Analyse von Paneldaten mit R

Hannes Weber
Universität Stuttgart, 14.06.2017

Kontakt:
hannes.weber@uni-tuebingen.de
hweber@startmail.com

Vorgehensweise

- „Verbale“ Hinführung auf diesen Folien (ohne Matrixalgebra, Beweise, etc.)

- Umsetzung in R mit Paket plm (siehe zugehöriges R-Skript)

Paneldaten mit R

1. Einleitung: Wozu Paneldaten?
2. Panel-Modelle
   a) Pooled OLS
   b) Lagged Dependent Variable
   c) Fixed Effects
   d) Random Effects
   e) First-difference
3. Ausblick

Eine klassische Fragestellung:

Wirkt sich der Bildungsstand auf den Demokratiegrad eines Landes aus?

Bildung und Demokratie

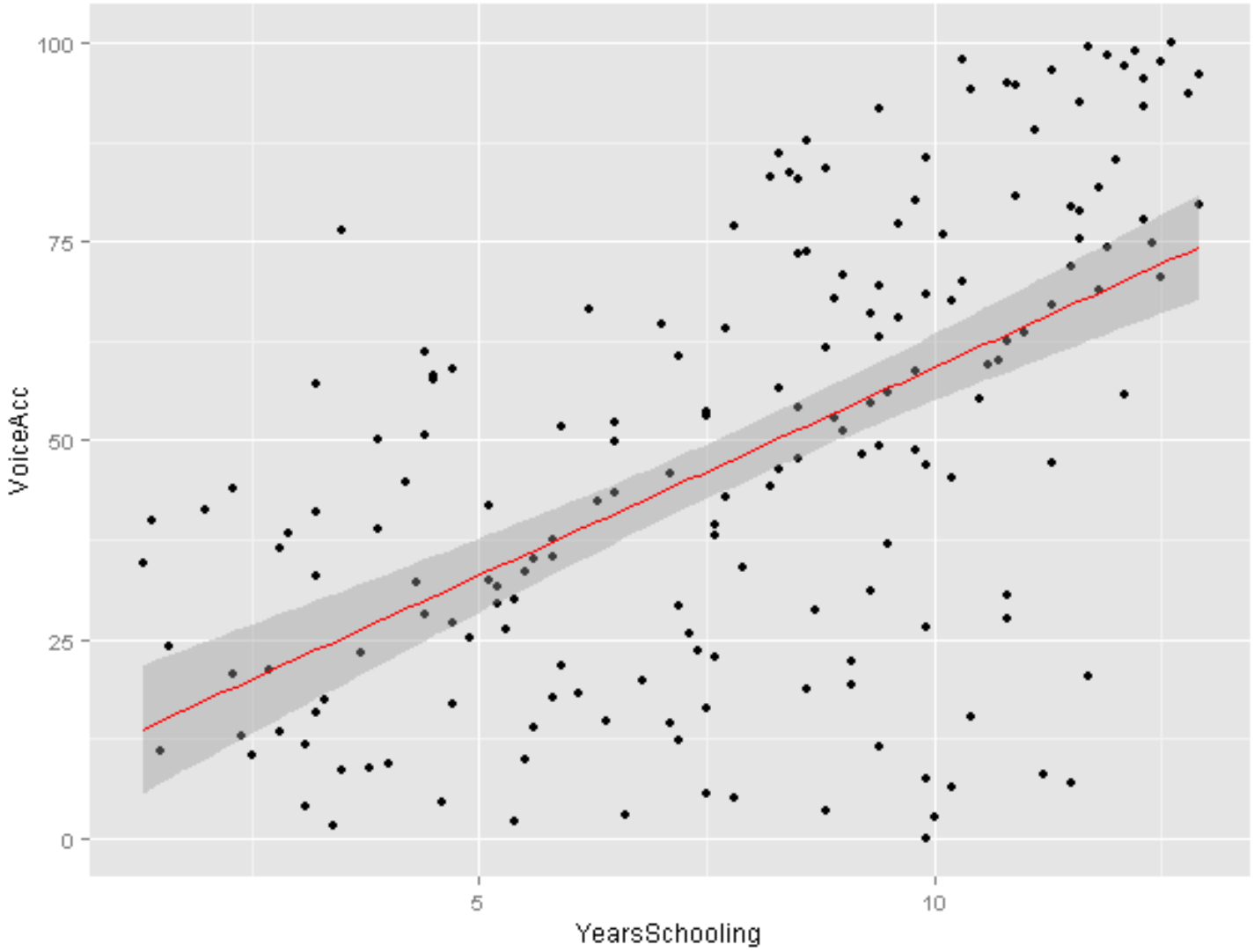
Querschnitt:

Bildung (durchschnittliche Anzahl Schuljahre) → Demokratie (Worldbank Voice & Accountability)

