Einführung III: Random Effects in Multiple-Country Repeated Cross-Sections Data

Kontext und Mehr-Ebenen-Datensätze

Was sind "Multiple-Country Repeated Cross-Sections Data"?

- Cross-Sections = kein Panel
- Repeated: Wiederholung gleicher Fragen, z.B. alle zwei Jahre
- Multiple-Country: gleiche Länder
- Beispiele: ESS, ISSP ...

Wo ist der Kontext?

- Land: alle Befragten aus *demselben Land* sind ähnlichen Einflüssen ausgesetzt

Wo ist der Kontext?

- Land: alle Befragten aus *demselben Land* sind ähnlichen Einflüssen ausgesetzt
- Jahr: alle *zum selben Zeitpunkt* befragten Personen sind ähnlichen Einflüssen ausgesetzt

Wo ist der Kontext?

- Land: alle Befragten aus *demselben Land* sind ähnlichen Einflüssen ausgesetzt
- Jahr: alle *zum selben Zeitpunkt* befragten Personen sind ähnlichen Einflüssen ausgesetzt
- Jahr $\times$ Land (country-year): alle innerhalb *derselben nationalen Welle befragten Personen sind ähnlichen Einflüssen* ausgesetzt

Wo ist der Kontext?

- Land: alle Befragten aus *demselben Land* sind ähnlichen Einflüssen ausgesetzt
- Jahr: alle *zum selben Zeitpunkt* befragten Personen sind ähnlichen Einflüssen ausgesetzt
- Jahr $\times$ Land (country-year): alle innerhalb *derselben nationalen Welle befragten Personen sind ähnlichen Einflüssen* ausgesetzt
- Zwei Ebenen oder drei Ebenen?
- Über-/Unter-/Gleichordnung?

Zur Erinnerung: Was ist ein random effect?

- Kontextspezifische Variation von intercept oder slope ...
- Wird nicht durch Serie von dummies abgebildet, sondern durch eine (Normal)Verteilung mit bestimmter Varianz
- Alle Einheiten im selben Kontext bekommen denselben random shock
- Sinnvoll/effizient, wenn es viele Kontexte gibt

Was ist ein Kontext / ein Kontextvariable?

- Übergeordnete Ebene, die untergeordnete Einheiten beeinflusst
- Mehrere Ebenen möglich (Hierarchie): Menschen in Gemeinden in Kreisen in Bundesländern ...
- Kontextvariable: Gemessene Eigenschaft des Kontextes
- *Innerhalb eines Kontextes konstant*

Umsetzung in Stata

- Kontextstruktur muss explizit angegeben werden
- Syntax ist verwirrend:
  mixed y x || obersteebene: || mittlereebene: ...

Was ist "cross-classification"?

- Manchmal keine klare Über-/Unterordnung von Kontexten
  - Schüler in Grundschulen/weiterführenden Schulen
  - Personen in Befragungsjahren / Befragungsländern
- Random effect für jede übergeordnete Ebene, die additiv auf untergeordnete Einheit wirken
- (Technisch: Zusammenfassung der Fälle; Bildung interner Indikatorvariablen für Gruppenzugehörigkeit, für die Effekte geschätzt werden)

Umsetzung in Stata

- Syntax ist noch etwas verwirrender wg interner Behandlung:
  mixed y x || _all: R.eineobereebene || _all: R.andereobereebene
- Oder kürzer und effizienter:
  mixed y x || _all: R.eineobereebene || andereobereebene:

