

# 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!) . . . . .	13
<b>3</b>	<b>Fixed effects Modelle</b>	<b>17</b>
3.1	Konzeptionelle Einführung . . . . .	17
3.2	Übungsaufgaben 1 . . . . .	22
3.3	<i>Fixed effects</i> Modelle in der praktischen Anwendung . . . . .	22
3.4	Conclusio: Vor- und Nachteile des <i>fixed effects</i> Modells . . . . .	29
3.5	Übungsaufgaben 2 . . . . .	30
<b>4</b>	<b>Random effects models</b>	<b>31</b>
<b>5</b>	<b>Within-between models</b>	<b>33</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 statistisch-philosophischen Überlegungen 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, gehen 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 aber keine komplizierten Dinge in R tun. Auch ohne weiterführende R-Kenntnisse sollten die Inhalte des Workshops in Bezug auf die datenanalytischen Verfahren klar werden.
- 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. Auch Modelle, in denen die Zeit-Variable als kontinuierlich (z.B. Tag der Erhebung im Gegensatz zu Indikator für Panelwelle) verwendet wird (z.B. Continuous Time Structural Equation Modeling), behandeln wir nicht.
- 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, broom, haven, plm, lmtest)
theme_set(theme_bw()) # ggplot theme

tibble(package = c("R", sort(pacman::p_loaded())) %>% mutate(version = map_chr(package,
  ~as.character(pacman::p_version(package = .x)))) %>% knitr::kable()
```

package	version
R	3.6.2
broom	0.5.4
dplyr	0.8.4
forcats	0.4.0
ggplot2	3.2.1
haven	2.2.0
lmtest	0.9.37
pacman	0.5.1
plm	2.2.3
purrr	0.3.3
readr	1.3.1
stringr	1.4.0
tibble	2.1.3
tidyr	1.0.2
tidyverse	1.3.0
zoo	1.8.7





## 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 logisch 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.
  - Konstruktstabilität: Wenn die Messungen über die Zeit ihre Bedeutung verändern, modellieren wir keine Veränderung des latenten Konstrukts von Interesse.
- Panelmortalität und Paneffekte
  - Panelmortalität: Einheiten (Befragte) fallen aus, möglicherweise systematisch mit Bezug auf die Konstrukte oder Effekte, die uns interessieren.
  - Paneffekte: 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()` (verwende ich in `R/data.R`) oder
  - `tidyr::pivot_longer()` und `tidyr::pivot_wider()`

## 2.2 Beispiel-Daten

- Titel: Soziale Normen im alltäglichen Umgang mit den Konsequenzen der Corona-Krise
- sponsored by Jule Scheper und Sophie Bruns
- Thema der Erhebung: Die Corona-Pandemie hat Regierungen auf der ganzen Welt dazu veranlasst, Regelungen zur Reduzierung der raschen Ausbreitung des Virus einzuführen. Die deutsche Bundesregierung hat am 22. März 2020 mehrere Maßnahmen zur Einschränkung sozialer Kontakte beschlossen. Diese Einschränkungen im sozialen Leben sind vollkommen neu und jede\*r Einzelne muss sich auf diese Regelungen und die neue Lebenssituation einstellen. Diese Studie beschäftigt sich mit der Frage, wie Menschen sich im Alltag mit der Corona-Pandemie beschäftigen und wie sie mit den Regelungen zur Beschränkung sozialer Kontakte umgehen. Im Mittelpunkt der Untersuchung steht die Entstehung und Veränderung von sozialen Normen und persönlichen Einstellungen zur Beschränkung sozialer Kontakte über die Zeit.
- Im Rahmen des Workshops steht der Einfluss der sozialen Normen und der eigenen Einstellung zum Verhalten auf das tatsächliche Social Distancing-Verhalten im Mittelpunkt.
- Zeitraum der Erhebung: 1.4.-28.4.2020
- Datum der Messzeitpunkte: Die Befragung besteht aus vier Wellen. Jede Welle war für eine Woche im Feld und bezog sich immer auf die vorherige Kalenderwoche.
  - Welle 1: Erhebungszeitraum vom 1.4.-7.4., Bezugszeitraum vom 23.3. bis 29.4.
  - Welle 2: Erhebungszeitraum vom 8.4.-14.4., Bezugszeitraum vom 30.3. bis 5.4.
  - Welle 3: Erhebungszeitraum vom 15.4.-21.4., Bezugszeitraum vom 6.4. bis 12.4.
  - Welle 4: Erhebungszeitraum vom 22.4.-28.4., Bezugszeitraum vom 13.4. bis 19.4.
- Nachvollziehen der Aufbereitung in `R/data.R`
- Direkt laden (z.B. für Übungen) aus `R/data/data.rds`

