Matching

Kai Arzheimer | Vorlesung Forschungsmethoden

Outline

Wiederholung: Kontrafaktische Definition von Kausalität

- Gedankenexperiment:
  1. Beobachte Wert Y_{i0} an Objekt i in Welt wo $X_i = 0$; z.B. X≡ Arbeitslosigkeit, Y≡ Rechtsextremismus, $i \equiv$ Petra Musterfrau
  2. Beobachte Y_{i1} in einer "closest possible world" wo $X_i = 1$ (ansonsten keine Veränderungen)
  3. Kausaler Effekt von X auf $Y = Y_{i|x=1} - Y_{i|x=0}$
- Kausaler Effekt *für einen einzelnen Fall*
- Randomisiertes Experiment als beste Annäherung in Sozialwissenschaften

Warum funktionieren randomisierte Experimente (meistens)?

- Randomisierung → Experimentalgruppe und Kontrollgruppe homogen bezüglich aller denkbaren Kovariaten (große Gruppen)
- Mittelwertunterschiede in Y unverzerrte Schätzung für kausalen Effekt von Y ...
- ... in *typischen Fällen*

Was sind mögliche Probleme beim randomisierten Experiment?

- Compliance/cross-over
  - Design nicht mehr balanciert
  - Wenn nicht zufällig, keine Randomisierung mehr
- Gültigkeit von SUTVA (Stable Unit Treatment Value Assumption)
  - Stable Unit Treatment Value Assumption
  - Es darf nur eine Variante des Treatments geben (keine Variation des Mechanismus innerhalb Experimentalgruppe)
  - Einheiten bzw. deren Zahl dürfen sich nicht gegenseitig beeinflussen
  - Berufsqualifikation

Was sind die Probleme des Ex-Post-Facto Designs?

- Gruppen sind nicht gleich groß (oversampling)
- Kontrollgruppe mit größerer Streuung → Extrapolation
- Anfälligkeit gegen Fehlspezifikation des Modells
- *Selbstselektion*
  - X=1 und höherer Wert von Y: beides Effekt einer Hintergrundvariable
  - Selbst wenn eigenständiger kausaler Effekt δ auf Y existiert ...
    - Möglicherweise stärker/schwächer bei Personen, die sich *typischerweise* in X=1 selektieren
    - Interaktion/indirekter Drittvariableneffekt

Average Treatment Effect etc.

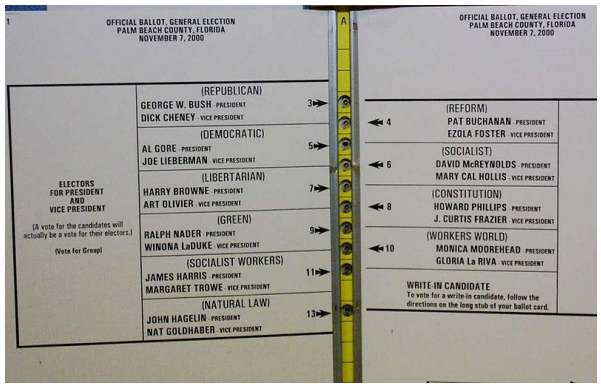
- Average Treatment Effect ATE von δ: $E[\delta] = E[Y^1] - E[Y^0]$
- Problem:
  - ATE (gewichtetes) Mittel von Effekt für Experimental- und Kontrollgruppe
  - Was ist, wenn sich Effekt in beiden Gruppen unterscheidet?
  - ATT: $E[\delta|D=1] = E[Y^1 - Y^0|D=1]$
  - ATC: $E[\delta|D=0] = E[Y^1 - Y^0|D=0]$
  - Randomisiertes Experiment: Warum sollte sich Effekt zwischen Gruppen unterscheiden?
  - Observational Data: Hilfsannahmen
  - Wenn Annahmen falsch, Über-/Unterschätzung des kausalen Effekts

Average Treatment Effect for the Treated

Manchmal nur ATT (oder ATC) interessant

Beispiel Butterfly Ballots

- Präsidentschaftswahl 2000: Manche counties in Florida verwenden „Butterfly Ballots“
- In diesen counties ungewöhnlich viele Stimmen für Buchanan
- Kausaler Effekt?
- Wie hoch wäre der Stimmenanteil für Buchanan
  - *in diesen counties* gewesen
  - *Wenn anderes Format*?



Was erhofft man sich von matching?