Daten:

ID	Country	VoiceAcc	FH	YearsSchooling	GDP_cap
1	AFGHANISTA	15.76	12	3.2	1945.50242
2	ALBANIA	54.68	6	9.3	10428.4559
3	ALGERIA	22.66	11	7.6	14167.3397
4	ANGOLA	16.75	11	4.7	7227.44077
5	ANTIGUA AN	67.98	4	8.9	21799.8001
6	ARGENTINA	58.62	4	9.8	12735.196
7	ARMENIA	30.54	9	10.8	8077.53329
8	AUSTRALIA	93.60	2	12.8	43901.5549
9	AUSTRIA	95.07	2	10.8	46164.9443
10	AZERBAIJAN	7.88	12	11.2	17515.6238

Querschnitt



Probleme mit Querschnittsdaten

- Repräsentativität in Zeitdimension
- Kausale Inferenz: Endogenität, unbeobachtete Heterogenität…
- Statistische Power bei oft kleiner Fallzahl

Paneldaten

ID	Country	Year	VoiceAcc	FH	YearsSchooling	GDP_cap
1	AFGHANISTAN	1996	1.92	14	1.86	NA
1	AFGHANISTAN	1997	1.20	14	1.92	NA
1	AFGHANISTAN	1998	0.48	14	1.98	NA
1	AFGHANISTAN	1999	0.96	14	2.04	NA
1	AFGHANISTAN	2000	1.44	14	2.1	NA
1	AFGHANISTAN	2001	4.09	14	2.18	NA
1	AFGHANISTAN	2002	6.73	12	2.26	NA
1	AFGHANISTAN	2003	12.02	12	2.34	NA
1	AFGHANISTAN	2004	13.94	11	2.42	940.476294
1	AFGHANISTAN	2005	13.46	10	2.5	1039.40824
1	AFGHANISTAN	2006	13.94	10	2.6	1095.65562
1	AFGHANISTAN	2007	16.35	10	2.8	1245.05922
1	AFGHANISTAN	2008	13.46	11	2.9	1283.04098
1	AFGHANISTAN	2009	8.53	12	3.1	1525.51704
1	AFGHANISTAN	2010	7.58	12	3.2	1629.16728
1	AFGHANISTAN	2011	9.39	12	3.2	1712.58872
1	AFGHANISTAN	2012	11.37	12	3.2	1933.39626
1	AFGHANISTAN	2013	13.27	12	3.2	1937.85596
1	AFGHANISTAN	2014	15.76	12	3.2	1945.50242
2	ALBANIA	1996	24.52	8	8.06	3245.55756
2	ALBANIA	1997	30.29	8	8.17	2982.9924
2	ALBANIA	1998	36.06	9	8.28	3419.77114

Brauche ich Paneldaten?

- Gibt es überhaupt Veränderungen über die Zeit? (Sonst einfach Vervielfachung der Fälle)
- Sind die Zeitintervalle sinnvoll? (Meist Jahre = willkürlich)
- Kosten der vergrößerten Informationsbasis: Fälle sind meistens nicht mehr voneinander unabhängig (serielle Korrelation der Fehlerterme in Regression).
  → Übliche Verfahren daher oft verzerrend.
  → Verfahren der Panelanalyse dagegen komplex!

Andere Beispiele

- Wie wirken sich Gesetzesänderungen/ demographischer Wandel/Bildungsexpansion… auf Kriminalitätsrate, Mietpreise, Arbeitslosigkeit... aus?
- Wie wirken sich Erfahrung von Arbeitslosigkeit/ Fortbildungen/ Geburt von Kindern… auf Arbeitszeit/ Einkommen/ politische Einstellungen… aus?
- Bekannte Datensätze z.B. SOEP, NEPS, GLES…