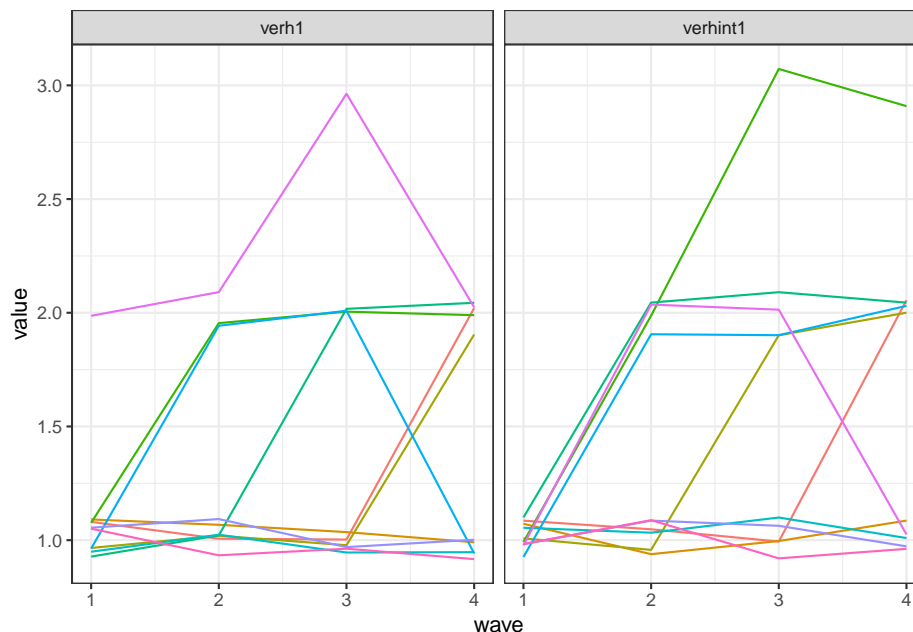
Variablenname	Label
alter	Alter
besorg1	Ich bin besorgt, wenn ich an Corona denke.
bildung	Bildungsabschluss
desnormp1	...sind in der letzten Woche rausgegangen, auch wenn es sich nicht um einen Arztbesuch handelt.
desnormp2	...haben sich in der letzten Woche in ihrer Freizeit mit mehr als einer anderen Person getroffen, die nicht in meinem Haushalt lebt.
desnormp3	...haben in der letzten Woche weniger als 1,5 Meter Abstand zu Personen gehalten, die nicht in meinem Haushalt leben.
desnormp4	...haben sich in der letzten Woche strikt an die Maßnahmen zur Beschränkung sozialer Kontakte gehalten.
ein1	Ich finde es in Ordnung, wenn man rausgeht, auch wenn es sich nicht um einen Arztbesuch handelt.
ein2	Ich finde es in Ordnung, wenn man sich in seiner Freizeit mit mehr als einer anderen Person trifft.
ein3	Ich finde es in Ordnung, wenn man weniger als 1,5 Meter Abstand zu Personen hält, die nicht in meinem Haushalt leben.
ein4	Ich finde es wichtig, dass die Empfehlung zur Beschränkung sozialer Kontakte strikt eingehalten wird.
ein5	Ich finde es richtig, dass generell Abstand gehalten werden soll.
injnorp1	...finden es in Ordnung, wenn man rausgeht, auch wenn es sich nicht um einen Arztbesuch handelt.
injnorp2	... finden es in Ordnung, wenn man sich in seiner Freizeit mit mehr als einer anderen Person trifft.
injnorp3	...finden es in Ordnung, wenn man weniger als 1,5 Meter Abstand zu Personen hält, die nicht in meinem Haushalt leben.
injnorp4	...finden es in Ordnung, wenn man sich strikt an die Maßnahmen zur Beschränkung sozialer Kontakte hält.
kompeer_s1	Freunde
kompeer_s2	Familie und Partner oder Partnerin
kompeer_s3	Bekannte (z.B. Arbeitskollegen und -kolleginnen, Vereinsmitglieder)
kompeer_s4	Prominente und/oder Influencer
med1	Zeitungen & Zeitschriften (z.B. Die ZEIT, Bild, Focus, der Spiegel)
med2	Öffentlich-rechtliche Fernsehsender (z.B. ARD, ZDF, h1)
med3	Private Fernsehsender (z.B. RTL, ProSieben)
med4	Öffentlich-Rechtliche Radiosender (z.B. DLF, n-joy, NDR)
med5	Private Radiosender (z.B. 89.0 RTL, fn)
sex	Geschlecht W4 dummy
stress	Ich fühle mich durch die Corona-Pandemie gestresst.
verh1	Ich bin rausgegangen, auch wenn es sich nicht um einen Arztbesuch, Arbeitsweg, Einkauf, Spaziergang handelt.
verh2	Ich habe mich mit mehr als einer Person getroffen, die nicht in meinem Haushalt lebt.
verh3	Ich habe weniger 1,5 Meter Abstand zu Personen gehalten, die nicht in meinem Haushalt leben.
verh4	Ich habe mich strikt an die Maßnahmen zur Beschränkung sozialer Kontakte gehalten.
verh5	Ich habe mich im Privaten mit Freunden oder Familienmitgliedern getroffen, die nicht in meinem Haushalt leben.
verh6	Ich war länger draußen als für einen üblichen Spaziergang (z.B. saß auf der Wiese oder im Park).
verhint1	Rausgehen, auch wenn es sich nicht um einen Arztbesuch, Arbeitsweg, Einkauf, Spaziergang handelt.
verhint2	Mich mit mehr als einer Person treffen, die nicht in meinem Haushalt lebt.
verhint3	Weniger als 1,5 Meter Abstand zu Personen halten, die nicht in meinem Haushalt leben.
verhint4	Mich strikt an die Maßnahmen zur Beschränkung sozialer Kontakte halten.
verhint5	Mich im Privaten mit Freunden oder Familienmitgliedern treffen, die nicht in meinem Haushalt leben.
verhint6	Mich länger draußen aufhalten als für einen üblichen Spaziergang (z.B. auf der Wiese oder im Park).
veruns	Ich bin verunsichert durch die Corona-Krise.

## 2.3 Pooled OLS (WRONG!)

- Als erstes Beispiel wollen wir uns einer klassischen Frage aus der Theory of Planned Behavior zuwenden. Wir interessieren uns für den Effekt der Verhaltensintention auf das (berichtete) Verhalten (schließlich würden wir zum Start des Workshops ja gerne etwas finden ;)). Konkret betrachten wir den Effekt des Vorhabens, entgegen der Empfehlungen ohne relevanten Grund die Wohnung zu verlassen, auf den Selbstbericht, dies auch zu tun. Die beiden relevanten Variablen sind `verh1` und `verhint1`. Die Abbildung zeigt ihre Entwicklung über die vier Wellen für 10 zufällig ausgewählte Personen.

