Mini-Workshop Panel Data Analysis

Marko Bachl (mit Material von Michael Scharkow)

Sommersemester 2020 | IJK Hannover

Contents

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 speziallisierter 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.
- Grundkentnisse 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 datenanalystischen 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 vielfälltigen 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 Anlehung 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)
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
pacman	0.5.1
purrr	0.3.3
readr	1.3.1
stringr	1.4.0
tibble	2.1.3
tidyr	1.0.2
tidyverse	1.3.0

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 Entfalltung 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 Paneleffekte
  - Panelmortalität: Einheiten (Befragte) fallen aus, möglicherweise systematisch mit Bezug auf die Konstrukte oder Effekte, die uns interessieren.
  - 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

i	t	y
1	1	6.55
1	2	6.68
1	3	6.77
2	1	5.55
2	2	6.01
...		

Wide Format

i	y_{t1}	y_{t2}	y_{t3}	y_{t4}
1	6.55	6.68	6.77	7.04
2	5.55	6.01	6.32	6.40
3	4.65	5.33	6.45	6.45
...				

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

- 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, Reglungen 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`

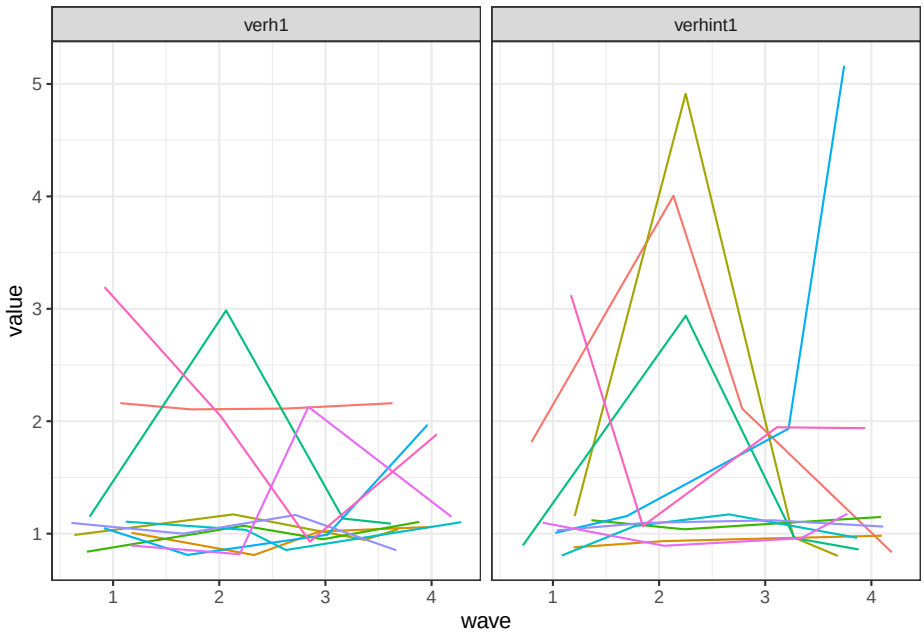
Variablenname	Label
alter	Alter
besorg1	Ich bin besorgt, wenn ich an Corona denke.
bildung	Bildungsabschluss
desnormp1	…sind in der letzten Woche rausgegangen, auch wenn es sich nicht um einen Arztbes
desnormp2	…haben sich in der letzten Woche in ihrer Freizeit mit mehr als einer anderen Perso
desnormp3	…haben in der letzten Woche weniger als 1,5 Meter Abstand zu Personen gehalten, d
desnormp4	...haben sich in der letzten Woche strikt an die Maßnahmen zur Beschränkung sozial
ein1	Ich finde es in Ordnung, wenn man rausgeht, auch wenn es sich nicht um einen Arzt
ein2	Ich finde es in Ordnung, wenn man sich in seiner Freizeit mit mehr als einer anderen
