var))`
  - To lag more periods: `lag(deprived, 2)` etc.
- A *lead variable* takes at t the value of $t+1$
- `data %<>%`
  - `group_by(id) %>%`
  - `mutate(var_lead = lead(var))`

LAG AND LEAD VARIABLES

- Lag and lead variables can be used for
  - Calculating transition tables by hand
  - Autoregressive/Lagged models
  - Calculating growth rates/differences over time

QUESTIONS OR COMMENTS?

EXERCISE 2: PANEL DATA MANAGEMENT

But first, 15 minutes break

EXOGENEITY ASSUMPTION FOR OLS

EXOGENEITY ASSUMPTION

- Assumption 5 means that the error term is independent from x

→Model includes all relevant variables and has correct functional form (*correctly specified*)

→Measurement error is random (does not depend on x)

- Ensures unbiased estimates
- Crucial assumption for estimating "true" (i.e. unbiased) parameters

MODEL SPECIFICATION

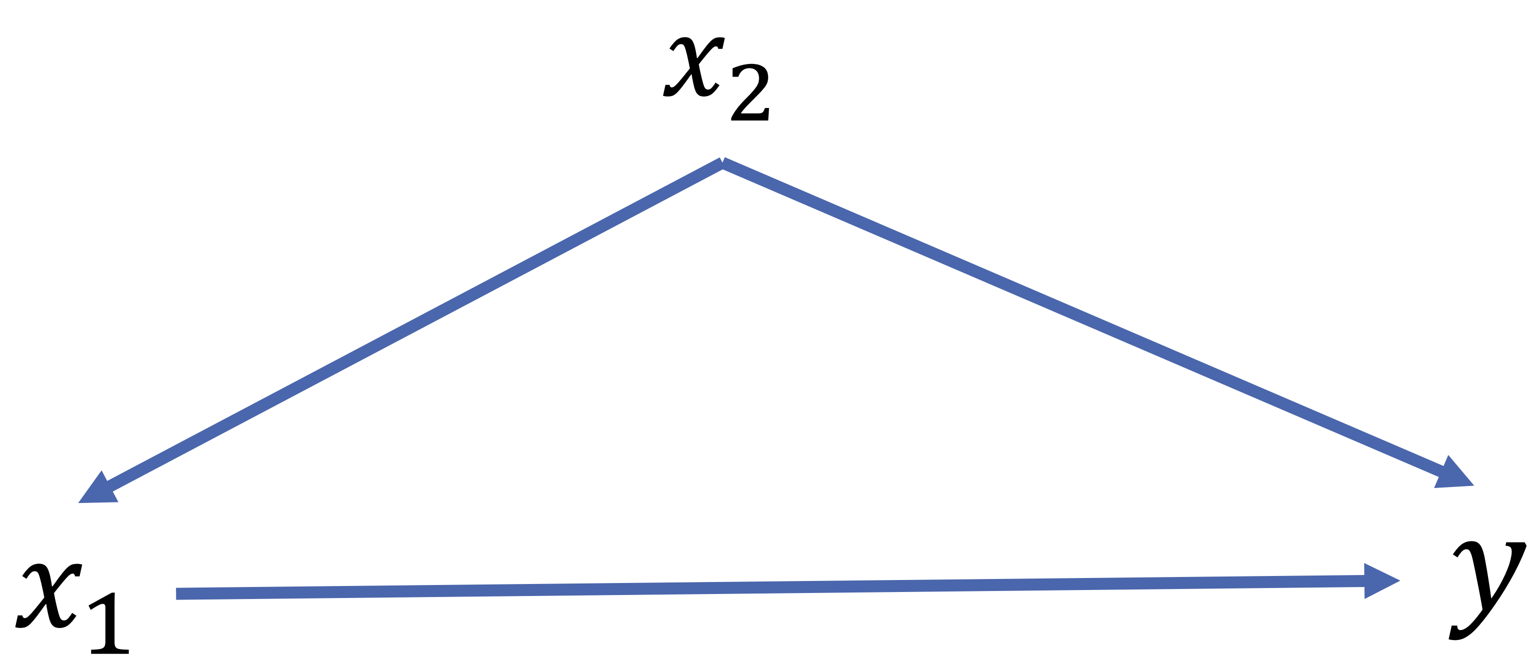
- A correctly specified model includes all relevant x
- Which x are relevant?
- Those that are conceptually or theoretically (!) cause both y and the x of interest
- Not including (omitting) relevant x_2 in a regression model will lead to a biased estimate of β_1
- This is because β_1 in this case carries part of the effect of β_2 on y
- *Avoiding bias is the main point of all statistical analyses!*