```
id_smple = sample(unique(d$IDSosci), 10)

d %>% filter(IDSosci %in% id_smple) %>% select(IDSosci, wave, verh1, verhint1) %>%
  gather(variable, value, -IDSosci, -wave) %>% ggplot(aes(wave, value, group = IDSosci,
    color = IDSosci)) + geom_line(position = position_jitter(height = 0.1, width = 0),
    show.legend = FALSE) + facet_wrap("variable")
```



- Das einfachste Modell, diesen Effekt zu schätzen, ist eine einfache OLS Regression der Verhaltensintention auf das Verhalten.

```
lm(verh1 ~ verhint1, data = d) %>% tidy() %>% mutate_if(is.numeric, round, 2)

## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>    <dbl>    <dbl>   <dbl>
```

## 1 (Intercept)	0.46	0.02	19.2	0
## 2 verhint1	0.59	0.01	53.8	0

- Das Modell besagt, dass die Häufigkeit, ohne triftigen Grund raus zu gehen, mit jedem Punkt auf der Intentionsskala um ca.  $b_{verhint1} = 0.6$  Punkte steigt.

### Warum ist Pooled OLS immer falsch? Statistische Theorie

- Wir nennen dieses Modell *pooled* OLS, da alle Beobachtungen einfach zusammengeworfen werden, ohne zu beachten, dass einige von ihnen zusammen gehören, da sie von denselben Personen stammen.
- 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*)

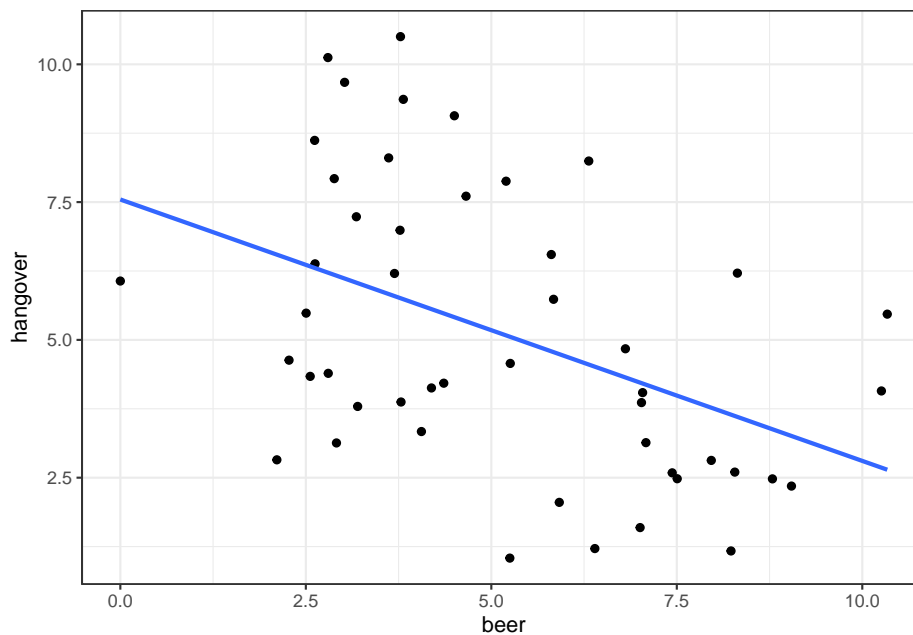
### Warum ist pooled OLS immer falsch? Inhaltliche Überlegungen

- 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?

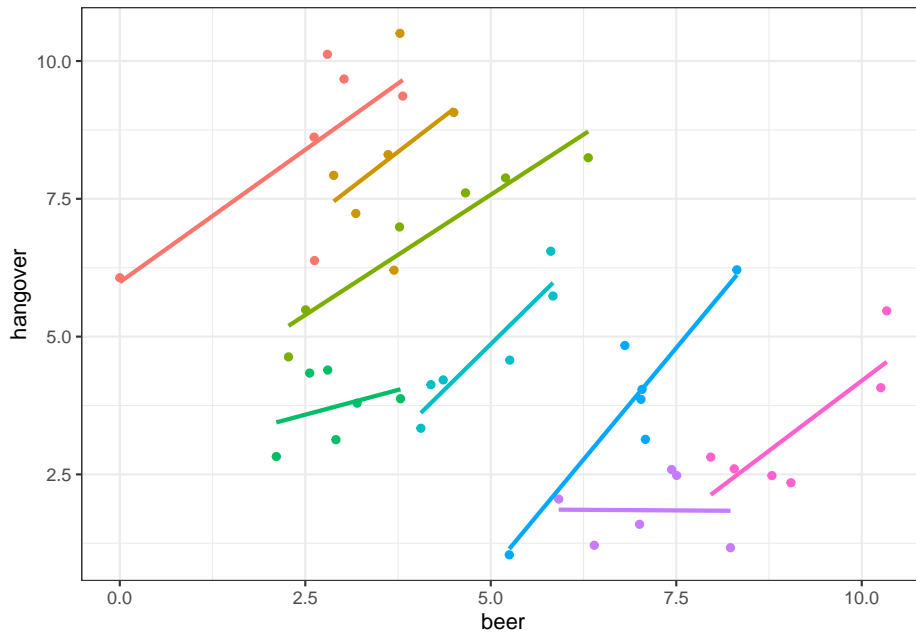
- Außerdem fällt auf, dass im einfachen OLS Modell nichts darauf hindeutet, dass es sich um Paneldaten handelt. Selbst wenn wir die genannten Probleme nicht hätten, hätten wir auch nichts durch die Paneldaten gewonnen.

### Pooled OLS, within und between - eine Illustration

- Zum Abschluss noch ein imaginäres Beispiel, um den Unterschied von intraindividuellen (within) Effekten und interindividuellen Unterschieden zu verdeutlichen. Wir führen eine Panel-Studie mit acht Personen und sechs Messzeitpunkten zum Zusammenhang von Bier-Konsum und Hangover durch. Wir interessieren uns für die kausale Frage, ob mehr Bier zu einem schlimmeren Kater führt.
- In der pooled OLS Analyse wird einfach die Regressionsgerade durch alle Beobachtung gelegt. Es zeigt sich ein negativer Zusammenhang. Je mehr Bier konsumiert wurde, desto schwächer fällt der Hangover aus.



- Wenn wir aber für alle acht Personen separat den Zusammenhang zwischen Bierkonsum und Kater berechnen (so genanntes no pooling Modell), ergibt sich ein anderes Bild. Für alle Personen gilt mehr oder weniger deutlich: Je mehr Bier konsumiert wurde, desto stärker fällt der Hangover aus (within).



- Dazu kommt ein systematischer Unterschied zwischen den Personen (between): Personen, die im Durchschnitt mehr Bier trinken, haben im Durchschnitt einen schwächeren Hangover. Dies könnte auf eine nicht beobachtete Drittvariable auf Ebene der Personen zurück gehen:
  - Vielleicht trinken Personen, die wissen, dass sie nicht so anfällig für einen Hangover sind, mehr, während Personen, die immer einen starken Kater haben, schon aus Angst vor dem nächsten Tag weniger trinken.
  - Oder es ist ein Gewöhnungseffekt: Personen, die häufig viel trinken, gewöhnen sich an den Kater und nehmen ihn als weniger schlimm wahr. Oder mit Lemmy: “A kid once said to me “Do you get hangovers?” I said, “To get hangovers you have to stop drinking.”
- Mit den vorliegenden Daten können wir die Frage nach dem Prozess nicht beantworten, da wir die Drittvariable nicht gemessen haben. Wir können aber *alle* Variablen kontrollieren, die auf Personenebene liegen, z.B., indem wir wie in der Abbildung für jede Person ein separates Modell schätzen. Dann können Unterschiede zwischen den Einheiten per Modelldefinition keinen Einfluss auf die Schätzung haben. Etwas ähnliches passiert im *fixed effects* Modell, das wir im nächsten Abschnitt besprechen.



## Chapter 3

# Fixed effects Modelle

### 3.1 Konzeptionelle Einführung

- Im ersten Teil des Abschnitts zu *fixed effects* Modellen beschäftigen wir uns mit den Grundlagen der Modellierung. Dazu nutzen wir `stats::lm()` (übliche OLS-Schätzung linearer Modelle in R).

#### Wie können wir den kausalen (within-person) Effekt mit Paneldaten schätzen?

- 1) Separate OLS Modelle für jede Person schätzen und Koeffizienten mitteln (no pooling).
  - 2) Alle  $X$  und  $Y$  Variablen um die Mittelwerte der Person zentrieren (within transformation).
  - 3) Dummy-Variablen für jede Person in das Regressionsmodell aufnehmen (least squares dummy variables [LSDV] estimation).
- Alle drei Varianten entfernen die (beobachteten und nicht beobachteten,) über die Zeit konstanten Unterschiede zwischen den Personen.
  - Varianten 2 und 3 entsprechen dem klassischen *fixed effects* Modell. Die Unterschiede zwischen den Personen werden kontrolliert, indem die personenspezifischen Mittelwerte vor der Schätzung entfernt werden (2) oder für jede Person im Modell geschätzt werden (3).
    - $y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + (u_{it} - \bar{u}_i)$  oder  $y_{it} = \beta' x'_{it} + \alpha_i + u_{it}$
  - In Variante 1 dürfen die kausalen within-person Effekte zwischen den Personen variieren. Unter der Annahme homogener Treatment-Effekte (entspricht der typischen Annahme im randomisierten Between-Subject-Experiment) entspricht das Ergebnis asymptotisch den Varianten 2 und 3.
    - Der Schätzer ist aber weniger effizient, da zufällige Unterschiede in

den Effekten zwischen den Personen aufgegriffen werden.

- Im letzten Teil des Abschnitts zum within-between-Modell kommen wir auf diesen Punkt zurück, wenn wir die Annahme homogener Treatment-Effekte lockern.

## No pooling

