

# Mini-Workshop Panel Data Analysis

Marko Bachl (mit Material von Michael Scharkow)

Sommersemester 2020 | IJK Hannover



# Contents

<b>1</b>	<b>Überblick</b>	<b>5</b>
1.1	Inhalt des virtuellen Mini-Workshops . . . . .	5
1.2	Welche Inhalte wir nicht behandeln . . . . .	6
1.3	Aufbau des Workshops . . . . .	6
<b>2</b>	<b>Einführung</b>	<b>9</b>
2.1	Längsschnittdaten . . . . .	9
2.2	Beispiel-Daten . . . . .	11
2.3	Pooled OLS (WRONG!) . . . . .	11
2.4	Warum ist Pooled OLS immer falsch? Statistische Theorie . . . .	11
2.5	Warum ist pooled OLS immer falsch? Inhaltliche Überlegungen .	12
<b>3</b>	<b>Fixed effects models</b>	<b>13</b>
<b>4</b>	<b>Random effects models</b>	<b>15</b>
<b>5</b>	<b>Within-between models</b>	<b>17</b>



# Chapter 1

## Überblick

### 1.1 Inhalt des virtuellen Mini-Workshops

- Der Mini-Workshop bietet eine *pragmatische* Einführung in die Analyse von Panel-Daten aus Erhebungen mit mindestens drei Wellen. Konkret liegt der Fokus auf sogenannten *micro panels*, also Datensätzen mit relativ vielen Fällen und relativ wenigen Messzeitpunkten (das klassische Befragungspanel).
- In der Analyse beschränken uns hier auf Varianten der *linearen* Regressionsmodelle. Wir beginnen mit den grundlegenden *fixed effects* und *random effects* Modellen. Dann betrachten wir das *within-between* Modell, das als eine Integration des *fixed effects* Modell in das *random effects* Modell verstanden werden kann. Dies ist auch eine gute Grundlage für den Einstieg in verschiedene Erweiterungen, zum Beispiel zu verallgemeinerten linearen Modellen oder zu Wachstumskurvenmodellen. Diese sind aber nicht Teil dieses Mini-Workshops.
- Wir schätzen die Modelle mit etablierten *least-squares* und *maximum likelihood* Methoden. Gerade bei den *within-between* Modellen sind bayesianische Schätzmethoden, z.B. *MCMC sampling* (implementiert in Stan), unabhängig von den statistisch-philosophischen, sehr interessant. Bei Interesse kann ich nur empfehlen, hier einen Einstieg zu finden.
- Zur Aufbereitung der Daten, Visualisierung und Modell-Schätzung verwenden wir R mit dem *tidyverse* und eine kleine Zahl spezialisierter Pakete für die Modellschätzung. Der Fokus des Workshops liegt aber auf der substantiellen Arbeit mit den Modellen, nicht auf der Umsetzung in R.

## 1.2 Welche Inhalte wir nicht behandeln

- Der Workshop ist kein Statistik- oder Ökonometrie-Kurs. Ich bin — wie auch ihr — ausgebildeter Sozialwissenschaftler. Die statistischen Grundlagen, auf denen der Workshop aufbaut, geht aus den Grundlagentexten (Bell and Jones, 2015; Vaisey and Miles, 2017) hervor.
- Grundkenntnisse in R setze ich voraus, insbesondere Datentransformationen innerhalb des `tidyverse`.
- Wir werden nicht viel Zeit auf die verschiedenen Schätzer, deren Effizienz und Bias, die verschiedenen Algorithmen und Datentransformationen verwenden.
- Wir werden keine Beweise oder Ableitungen besprechen. Wir setzen keine Kenntnisse in Matrixalgebra voraus — weder meiner- noch eurerseits.
- Wir behandeln einen sehr kleinen Ausschnitt möglicher Modelle für Panel-Daten. Der konzentrieren uns auf regressionsbasierte Modelle zur Schätzung kausaler Effekte. Damit behandeln wir insbesondere nicht die vielfältigen Verfahren, die in einem SEM-Framework verortet sind: längsschnittliche Messmodelle, Prozessmodelle, (random intercept) cross-lagged panel Modelle, Latent State-Trait Modelle, etc.
- Fehlende Daten (Panelmortalität, Ausfall von Einheiten in einzelnen Wellen) sind ein großes Thema in der Längsschnittanalyse. Wir werden es hier ignorieren, bis auf den Hinweis, dass alle Fälle, die in mindestens zwei bzw. drei Wellen Daten haben, grundsätzlich Informationen zur Schätzung beitragen.