Theorie ...

- Bessere Balancierung der Daten
- Realistischerer Vergleich/weniger Extrapolation
- Geringerer Modellabhängigkeit der Schätzungen für den fokalen Effekt
- *Annäherung* an Schätzung kausaler Effekte mit observational data

Was erhofft man sich von matching?

Theorie ...

- Bessere Balancierung der Daten
- Realistischerer Vergleich/weniger Extrapolation
- Geringerer Modellabhängigkeit der Schätzungen für den fokalen Effekt
- *Annäherung* an Schätzung kausaler Effekte mit observational data

... und Praxis

The Gathering of the Manna by James Tissot

Was ist matching?

- Kontrollierter Ausschluß von Fällen (i.d.R.) aus der Kontrollgruppe
- Ziele:
  - Daten balanciert bezogen auf (potentiell) kausale unabhängige Variable (treatment)
  - Korrelation zwischen treatment und anderen unabhängigen Variablen aufgebrochen
- Oberbegriff für eine Vielzahl von Methoden
- Einfachste Variante: Exaktes 1:1 matching
- Alternative Verfahren verlieren weniger Fälle

Was genau ist Balancierung?

- Wir haben eine Reihe von potentiell relevanten Drittvariablen ($\mathbf{X}$)
- Balancierung:
  - Nicht nur der Mittelwert ...
  - ... sondern die Dichte (Verteilung) dieser Variablen ist in Experimental- und Kontrollgruppe gleich
  - Univariat und multivariat
- Verschiedene Matching-Verfahren → verschiedene Kriterien/Ansätze für Balancierung

Vorsicht: Post-treatment Bias

- King: „Controlling away for the consequences of treatment, causal ordering among predictors ambiguous“
- Beispiel:
  - Kausaler Effekt PI → Wahlentscheidung
    - Für Rasse kontrollieren
    - Aber nicht für Wahlabsicht unmittelbar vor Wahl
- Oft ist die Lage unklar
- Demokratisierung → Bürgerkriege
  - Für GDP kontrollieren wg GDP → Demokratisierung
  - Aber: Wenn Demokratisierung → GDP, postreatment bias
- Mehr dazu hier gking.harvard.edu/talks/bigprobP.pdf

Warum funktioniert matching?

- Ziel: Reduktion von bias und Varianz
- Normalerweise: „first principle of statistics: more data is better“
- Gilt aber nur, wenn wir richtiges Modell/richtigen Schätzer haben
- Bei observational data Hauptproblem bias, nicht Varianz
- Ausschluß von Fällen bis zu einem Optimum

Warum nicht exaktes matching?

- 1:1 vermutlich ineffizient
- „Curse of dimensionality“ auch wenn 1:m
- In Politikwissenschaft momentan am weitesten verbreitet: propensity score matching
- (Mehr dazu gleich)

Matching vs randomisierte Experimente

Experimente

- Idealerweise zufällige Auswahl
- Zufällige Selektion
- Idealerweise (gleich) große Gruppen
- (In Erwartung): perfekte Balancierung für alle denkbaren X

Matching vs randomisierte Experimente

Experimente

- Idealerweise zufällige Auswahl
- Zufällige Selektion
- Idealerweise (gleich) große Gruppen
- (In Erwartung): perfekte Balancierung für alle denkbaren X

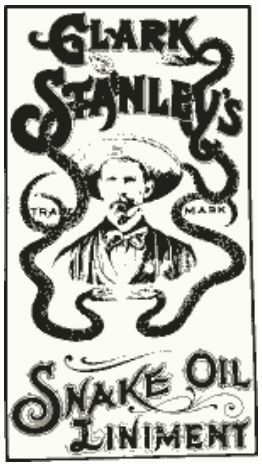
Matching

- Hoffentlich zufällige Auswahl
- Große Gruppen, aber Symmetrie?
- Selbstselektion
- *Nachträgliche Balancierung für einige X*
- Nicht grundsätzlich von multivariater Regression verschieden

Wie funktioniert propensity score matching?

PSM

- Grundidee: Fälle haben unterschiedliche Wahrscheinlichkeit (propensity), in treatment zu landen
- Schätze propensity auf Basis von **X** (Logit, alle Fälle)
- Balanciere Sample so, daß Verteilung von propensity für treatment und Kontrolle gleich
- (Indirekte Balancierung über Vielzahl von Variablen)
- Korrelation zwischen treatment und **X** aufgehoben, Unterschiede in Y (hoffentlich) treatment zurechenbar



Coarsened Exact Matching als Alternative?