```
d %>% group_by(IDsosci) %>% # mutate(chk = sd(verh1) != 0 & sd(verhint1) != 0) %>% filter(
  nest() %>% mutate(mdls = map(data, ~tidy(lm(verh1 ~ verhint1, data = .x)))) %>% unnest(
  ungroup() %>% select(-data) %>% na.omit() %>% filter(statistic != Inf) %>% filter(
    "verhint1") %>% mutate_if(is.numeric, round, 2) %>% print %>% summarise(estimate =
    std.error = sqrt(mean(std.error^2))) # simple approximation
```

```
## # A tibble: 232 x 6
##   IDsosci term      estimate std.error statistic p.value
##   <chr>   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 050IPY verhint1    1.25     0.56  2.24e+ 0    0.15
## 2 05J4R8 verhint1    0.45     0.18  2.50e+ 0    0.13
## 3 08BDZJ verhint1    0.33     0.53  6.30e- 1    0.59
## 4 0E09L2 verhint1    1.67     0.67  2.50e+ 0    0.13
## 5 0F5L9Z verhint1    0        0.71  0.        1
## 6 0KYAJ verhint1    0.45     0.18  2.50e+ 0    0.13
## 7 0ONV40 verhint1    1        0      9.01e+15    0
## 8 0ZCKB5 verhint1   -0.35     0.5   -6.90e- 1    0.56
## 9 114OWA verhint1    0.33     0.33  1.00e+ 0    0.42
## 10 16YGN0 verhint1    0.5      0.25  2.00e+ 0    0.18
## # ... with 222 more rows

## # A tibble: 1 x 2
##   estimate std.error
##   <dbl>    <dbl>
## 1    0.502    0.521
```

- Wir erhalten für jede Person einen Schätzer mit Standardfehler. Wir können diese mitteln, um einen Schätzer des durchschnittlichen kausalen Effekts zu erhalten.
- Wir müssen die Schätzer entfernen, bei denen es wegen eines perfekten Zusammenhangs oder wegen fehlender intraindividuellen Varianz keine OLS Lösung gibt.

## Within Transformation

```
d_wi = d %>% select(IDsosci, verh1, verhint1) %>% group_by(IDsosci) %>% mutate(verh1_wi =
  mean(verh1), verhint1_wi = verhint1 - mean(verhint1)) %>% ungroup()

d_wi %>% select(-IDsosci) %>% summary
```

```
##      verh1      verhint1      verh1_wi      verhint1_wi
## Min.    :1.000    Min.    :1.000    Min.    :-3.00    Min.    :-3.00
## 1st Qu.:1.000    1st Qu.:1.000    1st Qu.: -0.25    1st Qu.: -0.25
## Median :1.000    Median :1.000    Median :  0.00    Median :  0.00
## Mean    :1.529    Mean    :1.804    Mean    :  0.00    Mean    :  0.00
## 3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:  0.00    3rd Qu.:  0.00
## Max.    :5.000    Max.    :5.000    Max.    :  3.00    Max.    :  3.00

d_wi %>% lm(verh1_wi ~ verhint1_wi, data = .) %>% tidy() %>% mutate_if(is.numeric,
  round, 2)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      0         0.01         0         1
## 2 verhint1_wi    0.35         0.01     24.9         0
```

- Intuitive Interpretation: Eine Abweichung vom Personen-Durchschnitt in  $X$  um einen Punkt führt zu einer Abweichung vom Personen-Durchschnitt in  $Y$  um  $b_X$  Punkte.
- Hier: Wenn eine Person um einen Punkt wahrscheinlicher rausgehen möchte als üblich, dann wird sie 0.34 Punkte häufiger rausgehen (beides auf 5er Skalen).
- Das ist durchaus ein bedeutsamer Effekt. Aber zur Erinnerung: Der naiven pooled OLS Schätzung zufolge war der Effekt fast doppelt so groß.

## Least Squares mit Dummy Variablen (LSDV)

```
d %>% lm(verh1 ~ verhint1 + factor(IDsosci), data = .) %>% tidy() %>% mutate_if(is.numeric,
  round, 2) %>% print(n = 17)
```

```
## # A tibble: 577 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)      0.65      0.28      2.31    0.02
## 2 verhint1         0.35      0.02     21.5     0
## 3 factor(IDsosci)02E6C8 -0.35      0.4     -0.86    0.39
## 4 factor(IDsosci)050IPY  1.21      0.4      3.01     0
## 5 factor(IDsosci)05J4R8  0.32      0.4      0.79    0.43
## 6 factor(IDsosci)08BDZJ  0.96      0.4      2.39    0.02
## 7 factor(IDsosci)0BHGLF  0.570     0.4      1.42    0.16
## 8 factor(IDsosci)0EB6C1  0         0.4       0         1
## 9 factor(IDsosci)0E09L2  1.95      0.4      4.82     0
## 10 factor(IDsosci)0F5L9Z  1.64      0.4      4.06     0
## 11 factor(IDsosci)0KAKHF  2.2       0.4      5.44     0
## 12 factor(IDsosci)0KYAJ -0.01      0.4     -0.02    0.98
## 13 factor(IDsosci)00NV40  0.33      0.4      0.82    0.41
```