## 1.3 Aufbau des Workshops

- Inhaltlicher Aufbau: Siehe Kapitel-Gliederung

### Material

- Dieses Dokument + R Skripte: (Hoffentlich) mehr oder weniger selbsterklärendes Material
  - Kuratierte Form ist dieses HTML-Dokument
  - Es gibt auch ein PDF, das ich aber nicht formatiert habe
- Screencast: Ich gehe über das Material und erkläre es auf der Audio-Spur. Mal sehen, wie hilfreich das ist. Die Screencasts stelle ich über das LMS zur Verfügung.
- Übungen: Zu einigen Analysen gibt es Übungsaufgaben.

- Bei der *Wiederholung* geht es darum, die Modelle leicht zu verändern (durch Anpassen der R-Skripte aus dem Material) und die Ergebnisse der angepassten Modelle zu interpretieren.
- Bei der *Anwendung* geht es darum, in Anlehnung an die Beispiele eigene Modelle zu spezifizieren und diese zu interpretieren.

## Pakete

Wir verwenden die folgenden Pakete

```
if (!require("pacman")) install.packages("pacman")
pacman::p_load(tidyverse)

tibble(package = pacman::p_loaded(all = TRUE)) %>% mutate(version = map_chr(package,
  ~as.character(pacman::p_version(package = .x)))) %>% arrange(package) %>% knitr::kable()
```

package	version
base	3.6.2
datasets	3.6.2
dplyr	0.8.4
forcats	0.4.0
ggplot2	3.2.1
graphics	3.6.2
grDevices	3.6.2
methods	3.6.2
pacman	0.5.1
purrr	0.3.3
readr	1.3.1
stats	3.6.2
stringr	1.4.0
tibble	2.1.3
tidyr	1.0.2
tidyverse	1.3.0
utils	3.6.2





## Chapter 2

# Einführung

### 2.1 Längsschnittdaten

#### Begriffe

- Wiederholte Querschnittserhebungen (time series cross sectional, TSCS):  $n$  unabhängige Fälle (repräsentativ für dieselbe Grundgesamtheit) zu mehreren Messzeitpunkten  $t$ .
- Zeitreihe: Eine Einheit mit vielen Messzeitpunkten ( $n = 1, t > 30$ ).
- Paneldaten: Dieselben Einheiten mit wiederholten Messungen ( $n > 30, t \geq 2$ )
  - Macro panel:  $n$  klein,  $t$  groß (z.B. jährliche Untersuchung von Staaten, 1950–2015)
  - Micro panel  $n$  groß,  $t$  klein (typisches Befragungspanel)
- In diesem Workshop geht es um *micro panels* mit  $t > 2$

#### Vorteile von Paneldaten

- Paneldaten erlauben die Identifikation von kausalen Effekten unter schwächeren Annahmen (im Vergleich zu Querschnittsdaten).
  - Wir haben einige (aber nicht perfekte!) Informationen über die zeitliche Abfolge von Veränderungen.
  - Wir können untersuchen, ob, und wenn ja, wie ein Ereignis (eine Veränderung eines Prädiktors) das Kriterium verändert.
- Paneldaten erlauben die Untersuchung von individuellen Verläufen

#### Kausale Effekte mit Paneldaten schätzen

##### Bedingungen

1. Kovariation zwischen  $X$  und  $Y$  (bivariate Korrelation  $r_{XY}$ )

2.  $X$  muss logischer vor  $Y$  liegen
3. Keine (nicht beobachteten) Störvariablen (kein  $Z$  mit kausalem Effekt auf  $X$  und  $Y$ )

### Herausforderungen (auch bzw. gerade mit Paneldaten)

- Entsprechung der zeitlichen Entfaltung des Effekts und des Designs (Abstände, Verläufe)
- Reliabilität und Konstruktstabilität
  - Reliabilität: Bei geringer Reliabilität beobachten wir Veränderungen, die aber auf Rauschen in der Messung zurückgehen.
  - Reliabilität: Wenn die Messungen über die Zeit ihre Bedeutung verändern, modellieren wir keine Veränderung des latenten Konstrukts von Interesse.
- Panelmortalität und Paneleffekte
  - Panelmortalität: Einheiten (Befragte) fallen aus, möglicherweise systematisch mit Bezug auf die Konstrukte von Interesse
  - Paneleffekte: Einheiten (Befragte) verändern sich durch die Messung (z.B. Lernen von Wissensfragen, Anregung durch Fragen zu Medienangeboten)

### Format von Datensätzen mit Paneldaten

Long Format			Wide Format				
$i$	$t$	$y$	$i$	$y_{t1}$	$y_{t2}$	$y_{t3}$	$y_{t4}$
1	1	6.55	1	6.55	6.68	6.77	7.04
1	2	6.68	2	5.55	6.01	6.32	6.40
1	3	6.77	3	4.65	5.33	6.45	6.45
2	1	5.55	...				
2	2	6.01					
...							