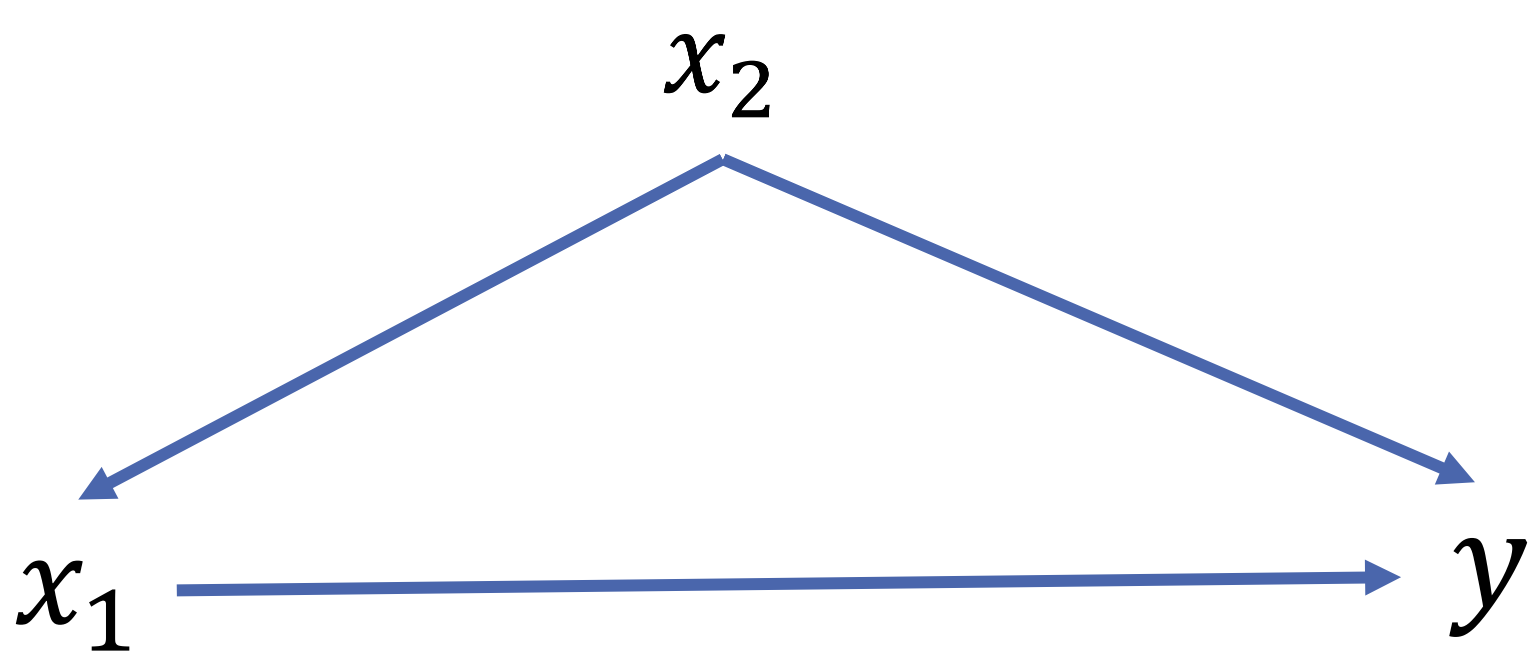
OMITTED VARIABLE BIAS

- True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
- Unbiased estimation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- New situation: x_2 unobserved
- Biased estimation: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
- Omitted variable bias: $Bias\left(\tilde{\beta}_1\right) = E\left(\tilde{\beta}_1\right) - \beta_1 = \beta_2 \frac{\widehat{cov}(x_1, x_2)}{\widehat{Var}(x_1)}$
- Hence no bias if
  - $\beta_2 = 0$
  - or $\frac{\widehat{cov}(x_1, x_2)}{\widehat{Var}(x_1)} = 0$

OMITTED VARIABLE BIAS

- $\beta_2 = 0$
- $\frac{\widehat{cov}(x_1, x_2)}{\widehat{Var}(x_1)} = 0$





LIMITS OF STATISTICAL CONTROLLING

- Within the standard linear regression framework, one can only control variables that are in the data
- Many things, however, are not observed
- Especially when working with secondary data
- Some techniques for longitudinal data analysis can tackle this problem
- Tbc.

ASSUMPTION OF UNCORRELATED ERRORS FOR OLS

PANEL DATA

- Panel data means the same individuals are observed over time (interviewed repeatedly)
- Person A is interviewed in time point 1 and in time point 2

→For each variable x, there are two data points for person A (x_{A1} and x_{A2})

→Same for person B (x_{B1} and x_{B2})

- In contrast to cross-sectional data analysis, the units of analysis are *not individuals*, but individual interviews!
- … because each individual is in the data multiple times (as often as she was interviewed)

OLS WITH PANEL DATA

- It is reasonable to assume that data points are not independent
- x_{A1} is likely to have more in common with x_{A2} than with x_{B1} (or x_{B2})
- For example, income of person A in 2015 is not independent from her income in 2014 (chances are high it's actually the same)
- Put differently, observations (interviews) cluster within individuals
- … which separates them from interviews of other individuals

→Likely a violation of the assumption of independent errors

ASSUMPTION OF INDEPENDENT ERRORS