ein3	Ich finde es in Ordnung, wenn man weniger als 1,5 Meter Abstand zu Personen hält,
ein4	Ich finde es wichtig, dass die Empfehlung zur Beschränkung sozialer Kontakte strikt
ein5	Ich finde es richtig, dass generell Abstand gehalten werden soll.
injnormp1	…finden es in Ordnung, wenn man rausgeht, auch wenn es sich nicht um einen Arztb
injnormp2	… finden es in Ordnung, wenn man sich in seiner Freizeit mit mehr als einer anderen
injnormp3	…finden es in Ordnung, wenn man weniger als 1,5 Meter Abstand zu Personen hält,
injnormp4	...finden es in Ordnung, wenn man sich strikt an die Maßnahmen zur Beschränkung
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, ffn)
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, Ei
verh2	Ich habe mich mit mehr als einer Person getroffen, die nicht in meinem Haushalt leb
verh3	Ich habe weniger 1,5 Meter Abstand zu Personen gehalten, die nicht in meinem Hau
verh4	Ich habe mich strikt an die Maßnahmen zur Beschränkung sozialer Kontakte gehalte
verh5	Ich habe mich im Privaten mit Freunden oder Familienmitgliedern getroffen, die nic
verh6	Ich war länger draußen als für einen üblichen Spaziergang (z.B. saß auf der Wiese o
verhint1	Rausgehen, auch wenn es sich nicht um einen Arztbesuch, Arbeitsweg, Einkauf, Spa
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 le
verhint4	Mich strikt an die Maßnahmen zur Beschränkung sozialer Kontakte halten.
verhint5	Mich im Privaten mit Freunden oder Familienmitgliedern treffen, die nicht in meine
verhint6	Mich länger draußen aufhalten als für einen üblichen Spaziergang (z.B. auf der Wies
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 = 0.2), 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)     0.47       0.02       19.6       0
## 2 verhint1        0.59       0.01       53.8       0
```

- Das Modell besagt, dass die Häufigkeit, nach 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 Korreationen durch die zeitliche Abhängigkeit der Messungen
3) Annahme der Unabhängigkeit der Bebobachtungen 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