Figure 2.1:  $i$  indiziert Einheiten,  $t$  indiziert Messzeitpunkte,  $y$  ist eine Variable

- Die Modelle in diesem Workshop nutzen das *long format*
- Datensätze können von einem ins andere Format transformiert werden, z.B. im `tidyverse`:
  - `tidyr::gather()` und `tidyr::spread()` oder
  - `tidyr::pivot_longer()` und `tidyr::pivot_wider()`

## 2.2 Beispiel-Daten

- sponsored by Jule Scheper und Sophie Bruns
- KURZE INHALTLICHE BESCHREIBUNG
  - Erhebungszeitraum, Messzeitpunkte
  - Repeated Measures
  - Constant Measures
- Kurzer Auszug mit `summary` und `print`

```
# source('R/packages.R')
d = tibble(x = rnorm(10), y = rnorm(10))
```

## 2.3 Pooled OLS (WRONG!)

- Untersuchung des Effekts  $X \rightarrow Y$
- Einfachstes Modell: Regression  $X \rightarrow Y$

```
lm(y ~ x, data = d) %>% summary(show.resid = F)

##
## Call:
## lm(formula = y ~ x, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25418 -0.37140  0.07569  0.42561  1.00785
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52744    0.23979   2.200   0.059 .
## x            0.02312    0.17617   0.131   0.899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6989 on 8 degrees of freedom
## Multiple R-squared:  0.002148,    Adjusted R-squared:  -0.1226
## F-statistic: 0.01722 on 1 and 8 DF,  p-value: 0.8988
```

## 2.4 Warum ist Pooled OLS immer falsch?

### Statistische Theorie

- 1) Exogenitätsannahme ist verletzt,  $E(u_i|x_i) \neq 0$ , da

- Korrelationen zwischen den Variablen  $x$  gehen auf nicht gemessene Eigenschaften der Einheiten zurück, z.B. Eigenschaften der Person  $z_i$ , die sowohl  $x_i$  als auch  $y_i$  beeinflussen.
  - Auch bekannt als *omitted variable bias*
  - Könnte behoben werden, wenn alle  $z_i$  im Modell wären; diese Idee wird später wichtig
- 2) Annahmen Homoskedastizität und unkorrelierte Residuen sind (wahrscheinlich) verletzt
- Systematische Variation der Residuen zwischen Einheiten
  - Wahrscheinlich serielle Korrelationen durch die zeitliche Abhängigkeit der Messungen
- 3) Annahme der Unabhängigkeit der Beobachtungen verletzt
- Überschätzung der Information von abhängigen Fällen (dieselbe Information ist mehrmals im Datensatz)
    - Zu kleine Standardfehler, zu große Zahl der Freiheitsgrade in Signifikanz-Tests
  - Die wahre Fallzahl (effective sample size) ist kleiner als Zahl der Zeilen im Datensatz (*long format*)

## 2.5 Warum ist pooled OLS immer falsch? Inhaltliche Überlegungen

- Unser Ziel ist es, den wahren kausalen Effekt von  $X$  auf  $Y$  zu schätzen.
- Pooled OLS vermischt aber zwei Quellen von Unterschieden in den Daten: Den (kausalen) Effekt innerhalb der Personen (within) und die Unterschiede zwischen Personen (between).
- Within und between Effekte können sich in Größe und sogar in der Richtung unterscheiden!
- Die Schätzung aus einem pooled OLS Modell vermischt den kausalen Effekt und die interindividuellen Unterschiede.
- In der Sprache von Interventionsstudien ist das ein Selbstselektions-Problem: Was passiert, wenn Personen, die vor dem Treatment  $x$  schon höhere Werte in  $y$  haben? *X Evtl. anpassen an Datenbeispiel X*
- Außerdem fällt auf, dass im einfachen OLS Modell nichts darauf hindeutet, dass es sich um Paneldaten handelt. Selbst wenn wir die genannten Probleme nicht hätten, hätten wir auch nichts durch die Paneldaten gewonnen.

## Chapter 3

# Fixed effects models

xxx



## Chapter 4

# Random effects models

xxx





## Chapter 5

# Within-between models

xxx



# Bibliography

Bell, A. and Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1):133–153.

Vaisey, S. and Miles, A. (2017). What you can—and can’t—do with three-wave panel data. *Sociological Methods & Research*, 46(1):44–67.