```
## 14 factor(IDsosci)OPKFWT   -0.09      0.4      -0.22     0.83
## 15 factor(IDsosci)OZCKB5    0.32      0.4       0.79     0.43
## 16 factor(IDsosci)114OWA    0.33      0.4       0.82     0.41
## 17 factor(IDsosci)11KVRK   -0.17      0.4      -0.43     0.67
## # ... with 560 more rows
```

- Der Punktschätzer  $b_X$  entspricht genau dem Punktschätzer nach der within-person Transformation.
- Zusätzlich gibt die Regressionskonstante den Mittelwert für Person 1 an und die  $n - 1$  Koeffizienten der Dummy-Variablen die Abweichung der übrigen Personen von diesem Mittelwert. Es gelten die üblichen Regeln für die Interpretation solcher Koeffizienten.

### Welche Modellspezifikation soll ich nutzen?

- 1) Der Schätzer des durchschnittlichen kausalen Effekts in der no pooling Spezifikation ist im Vergleich zu den beiden anderen Varianten weniger effizient. Außerdem ist er praktisch schwieriger zu ermitteln, da er erst aus den Schätzern der Einzel-Modelle berechnet werden muss. Wenn wir die Annahme eines homogenen kausalen Effekts treffen (und das tun wir üblicherweise), dann gibt es keinen Grund, das no pooling Modell in der Praxis zu verwenden.
  - 2) Die Spezifikationen mit within-person Transformation und LSDV ergeben dieselben Punktschätzer für den kausalen Effekt und sind insofern austauschbar.
  - 3) Die Standardfehler des Modells mit einer naiven within-person Transformation (wie oben dargestellt) sind zu klein, da wir die Stichprobenmittelwerte und nicht die (mit Unsicherheit behafteten) Schätzer der Populationsmittelwerte zur Zentrierung verwenden. Die Standardfehler müssen daher angepasst werden (passiert in spezialisierten Software-Paketen automatisch).
  - 4) Die LSDV Spezifikation ist in fast jedem Softwarepaket einfach umzusetzen. Mit großen Datensätzen wird aber die Schätzung langsam und der Output unübersichtlich.
- Unabhängig von der Spezifikation gelten weiterhin alle Annahmen der (OLS) Regression. Besonders gern vergessen wird der *omitted variable bias* durch nicht gemessene, über die Zeit variierende  $Z$ . *Fixed effects* Modelle kontrollieren nur die  $Z$ , die auf konstante Merkmale der als *fixed effects* spezifizierten Einheiten zurückgehen.
  - Insgesamt sind viele quantitative Sozialforscher (v.a. die mit einer Ökonometrie-Ausbildung) der Ansicht, dass *fixed effects* Modelle die beste Methode sind, um kausale Effekte aus nicht-experimentellen Daten zu schätzen.

### 3.1.1 Mehre fixed effects in einem Modell – Perioden-Effekte

- Grundsätzlich können in einem Modell beliebig viele *fixed effects* spezifiziert werden.
- In Paneldaten ist der Erhebungszeitpunkt bzw. die Erhebungsperiode (Panelwelle) eine typische Variable, über die verschiedene, für alle Personen konstante Effekte kontrolliert werden können.
- Einige Lehrbücher empfehlen, dies *immer* zu tun, da kausale Effekte von Ereignissen, die für alle Einheiten konstant sind, statistisch nicht identifiziert sind.
- Eine typische Spezifikation ist die Aufnahme eines *fixed effects* für den Indikator der Panelwelle.
- In der LSDV-Spezifikation kann einfach ein weiterer Dummy-Faktor hinzugefügt werden. Die within-person Transformation ist mathematisch komplizierter, wird aber in spezialisierten Software-Paketen im Hintergrund erledigt. Es können auch beide Spezifikationen kombiniert werden, wenn z.B. die Periodeneffekte von inhaltlichem Interesse sind und im Output angezeigt werden sollen (siehe nächsten Teilabschnitt).

#### Ein Beispiel mit *fixed effects* für Personen und Perioden

```
d %>% lm(verh1 ~ verhint1 + factor(wave) + factor(IDsosci), data = .) %>% tidy() %>%
  mutate_if(is.numeric, round, 2) %>% print(n = 17)
```

```
## # A tibble: 580 x 5
##   term                estimate std.error statistic p.value
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)         0.6      0.28     2.13     0.03
## 2 verhint1            0.33     0.02    19.8      0
## 3 factor(wave)2       0.02     0.03     0.6     0.55
## 4 factor(wave)3       0.14     0.03     4.15     0
## 5 factor(wave)4       0.12     0.03     3.41     0
## 6 factor(IDsosci)02E6C8 -0.33    0.4     -0.83    0.41
## 7 factor(IDsosci)050IPY  1.26    0.4      3.15     0
## 8 factor(IDsosci)05J4R8  0.34    0.4      0.85    0.39
## 9 factor(IDsosci)08BDZJ  1.01    0.4      2.53    0.01
## 10 factor(IDsosci)0BHGLF  0.59    0.4      1.48    0.14
## 11 factor(IDsosci)0EB6C1  0       0.4      0       1
## 12 factor(IDsosci)0E09L2  2.02    0.4      5.01     0
## 13 factor(IDsosci)0F5L9Z  1.68    0.4      4.2      0
## 14 factor(IDsosci)0KAKHF  2.27    0.4      5.63     0
## 15 factor(IDsosci)0KYAJ  0       0.4      0.01    0.99
## 16 factor(IDsosci)00NV40  0.34    0.4      0.84     0.4
## 17 factor(IDsosci)0PKFWT -0.08    0.4     -0.21    0.84
## # ... with 563 more rows
```

- $b_{verhint1}$  quantifiziert weiterhin den kausalen Effekt von Interesse. Er ist robust gegen die Kontrolle des Periodeneffekts.
- Die  $b_{wave_t}$  zeigen den Kontrast zur ersten Welle. In diesem Fall sind liegen in der dritten und vierten Welle die Häufigkeiten des Rausgehens höher als noch in den ersten beiden Wellen.
- Die  $b_{id_i}$  zeigen weiterhin den Kontrast zu Person 1 (substantiell nicht sonderlich interessant).

## 3.2 Übungsaufgaben 1

- 1) Schätze den kausalen Effekt der Informationshäufigkeit aus öffentlich-rechtlichen TV-Programmen (`med2`) auf die Intention, weniger als 1.5m Abstand zu einer Person zu halten, die nicht im eigenen Haushalt lebt (`verhint3`).
  - Schätze zuerst das *falsche* pooled OLS Modell.
  - Schätze dann das einfache *fixed effects* Modell mit einer Spezifikation freier Wahl.
  - Vergleiche schließlich die Modelle mit und ohne Periodeneffekt.
- 2) Spezifiziere, schätze und interpretiere ein eigenes bivariates *fixed effects* Modell mit Daten aus dem Beispieldatensatz.

## 3.3 *Fixed effects* Modelle in der praktischen Anwendung

- Auch wenn wir das *fixed effects* Modell nur mit `stats::lm()` und der LSDV-Spezifikation schätzen können, ist die weitere Arbeit mit diesen Modellen nicht ideal - besonders, wenn wir tiefer in Detail-Anpassungen einsteigen.
- Zudem wird das Schätzen mit `stats::lm()` und LSDV bei großen Datensätzen und mit vielen *fixed effects* langsam.
- `plm` (Croissant et al., 2020) ist das etablierte Paket für das Schätzen von ökonometrischen Panel-Modellen in R. Es bietet ein einfaches Interface zu allen Standardmodellen (und zu den übrigen Klassikern der Ökonometrie, *instrumental variables*, *differences in differences*).
- Das Schätzen der Modelle basiert auf OLS mit Datentransformationen im Hintergrund. Dadurch ist das Schätzen wesentlich schneller als mit einer LSDV-Spezifikation. Die notwendigen Anpassungen der Standardfehler werden ebenfalls vorgenommen.

### Spezifikation eines einfachen *fixed effects* Modells mit `plm`