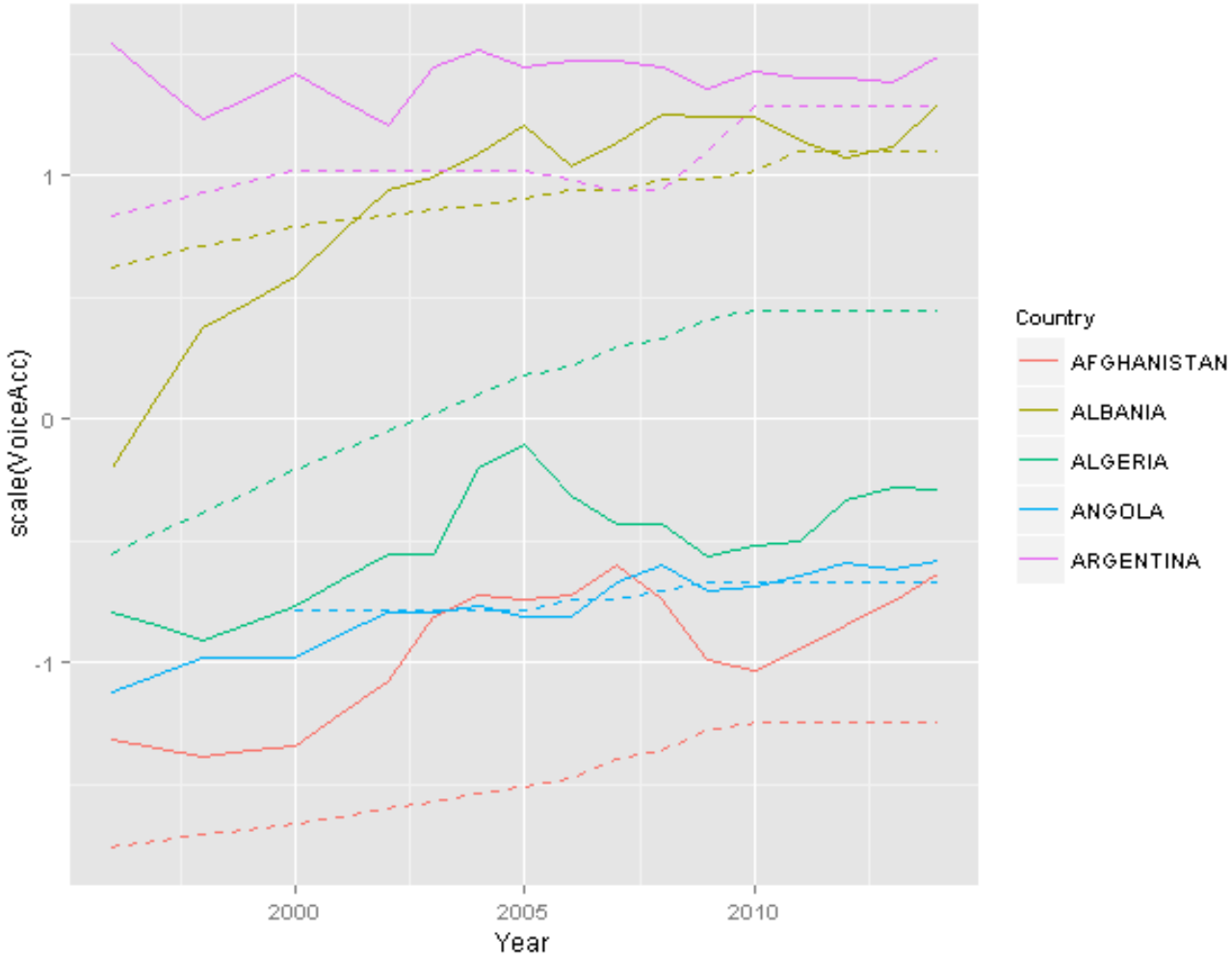
Bildung und Demokratie

Längsschnitt:

Bildung (---) vs.

Demokratie (__)

in fünf Ländern.



a) Pooled OLS

Wir haben 183 Länder über 19 Jahre (1996-2014) und daher statt 183 nun 183*19=3477 Fälle.

Wir rechnen eine normale OLS-Regression.

Wie angesprochen werden hier aber in vielen Fällen Grundvoraussetzungen für OLS verletzt.

Siehe R-Code **>>>**

b) OLS mit LDV

Wir nehmen die abhängige Variable zu t-1 (lagged dependent variable) als unabhängige Variable.

→Soll serielle Korrelation/ Pfadabhängigkeit/ „Matthäus-Effekt“ etc. kontrollieren.

→Andere UV erklären jetzt Veränderung in AV.

→Nachteil: Großteil der Varianz geht verloren, v.a. bei stark pfadabhängigen Variablen. Modell wird größtenteils tautologisch.

→Gemeinsame Ursachen von Demokratie zu t & t-1 ? (meist keine Markov-Ketten wie z.B. Wetter…)

b) OLS mit LDV (II)

→Lag kann auch erhöht werden, z.B. t-2, t-3, etc., wenn man vermutet, dass der Effekt (auch) nach längerer Zeit eintritt.

→Änderungsrate über größeren Zeitraum zu messen, lässt i.d.R. etwas mehr Varianz übrig.

→Theoretische Begründung/ empirischer Test hierbei nötig. Nachteil: Fälle gehen verloren.

→Auch UV können Lags enthalten (Autoregressive Distributed Lag (ADL)-Modell, Beck/Katz 2011).

c) Fixed Effects

„Fixed effects“ im Kontext von Panel-Modellen: wie Dummy-Variablen für alle Länder/Personen.

→Soll unbeobachtete „Eigenheiten“ eines Landes kontrollieren (Varianz auf „within“ beschränkt).