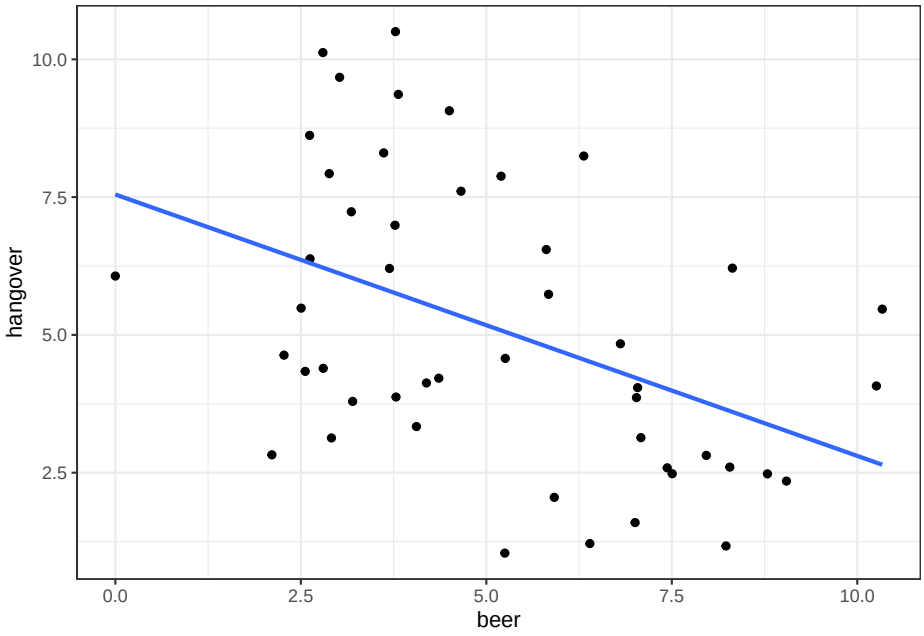
- Unser Ziel ist es, den wahren kausalen Effekt von X auf Y zu schätzen.
- Pooled OLS vermischt aber zwei Quellen von Unterschieden in den Daten: Den (kausalen) Effekt innerhalb der Personen (within) und die Unterschiede zwischen Personen (between).
- Within und between Effekte können sich in Größe und sogar in der Richtung unterscheiden!
- Die Schätzung aus einem poold 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.

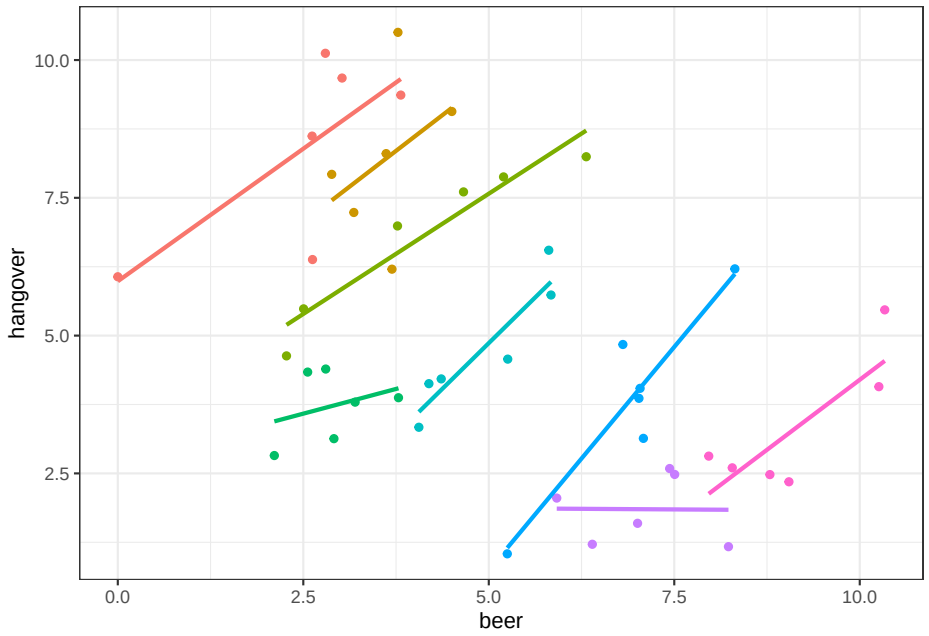
Pooled OLS, within und between - eine Illustration

- Zum Abschluss noch ein imagninäres 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 Rergressionsgerade durch alle Beobachung 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 beobachte 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 kauselen (within-person) Effekt mit Paneldaten schätzen?

1) Separate OLS Modelle für jede Person schätzen und Koeffizeinten 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 beobachten,) 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) %>% fil
nest() %>% mutate(mdls = map(data, ~tidy(lm(verh1 ~ verhint1, data = .x)))) %>% unnest
    ungroup() %>% select(-data) %>% na.omit() %>% filter(statistic != Inf) %>% filter(t
    "verhint1") %>% mutate_if(is.numeric, round, 2) %>% print %>% summarise(estimate =
    std.error = sqrt(mean(std.error^2)))  # simple approximation
```

```
## # A tibble: 236 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 0EO9L2  verhint1     1.67      0.67  2.50e+ 0    0.13
##  5 0F5L9Z  verhint1     0         0.71  0.          1
##  6 0KYYAJ  verhint1     0.45      0.18  2.50e+ 0    0.13
##  7 0ONV4O  verhint1     1         0     9.01e+15    0
##  8 0Q5XIM  verhint1    -0.27      0.31 -8.70e- 1    0.48
##  9 0ZCKB5  verhint1    -0.35      0.5  -6.90e- 1    0.56
## 10 114OWA  verhint1     0.33      0.33  1.00e+ 0    0.42
## # ... with 226 more rows
```

```
## # A tibble: 1 x 2
##   estimate std.error
##      <dbl>     <dbl>
## 1    0.493     0.520
```