```
d %>% plm(verh1 ~ verhint1, data = ., index = "IDSosci", model = "within") %>% summary

## Oneway (individual) effect Within Model
```

### 3.3. FIXED EFFECTS MODELLE IN DER PRAKTISCHEN ANWENDUNG23

```
##
## Call:
## plm(formula = verh1 ~ verhint1, data = ., model = "within", index = "IDSosci")
##
## Balanced Panel: n = 576, T = 4, N = 2304
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -3.000000 -0.163516  0.000000  0.095937  3.000000
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## verhint1 0.345937    0.016061  21.539 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    702.5
## Residual Sum of Squares: 553.75
## R-Squared:    0.21175
## Adj. R-Squared: -0.051155
## F-statistic: 463.924 on 1 and 1727 DF, p-value: < 2.22e-16
```

- Das *fixed effects* Modell wird über `model = "within"` angefordert. Mit `index = "IDSosci"` wird der Indikator für die Einheiten angegeben.
- Der Output von `summary()` liefert eine korrekte Beschreibung der Fallzahlen im Datensatz.
- Beachte: Das angepasste  $R^2$  ist hier (wie in vielen *fixed effects* Modellen) negativ. Das ist kein Grund zur Beunruhigung. Die Logik dahinter kann gut nachvollzogen werden, wenn wir uns die LSDV-Spezifikation in Erinnerung rufen. Zusätzlich zu den inhaltlich relevanten Prädiktoren enthält das Modell  $n - 1$  Prädiktoren für die Einheiten.

#### 3.3.1 Mehre *fixed effects* in einem Modell – Perioden-Effekte mit `plm`

- `plm` bietet zwei Möglichkeiten, die Perioden-Effekte zu spezifizieren (identische Ergebnisse, anderer Output):
- 1) Zwei Indices `index=c("IDSosci", "wave")` und `effect = "twoways"` für die within-Transformation.
    - Es wird "still" für Personen und Perioden kontrolliert, beide werden nicht im Output angezeigt.
    - Das  $R^2$  bezieht sich nur auf die Varianzaufklärung durch die Prädiktoren.
  - 2) Perioden-Effekt als Dummies hinzufügen.
    - Praktisch, wenn es nur wenige Perioden gibt und wir die Ergebnisse dazu direkt im Output sehen wollen.

- Das  $R^2$  bezieht sich auf die Varianzaufklärung durch die Prädiktoren und den Perioden-Effekt.

```
d %>% plm(verh1 ~ verhint1, data = ., index = c("IDSosci", "wave"), model = "within",
  effect = "twoways") %>% summary()
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = verh1 ~ verhint1, data = ., effect = "twoways",
##      model = "within", index = c("IDSosci", "wave"))
##
## Balanced Panel: n = 576, T = 4, N = 2304
##
## Residuals:
##      Min.    1st Qu.    Median    3rd Qu.    Max.
## -3.070526 -0.180571  0.011669  0.131626  3.049432
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## verhint1 0.328779    0.016614  19.789 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    669.68
## Residual Sum of Squares: 545.72
## R-Squared:    0.1851
## Adj. R-Squared: -0.088576
## F-statistic: 391.609 on 1 and 1724 DF, p-value: < 2.22e-16
```

```
mdl_pfe_pdv = d %>% plm(verh1 ~ verhint1 + factor(wave), data = ., index = "IDSosci",
  model = "within")
mdl_pfe_pdv %>% summary()
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = verh1 ~ verhint1 + factor(wave), data = ., model = "within",
##      index = "IDSosci")
##
## Balanced Panel: n = 576, T = 4, N = 2304
##
## Residuals:
##      Min.    1st Qu.    Median    3rd Qu.    Max.
## -3.070526 -0.180571  0.011669  0.131626  3.049432
##
## Coefficients:
```



### 3.3. FIXED EFFECTS MODELLE IN DER PRAKTISCHEN ANWENDUNG 25

```
##              Estimate Std. Error t-value Pr(>|t|)
## verhint1      0.328779   0.016614 19.7891 < 2.2e-16 ***
## factor(wave)2 0.019997   0.033589   0.5954 0.5516856
## factor(wave)3 0.139955   0.033691   4.1540 3.426e-05 ***
## factor(wave)4 0.117763   0.034484   3.4150 0.0006526 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    702.5
## Residual Sum of Squares: 545.72
## R-Squared:              0.22318
## Adj. R-Squared:        -0.037717
## F-statistic: 123.824 on 4 and 1724 DF, p-value: < 2.22e-16
```

### Robuste Standardfehler

- In der ökonometrischen Diskussion ist die Wahl der korrekten (robusten) Standardfehler sehr prominent. Diese sind robust gegen Verletzung verschiedener Annahmen, z.B. durch serielle Autokorrelation oder Heteroskedastizität.
- Das `lmtest` Paket (Hothorn et al., 2019) ist kompatibel mit Modellen aus `plm`. Es stellt zahlreiche robuste Standardfehler zur Verfügung.
- Hier die “normalen” Standardfehler und bei Heteroskedastizität robuste Standardfehler sowie die darauf basierenden Konfidenzintervalle im Vergleich.
- Weiter wollen wir dieses Thema hier nicht vertiefen. Ich empfehle unter anderem King and Roberts (2015) zur kritischen Lektüre.