- Iacus et al. 2011: Größte Probleme bias und Modellabhängigkeit
- Oft wird balance für manche X verbessert, für andere verschlechtert → Schrauberei
- Coarsening: X-Variablen zu (inhaltlich sinnvollen) Kategorien zusammenfassen (z.B. Schulbildung)
- Exaktes matching mit diesen „coarsened variables“; Ausschluß von Fällen, die nicht gut zu matchen sind
- Anschließend normales Modell (z.B.) Regression mit *ursprünglichen* Variablen als Kontrollvariablen
- In Computersimulationen besser, schneller, flexibler als existierende Techniken

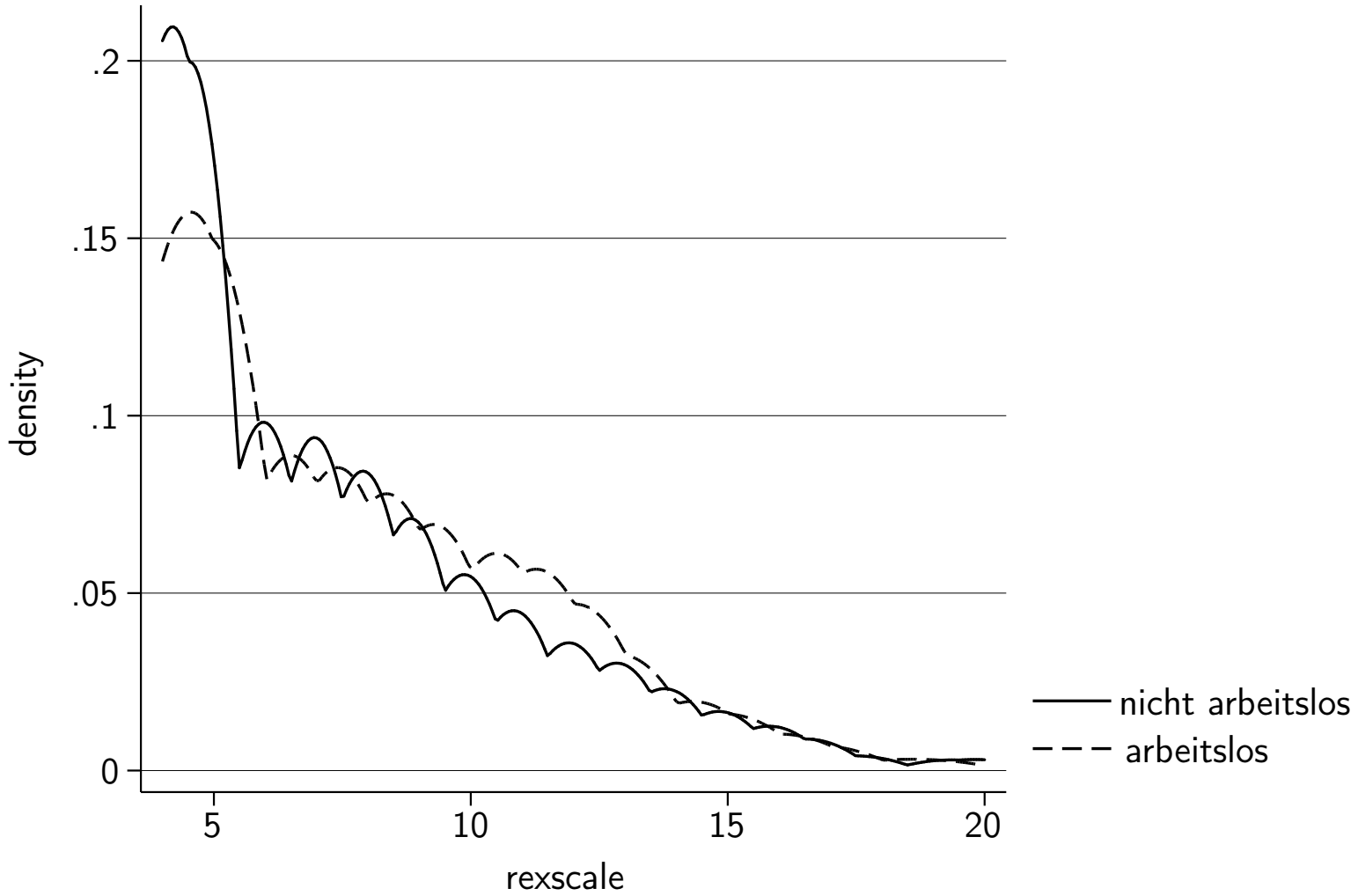
Coarsened Exact Matching als Alternative?