- 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 das Modell wegen eines perfekten Zusammenhangs oder wegen fehlender *within-person* Varianz keine OLS Lösung hat.

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.   :1.0   Min.   :-3.00   Min.   :-3.00
##  1st Qu.:1.000   1st Qu.:1.0   1st Qu.:-0.25   1st Qu.:-0.25
##  Median :1.000   Median :1.0   Median : 0.00   Median : 0.00
##  Mean   :1.526   Mean   :1.8   Mean   : 0.00   Mean   : 0.00
##  3rd Qu.:2.000   3rd Qu.:2.0   3rd Qu.: 0.00   3rd Qu.: 0.00
##  Max.   :5.000   Max.   :5.0   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     0.34      0.01      24.7       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: 586 x 5
##    term                   estimate std.error statistic p.value
##    <chr>                     <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)               0.66       0.28      2.32    0.02
##  2 verhint1                  0.34       0.02     21.4     0
##  3 factor(IDsosci)02E6C8    -0.34       0.4      -0.85    0.39
##  4 factor(IDsosci)050IPY     1.23       0.4       3.04    0
##  5 factor(IDsosci)05J4R8     0.32       0.4       0.81    0.42
##  6 factor(IDsosci)08BDZJ     0.98       0.4       2.42    0.02
##  7 factor(IDsosci)0BHGLF     0.570      0.4       1.43    0.15
##  8 factor(IDsosci)0EB6C1     0          0.4       0       1
##  9 factor(IDsosci)0EO9L2     1.97       0.4       4.87    0
## 10 factor(IDsosci)0F5L9Z     1.65       0.4       4.09    0
## 11 factor(IDsosci)0KAKHF     2.22       0.4       5.49    0
## 12 factor(IDsosci)0KYYAJ    -0.01       0.4      -0.01    0.99
## 13 factor(IDsosci)0ONV4O     0.33       0.4       0.82    0.41
```

```
## 14 factor(IDsosci)0PKFWT   -0.09       0.4      -0.21    0.83
## 15 factor(IDsosci)0Q5XIM    0.32       0.4       0.81    0.42
## 16 factor(IDsosci)0ZCKB5    0.32       0.4       0.81    0.42
## 17 factor(IDsosci)114OWA    0.33       0.4       0.82    0.41
## # ... with 569 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 variierdene 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 Ergebungsperiode (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-Pakten 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: 589 x 5
##    term                  estimate std.error statistic p.value
##    <chr>                    <dbl>     <dbl>     <dbl>   <dbl>
##  1 (Intercept)               0.61      0.28      2.15    0.03
##  2 verhint1                  0.32      0.02     19.6     0
##  3 factor(wave)2             0.02      0.03      0.66    0.51
##  4 factor(wave)3             0.14      0.03      4.25    0
##  5 factor(wave)4             0.12      0.03      3.53    0
##  6 factor(IDsosci)02E6C8    -0.32      0.4      -0.81    0.42
##  7 factor(IDsosci)050IPY     1.28      0.4       3.19    0
##  8 factor(IDsosci)05J4R8     0.35      0.4       0.87    0.39
##  9 factor(IDsosci)08BDZJ     1.03      0.4       2.57    0.01
## 10 factor(IDsosci)0BHGLF     0.6       0.4       1.5     0.13
## 11 factor(IDsosci)0EB6C1     0         0.4       0       1
## 12 factor(IDsosci)0EO9L2     2.04      0.4       5.06    0
## 13 factor(IDsosci)0F5L9Z     1.69      0.4       4.23    0
## 14 factor(IDsosci)0KAKHF     2.29      0.4       5.68    0
## 15 factor(IDsosci)0KYYAJ     0.01      0.4       0.02    0.99
## 16 factor(IDsosci)0ONV4O     0.34      0.4       0.85    0.4
## 17 factor(IDsosci)0PKFWT    -0.08      0.4      -0.2     0.84
## # ... with 572 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).

Übungsaufgaben zur grundsätzlichen Spezifikation von *fixed effects* Modellen

1) Schätze den kausalen Effekt von X auf Y mit einem *fixed effects* Modell (andere Beispiele aus dem Datensatz).
   - INHALTLICH FORMULIERTE FRAGE
   - Vergleiche die Modelle mit und ohne Periodeneffekte.
2) Spezifiziere, schätze und interpretiere ein eigenes *fixed effects* Modell mit dem Beipsieldatensatz.

3.2 Fixed effects Modelle in der praktischen Anwendung

- mit `plm`

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.