```
# Normale SE und CI
mdl_pfe_pdv %>% coeftest() %>% round(3)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## verhint1      0.329      0.017  19.789  <2e-16 ***
## factor(wave)2  0.020      0.034   0.595    0.552
## factor(wave)3  0.140      0.034   4.154  <2e-16 ***
## factor(wave)4  0.118      0.034   3.415   0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mdl_pfe_pdv %>% coefci() %>% round(3)

##              2.5 % 97.5 %
## verhint1      0.296  0.361
```

```
## factor(wave)2 -0.046 0.086
## factor(wave)3 0.074 0.206
## factor(wave)4 0.050 0.185

# Heteroskedasticity-robust SE and CI
mdl_pfe_pdv %>% coeftest(vcov. = vcovHC) %>% round(3)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## verhint1          0.329      0.028  11.868  <2e-16 ***
## factor(wave)2       0.020      0.029   0.688    0.491
## factor(wave)3       0.140      0.036   3.849  <2e-16 ***
## factor(wave)4       0.118      0.032   3.701  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mdl_pfe_pdv %>% coefci(vcov. = vcovHC) %>% round(3)

##              2.5 % 97.5 %
## verhint1          0.274 0.383
## factor(wave)2     -0.037 0.077
## factor(wave)3      0.069 0.211
## factor(wave)4      0.055 0.180
```

## Aufnahme weiterer über die Zeit variierender Prädiktoren

- Die Aufnahme weiterer Prädiktoren, die über die Zeit variieren, erfolgt prinzipiell wie im bekannten OLS Modell.
- Wichtig ist, dass es bei *fixed effects* Modellen explizit um das Schätzen von kausalen Effekten geht. Entsprechend bedacht sollte die Auswahl von weiteren Prädiktoren sein. Ein “kitchen sink” Ansatz, den man vor allem in OLS mit Querschnittsdaten sieht, ist hier nicht angebracht. Es muss (wie eigentlich immer) darauf geachtet werden, welche Koeffizienten eines Regressionsmodells kausal interpretiert werden dürfen (Keele et al., 2019). Im *fixed effects* Modell müssen wir uns das ganz explizit vergegenwärtigen und in der Ergebnisdarstellung berücksichtigen, da die Modellklasse kausale Effekte impliziert.
- Nach der TPB dürfen wir dieses Modell annehmen, da die drei Prädiktoren auf derselben kausalen Stufe stehen: Verhaltensintention ~ Einstellung + Deskriptive Norm + Injunktive Norm. Hier schätzen wir das Modell für die Verhaltensintention *Rausgehen ohne triftigen Grund*.

```
d %>% plm(verhint1 ~ ein1 + desnormp1 + injnormp1 + factor(wave), data = ., index = "ID",
model = "within") %>% tidy() %>% mutate_if(is.numeric, round, 2)
```

### 3.3. FIXED EFFECTS MODELLE IN DER PRAKTISCHEN ANWENDUNG27

```
## # A tibble: 6 x 5
##   term                estimate std.error statistic p.value
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>
## 1 ein1                0.31      0.02     12.4      0
## 2 desnormp1           0.05      0.03      1.56    0.12
## 3 injnormp1           0.1       0.03      3.31     0
## 4 factor(wave)2        0.21      0.05      4.69     0
## 5 factor(wave)3        0.24      0.05      5.3      0
## 6 factor(wave)4        0.39      0.05      8.32     0
```

- Vor allem die Einstellung zum Verhalten und die wahrgenommenen normativen Erwartungen haben stärkere kausale Effekte auf die Verhaltensintention.

**Aufnahme eines Personenmerkmals (funktioniert nicht, ohne Warnung!)**

```
d %>% plm(verh1 ~ verhint1 + C_sex + factor(wave), data = ., index = "IDSosci", model = "within")
summary
```

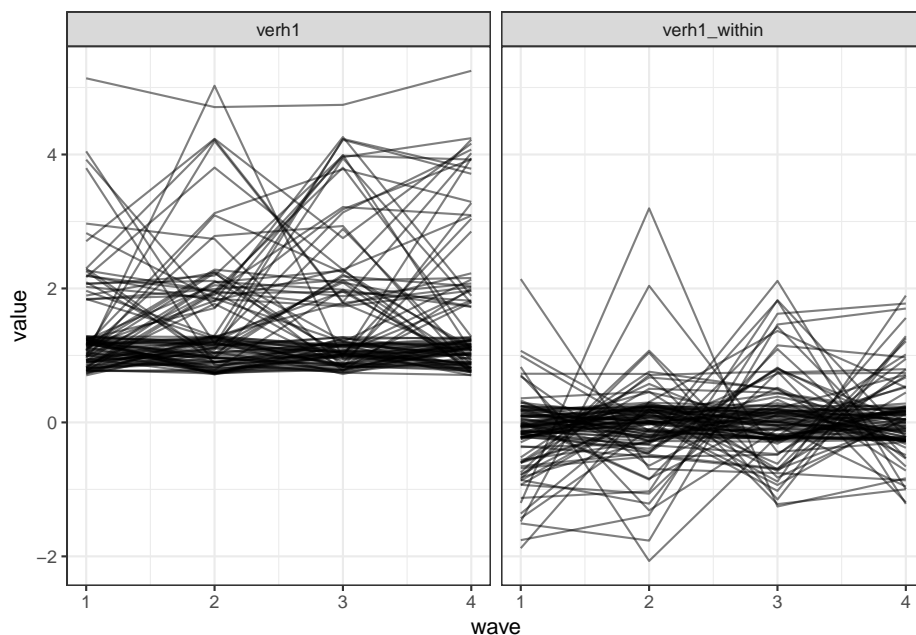
```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = verh1 ~ verhint1 + C_sex + factor(wave), data = .,
##      model = "within", index = "IDSosci")
##
## Balanced Panel: n = 576, T = 4, N = 2304
##
## Residuals:
##      Min.      1st Qu.      Median      3rd Qu.      Max.
## -3.070526 -0.180571  0.011669  0.131626  3.049432
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## verhint1      0.328779   0.016614 19.7891 < 2.2e-16 ***
## factor(wave)2  0.019997   0.033589  0.5954 0.5516856
## factor(wave)3  0.139955   0.033691  4.1540 3.426e-05 ***
## factor(wave)4  0.117763   0.034484  3.4150 0.0006526 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    702.5
## Residual Sum of Squares: 545.72
## R-Squared:              0.22318
## Adj. R-Squared:        -0.037717
## F-statistic: 123.824 on 4 and 1724 DF, p-value: < 2.22e-16
```

- Geschlecht wird nicht in das Modell aufgenommen. Vorsicht: Es taucht einfach nicht im Ergebnis auf, obwohl es in der Formel steht (siehe Call in der Summary)

### Warum wird das Personenmerkmal nicht ins Modell aufgenommen?

- Within-person Transformation entfernt die gesamte between-person Varianz aus den Daten:  $\bar{y}_i = 0$ .
- Daher können innerhalb der Personen invariante Merkmale keine Unterschiede erklären.