- Violation of the assumption of independent errors means observations are not statistically independent
- Sample size is inflated
- There is less information in the data than it seems (because it is partly correlated)

→More data leads to lower standard errors (erroneously, in this case)

- Underestimated standard errors lead to wrong p-values and confidence intervals
- Results look “too significant”
- Should be modelled

SUMMARY

- OLS regression yields biased estimates if there are unobserved confounders
- OLS regression can be used with panel data if all OLS assumptions are met
- However, if there are unobserved confounders there is also very likely serial correlation (because the error contains systematic components, i.e. is not random)

PANEL DATA MODEL

- How can we use panel data if there are unobserved confounders?
- Adding an index for time: $y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \varepsilon_{it}$
- Differentiating between time-constant and time-variant variables:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \gamma_1 z_{1i} + \cdots + \gamma_l z_{li} + u_i + e_{it}$$

with
  - $i=1,\ldots,n$ units
  - $t=1,\ldots T$ observations
  - k time-varying variables x
  - l time-constant variables z
- Decomposition of the error term: $\varepsilon_{it} = u_i + e_{it}$

THE UNOBSERVED EFFECTS MODEL

- $y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \gamma_1 z_{1i} + \cdots + \gamma_k z_{li} + u_i + e_{it}$
- The error term u_i captures all time-constant unobserved characteristics of the units of analysis
- The model yields biased estimates if the error terms are correlated with the variables in the model

QUESTIONS OR COMMENTS?

Thanks for your attention!

LITERATURE

- Wickham & Grolemund (2017). R for Data Science. O'Reilly.
- Andreß, Golsch & Schmidt (2014). Applied panel data analysis for economic and social surveys. *Chapter 2 (pages 15 - 48)*. Springer Science & Business Media.
- Elwert (2013). Graphical causal models. In: Handbook of causal analysis for social research (245 - 273). Springer Science & Business Media.