- Iacus et al. 2011: Größte Probleme bias und Modellabhängigkeit
- Oft wird balance für manche X verbessert, für andere verschlechtert → Schrauberei
- Coarsening: X-Variablen zu (inhaltlich sinnvollen) Kategorien zusammenfassen (z.B. Schulbildung)
- Exaktes matching mit diesen „coarsened variables“; Ausschluß von Fällen, die nicht gut zu matchen sind
- Anschließend normales Modell (z.B.) Regression mit *ursprünglichen* Variablen als Kontrollvariablen
- In Computersimulationen besser, schneller, flexibler als existierende Techniken
- What's next?

Software

- R: matchit (King und Freunde, Vielzahl von Prozeduren inkl. cem)
- Stata
  - psmatch2 (Nichols)
  - cem (King und Freunde)

Beispiel: Arbeitslosigkeit und Neonazismus



Beispiel: Arbeitslosigkeit und Neonazismus

```
. * Einfacher Mittelwertunterschied
. reg rexscale arbeitslos
```

Source	SS	df	MS
Model	45.8751965	1	45.8751965
Residual	33160.3744	2750	12.058318
Total	33206.2496	2751	12.0706106

Number of obs	= 2752
F(1, 2750)	= 3.80
Prob > F	= 0.0512
R-squared	= 0.0014
Adj R-squared	= 0.0010
Root MSE	= 3.4725

rexscale	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
arbeitslos	.3374962	.1730307	1.95	0.051	-.0017872	.6767795
_cons	7.064545	.0730125	96.76	0.000	6.92138	7.207709

Beispiel: Arbeitslosigkeit und Neonazismus

```
. * Propensity score matching
. * Arbeitslosigkeit vorhersagen
. logit arb male alter bildung ost
Iteration 0:   log likelihood =  -1576.285
Iteration 1:   log likelihood = -1492.1533
Iteration 2:   log likelihood = -1489.0855
Iteration 3:   log likelihood = -1489.0801
Iteration 4:   log likelihood = -1489.0801
Logistic regression                             Number of obs     =      3,398
                                                LR chi2(4)        =     174.41
                                                Prob > chi2       =     0.0000
Log likelihood = -1489.0801                     Pseudo R2         =     0.0553
```

arbeitslos	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
male	.0524309	.0929082	0.56	0.573	-.1296659	.2345277
alter	-.0311126	.002903	-10.72	0.000	-.0368024	-.0254229
bildung	-.2130927	.0617395	-3.45	0.001	-.3341	-.0920855
ost	.7822706	.0957406	8.17	0.000	.5946224	.9699188
_cons	.0382322	.2105266	0.18	0.856	-.3743924	.4508568

```
. predict propensity
(option pr assumed; Pr(arbeitslos))
(71 missing values generated)
```

Beispiel: Arbeitslosigkeit und Neonazismus

```
. * PS matching
. psmatch2 arb, pscore(propensity) out(rexscale)
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
rexscale	Unmatched	7.40368852	7.06865402	.335034505	.173877766	1.93
	ATT	7.40368852	6.86270492	.540983607	.283377311	1.91

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	2,214	2,214
Treated	488	488
Total	2,702	2,702

Beispiel: Arbeitslosigkeit und Neonazismus

```
. * Coarsened exact matching - Vorbereitung
. cem alter bildung ost male, treatment(arb)

Matching Summary:
-----------------
Number of strata: 163
Number of matched strata: 97

                0     1
      All  2870   599
  Matched  2290   597
Unmatched   580     2

Multivariate L1 distance: .1993368

Univariate imbalance:

               L1      mean       min       25%       50%       75%       max
  alter    .10521   -.23474         1        -1         0         0        -3
bildung   4.9e-16  -1.4e-14         0         0         0         0         .
    ost   2.8e-16  -2.2e-16         0         0         0         0         0
   male   4.4e-16  -1.1e-16         0         0         0         0         0
```

Beispiel: Arbeitslosigkeit und Neonazismus