```
id_sample = sample(unique(d$IDSosci), 100)
d %>% filter(IDSosci %in% id_sample) %>% select(IDSosci, wave, verh1) %>% group_by(IDSosci) %>%
  mutate(verh1_within = verh1 - mean(verh1)) %>% ungroup() %>% gather(transformation =
    value, -IDSosci, -wave) %>% ggplot(aes(wave, value, group = IDSosci)) + geom_line(
    height = 0.3), show.legend = FALSE, alpha = 0.5) + facet_wrap("transformation")
```



- Die Abbildung verdeutlicht dies anhand von 100 zufällig ausgewählten Personen aus dem Datensatz. Vor der Transformation gibt es (etwas) Varianz im Level der Verhaltensintention zwischen den Personen. Durch die Transformation verschwinden diese Unterschiede, es bleibt nur die Variation innerhalb der Personen über die Zeit.
- Das gleiche gilt für Prädiktoren, die als Merkmale anderer Einheiten, die wir als *fixed effects* spezifiziert haben, konstant sind. In diesem Beispiel wären dies Eigenschaften der Perioden, also z.B. neue Schutzmaßnahmen

### 3.4. CONCLUSIO: VOR- UND NACHTEILE DES FIXED EFFECTS MODELLS<sup>29</sup>

bzw. deren Lockerung, soweit sie alle Personen gleichermaßen im gleichen Zeitraum betreffen.

#### 3.3.2 Interaktionen mit Personenmerkmalen

- Wir können jedoch Interaktionen zwischen über die Zeit variierenden Prädiktoren und Personenmerkmalen (oder Merkmalen anderer *fixed effects* Einheiten) ins Modell aufnehmen.
- Bei kategoriellen Moderator-Variablen erhalten wir Schätzer der gruppenspezifischen Effekte, z.B. den Unterschied zwischen den Effekten der Verhaltensintention auf das Verhalten für Frauen und Männer.
- Bei kontinuierlichen Moderator-Variablen gelten die üblichen Fallstricke: Der Koeffizient des Prädiktors ist nun der einfache Effekt für den Fall, dass der Moderator gleich 0 ist.

```
d %>% plm(verh1 ~ verhint1 * C_sex + factor(wave), data = ., index = "IDsossci", model = "within")
tidy() %>% mutate_if(is.numeric, round, 2)
```

```
## # A tibble: 5 x 5
##   term                estimate std.error statistic p.value
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>
## 1 verhint1           0.34      0.03     13.2      0
## 2 factor(wave)2       0.02      0.03      0.59     0.55
## 3 factor(wave)3       0.14      0.03      4.14      0
## 4 factor(wave)4       0.12      0.03      3.42      0
## 5 verhint1:C_sex     -0.01      0.03     -0.39     0.7
```

- Der Effekt ist in der Stichprobe für Frauen minimal schwächer als für Männer. Der Unterschied ist jedoch weder substantiell noch statistisch bedeutsam.

### 3.4 Conclusio: Vor- und Nachteile des *fixed effects* Modells

In many applications the whole point of using panel data is to allow for  $a_i$  to be arbitrarily correlated with the  $x_{it}$ . A fixed effects analysis achieves this purpose explicitly. — Wooldridge (2010), p. 300

By controlling out context, FE models effectively cut out much of what is going on — goings-on that are usually of interest to the researcher, the reader and the policy maker. We contend that models that control out, rather than explicitly model, context and heterogeneity offer overly simplistic and impoverished results that can lead to misleading interpretations. — Bell and Jones (2015), p. 134

- Das *fixed effects* Modell ist nützlich, wenn wir einen kausalen Effekt, der sich innerhalb von Einheiten (Personen) abspielt, schätzen wollen.

- Das *fixed effects* Modell kann keine Merkmale der Einheiten (Personen) als Prädiktoren berücksichtigen, da die gesamten einheiten(personen)spezifischen Unterschiede bereits durch die *fixed effects* erklärt werden.
- Wir interessieren uns aber häufig (auch) für die Unterschiede zwischen Einheiten (Personen). Das *fixed effects* Modell macht Antworten auf solche Fragen unmöglich.
- Ein weiterer, damit unverbundener Nachteil des *fixed effects* Modells ist die starke Anfälligkeit für Messfehler. Die Transformation verringert die wahre Varianz deutlich, während große Teile der Messfehlervarianz erhalten bleiben (sie sind nicht personenspezifisch).

### 3.5 Übungsaufgaben 2

- 1) Schätze den kausalen Effekt der Informationshäufigkeit aus öffentlich-rechtlichen TV-Programmen (`med2`) auf die Intention, weniger als 1.5m Abstand zu einer Person zu halten, die nicht im eigenen Haushalt lebt (`verhint3`). Berücksichtige dabei auch die Periodeneffekte der Panelwellen. Siehe dazu auch Übung 1.
  - Verwende jetzt `plm` für die Schätzung.
  - Nimm zusätzlich die Information aus Zeitungen und Zeitschriften (`med1`) in das Modell auf.
  - Prüfe, ob sich der Effekt der Information aus Zeitungen und Zeitschriften nach Geschlecht (`C_sex`) unterscheidet.
- 2) Spezifiziere, schätze und interpretiere ein eigenes *fixed effects* Modell mit Daten aus dem Beipsioldatensatz. Nutze dabei alle Techniken (unterschiedliche Spezifikation, Standardfehler, Moderation, mehrere Prädiktoren), die du ausprobieren und zu denen du ggf. Fragen stellen willst.

## 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.
- Croissant, Y., Millo, G., and Tappe, K. (2020). *plm: Linear Models for Panel Data*. R package version 2.2-3.
- Hothorn, T., Zeileis, A., Farebrother, R. W., and Cummins, C. (2019). *lmtest: Testing Linear Regression Models*. R package version 0.9-37.
- Keele, L., Stevenson, R. T., and Elwert, F. (2019). The causal interpretation of estimated associations in regression models. *Political Science Research and Methods*, pages 1–13.
- King, G. and Roberts, M. E. (2015). How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It. *Political Analysis*, 23(2):159–179.
- 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.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.