→z.B.: Aufgrund von nicht quantifizierbaren historischen, politischen, kulturellen o.a. Gründen hat Land A generell höheres Demokratieniveau (nicht nur „wegen Demokratie im Vorjahr“).

c) Fixed Effects (II)

- Nachteil: Keine wirkliche „Erklärung“ (Die schwedische Zeitreihe ist demokratischer als die somalische wegen dem Faktor Schweden bei der ersten und dem Faktor Somalia bei der zweiten Zeitreihe.)

→Es geht ebenfalls ein Großteil der Varianz verloren. Vor allem bei kleinem T (und großem n) problematisch!

→Bei konstanten/ langsam ändernden Variablen ungeeignet.

c) Fixed Effects (II)

Neben Länder- kann man auch Zeit-FEs aufnehmen.

→Kontrolle der unbeobachteten Eigenheiten eines Jahres (Periodeneffekte, Schocks…).

→Wiederum keine substantielle Erklärung. Hoher Tautologiegrad, hohe Multikollinearität.

<u>Trotzdem:</u> Wenn Effekt unter Fixed-Effects-Spezifikation bestehen bleibt (und diese theoretisch rechfertigbar ist), kann evtl. (vorsichtig) Kausalität unterstellt werden.

d) Random Effects

Anstatt n-1 Länder-Dummies wird ein länder-spezifischer Fehlerterm, der zufällig variiert, eingefügt. (Ähnlich wie „Random Intercept“/Mehrebenen-modelle)

→Sparsamer als FE-Modell. V.a. bei zeitunabhängigen/ kaum ändernden Faktoren besser.

→Aber auch nur dann unverzerrt, wenn die Länder-Abweichungen wirklich zufällig sind und nicht mit den UV korrelieren (z.B. mit Wohlstand).

→Hausman-Test, ob FE vorzuziehen ist.

e) First-Difference Model

- → Wenn X sich ändert, ändert Y sich entsprechend?
- → Ähnlich dem Modell mit LDV, aber hier werden von allen Variablen die Änderungsraten genommen.
- → Gegenüber FE vorzuziehen wenn Prozess nicht-stationär (z.B. Random Walk) ist. (Stationär = pendelt um den Langzeit-Durchschnitt. Bei nicht-stationären Zeitreihen besteht das Risiko von Scheinkorrelationen. Die 1. Ableitung einer nicht-stationären Variable ist häufig eine stationäre Zeitreihe – z.B. BIP, Bildung, Geburtenrate, etc. kann einem Trend unterliegen, aber die Änderungsrate evtl. nicht.)
- → Nachteil: Absolute Höhe der UV ohne Effekt – plausibel?

Mögliche Probleme (Auswahl)

- Heterogenität in Bezug auf Regressionskoeffizienten zwischen Ländern.
- Endogenität bei pfadabhängigen Variablen und begrenztem T.
- Geographische o.a. Beeinflussung (cross-sectional dependence).

→ Fortgeschrittene, kompliziertere Verfahren…

Zu unserem Beispiel:

Effekt von Bildung auf Demokratie häufig vorgefunden (z.B. Glaeser et al.).

- Acemoglu et al. (2005) wenden Difference-GMM an (Arellano/Bond 1991) und finden keinen Effekt.
- Bobba und Coviello (2007) sagen, dieser Schätzer sei unangebracht, wenden System-GMM an (Blundell/Bond 1998) und finden Effekt.
- Wie beurteilen wir als „anwendende“ Forscher das…?

Unsere Ergebnisse:

Sieben Modelle (3 OLS, 2 FE, 1 RE, 1 FD):

- 4 mal positiv und signifikanter Bildungseffekt
- 1 mal negativ und signifikant
- 3 mal nicht signifikant

→ Im (mutmaßlich) „besten“ Modell (FE) negativ bzw. insignifikant!

→ Keine (kurzfristigen) Demokratie-Effekte bei Ausbau des Bildungssystems.

Welches Verfahren?

Wie gesehen, kann die Wahl des Modells einen gewaltigen Einfluss auf die Ergebnisse haben.

- Theorie, Datenstruktur + Tests können Hinweise auf geeignete Modellierung geben.
- Mehrere in Frage kommende Modelle als Robustheits-Test rechnen.
- Es kann sein, dass keines der vorgestellten Verfahren wirklich angemessen ist…

Literatur

- Croissant/ Millo 2008: „Panel Data Econometrics in R: The plm Package“. Journal of Statistical Software.
- Beck/ Katz 2011: „Modelling Dynamics in Time-Series-Cross-Section Political Economy Data“. Annual Review of Political Science.
- Baltagi 2005: „Econometric Analysis of Panel Data“ (Wiley).
- Woolridge 2010: „Econometric Analysis of Cross Section and Panel Data“ (MIT Press).