```
. * Coarsened exact matching - Auswertung
. reg rexscale arb alter bildung ost male [pw=cem_weights]
(sum of wgt is   2.2628e+03)
```

Linear regression

Number of obs	= 2271
F(5, 2265)	= 38.43
Prob > F	= 0.0000
R-squared	= 0.0924
Root MSE	= 3.2718

rexscale	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
arbeitslos	.2685681	.1852787	1.45	0.147	-.0947657	.6319019
alter	-.0194236	.0067362	-2.88	0.004	-.0326333	-.0062139
bildung	-1.372526	.107456	-12.77	0.000	-1.583248	-1.161803
ost	.7271412	.1816283	4.00	0.000	.3709658	1.083317
male	.4850891	.1692314	2.87	0.004	.1532243	.8169539
_cons	10.04662	.4455583	22.55	0.000	9.172874	10.92036

Beispiel: Arbeitslosigkeit und Neonazismus

```
. * Normale Regression
. reg rexscale arb alter bildung ost male
```

Source	SS	df	MS
Model	2977.59488	5	595.518976
Residual	29708.327	2696	11.0194091
Total	32685.9219	2701	12.101415

Number of obs	= 2702
F(5, 2696)	= 54.04
Prob > F	= 0.0000
R-squared	= 0.0911
Adj R-squared	= 0.0894
Root MSE	= 3.3195

rexscale	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
arbeitslos	.2397376	.1709481	1.40	0.161	-.0954651	.5749403
alter	.0037525	.003964	0.95	0.344	-.0040202	.0115253
bildung	-1.216469	.0835622	-14.56	0.000	-1.380321	-1.052616
ost	.4351661	.13738	3.17	0.002	.1657854	.7045469
male	.4684388	.1279237	3.66	0.000	.2176003	.7192773
_cons	8.891038	.318505	27.91	0.000	8.266499	9.515577

Beispiele

Gilligan/Sergenti 2008: Do UN Interventions Cause Peace?

„Previous statistical studies of the effects of UN peacekeeping have generally suggested that UN interventions have a positive effect on building a sustainable peace after civil war. Recent methodological developments have questioned this result because the cases in which the United Nations intervened were quite different from those in which they did not. (. . .) We correct for the effects of nonrandom assignment with matching techniques on a sample of UN interventions in post-Cold-War conflicts . . .“

Beispiele

Mayer 2011 (JOP): Does Education Increase Political Participation?

„The consensus among scholars has long held that educational advancement causes greater political participation. (. . .) This recent work strongly suggests that selection mechanisms confound previous results, and it employs propensity score matching to argue that education has no effect. In this article I show how propensity score matching, . . . , introduces bias by creating poorly matched treatment and control groups. (. . .) I use genetic matching to create balanced treatment and control groups.“

Should I or Shouldn't I?

Ho et al. 2006

Matching methods, which offer the promise of causal inference with fewer assumptions, constitute one possible way forward, but crucial results in this fast-growing methodological literature are often grossly misinterpreted.

Should I or Shouldn't I?

Iacus et al. 2011

> *. . . widely used current methods, such as propensity score and Mahalanobis matching . . . [do] not guarantee any level of imbalance reduction in any given data set . . . . In any application a single use of these techniques can increase imbalance and model dependence by any amount.*

Should I or Shouldn't I?

Sekhon 2009

"Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive"

Zusammenfassung

- Matching: Selektives Entfernen von Fällen → Ex-post balancierte(re) Daten
- Hilft „extreme counterfactuals“ zu vermeiden (CEM)
- Funktioniert nur, wenn relevante X-Variablen erhoben
- Vor allem interessant für Quasi-Experimente (Evaluation)
- Keine Wunderwaffe (schade!)

Literatur für die beiden nächsten Sitzungen

- Berning (2018), *Strukturgleichungsmodelle* In: Wagemann C., Goerres A., Siewert M. (eds) Handbuch Methoden der Politikwissenschaft. Springer Reference Sozialwissenschaften. Springer VS, Wiesbaden, `http://www.doi.org/10.1007/978-3-658-16937-4_30-2`
- Arzheimer (2015), *Strukturgleichungsmodelle. Eine anwendungsorientierte Einführung*. Wiesbaden: Springer VS, `http://www.springer.com/us/book/9783658096083` bzw. `http://www.kai-arzheimer.com/beispiele-sem/`