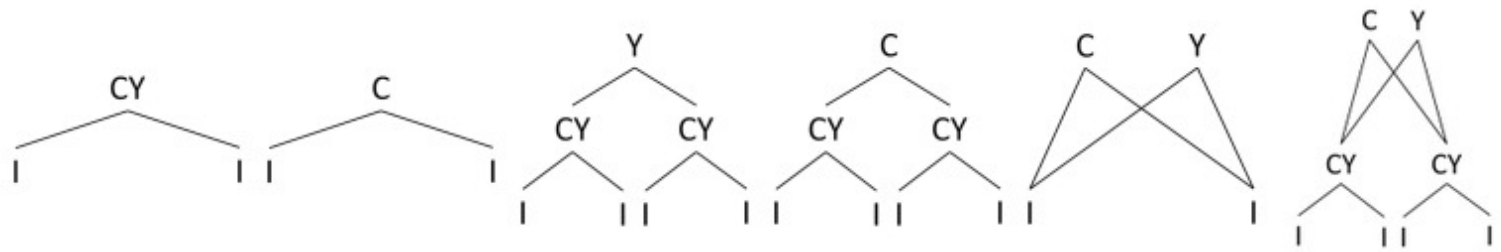
Anlaß

- Zahlreiche Veröffentlichung auf Basis von ESS-ähnlichen Designs
  - Standardisierte Querschnittsbefragungen
  - Relativ viele Länder (20-30)
  - Regelmäßig wiederholt (5-7 mal)
- Befragte sind nicht unabhängig voneinander, sondern "nested"
- Aber wie? CY, C, Y

Simulierte Daten

- 20 Länder
- 5 Wiederholungen (Befragungsjahre)
- 100 country years

Typische Vorgehensweisen (Modelle A-F)



Probleme?

- A+C (Befragte in country years):

Probleme?

- A+C (Befragte in country years):
- Beobachtung von (statischen) Ländervariablen nicht unabhängig (20 statt 100)

Probleme?

- A+C (Befragte in country years):
- Beobachtung von (statischen) Ländervariablen nicht unabhängig (20 statt 100)
- B+E (Befragte in countries oder countries $\times$ years)

Probleme?

- A+C (Befragte in country years):
- Beobachtung von (statischen) Ländervariablen nicht unabhängig (20 statt 100)
- B+E (Befragte in countries oder countries $\times$ years)
- CY variables nicht konstant innerhalb von Jahren/Ländern; werden wg Variation implizit als Individualvariablen betrachtet

Probleme II

- A + B + D

Probleme II

- A + B + D
- Ignorieren Variation auf Ebene der Jahre
- F theoretisch am besten, D als mögliche Alternative bei Konvergenzproblemen

Noch eine Komplikation

- Wirkt die Veränderung einer Kontextvariable (z.B. ALQ) *innerhalb* eines Landes genau so . . .
- . . . wie Unterschiede in derselben Variable *zwischen* Ländern?
- Mögliche Lösung:
  - Zentrieren am Mittelwert des Landes
  - Zentrierte Werte (within) +
  - Ländermittelwerte in das Modell packen (between)

Simulation

- Nehme bestimmte wahre Zusammenhänge an ("DGP")
- Generiere (mit zufälliger Variation) Datensätze
- Analysiere, zeichne Ergebnisse auf, studiere Verzerrungen unter Modellen A-F
- Koeffizienten, Standardfehler: bias und Varianz
- /In vielen Fällen führt das Fehlen von Random Effects zu verzerrten Schätzungen für fixed effects (Koeffizienten)

DGP 15

- 30 Länder, 25 Jahre
  - i individuals
  - j country years
  - k years
  - l countries
- Q_k : year level variable
- "Z varies at two levels, the country and the country-year level, with Z_l the between-country component and Z_{jkl} the within-country component."

$$y_{ijkl} = 1 \times X_{ijkl} - 1 \times Z_{jkl} + 1 \times Z_l + 1 \times Q_k + u_{jkl} + u_l + u_k + e_{ijkl} \quad (1)$$

Modell A

```
use dgp15, replace
ren X_jkl Z_jkl
ren X_k Q_k
ren X_l Z_l
mixed y X_ijkl Z_jkl Z_l Q_k  || cyear: , mle var nolog


Mixed-effects ML regression                     Number of obs     =      7,500
Group variable: cyear                           Number of groups  =        750

                                                Obs per group:
                                                              min =         10
                                                              avg =       10.0
                                                              max =         10

                                                Wald chi2(4)      =    2139.98
Log likelihood = -16679.041                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9765167   .0241491    40.44   0.000     .9291853    1.023848
       Z_jkl |  -.9618794   .0646534   -14.88   0.000    -1.088598   -.8351611
         Z_l |   .8708592   .0809152    10.76   0.000     .7122683     1.02945
         Q_k |   .9856599   .0766937    12.85   0.000      .835343    1.135977
       _cons |    1.19071   .0646094    18.43   0.000     1.064078    1.317342
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
cyear: Identity              |
                  var(_cons) |   2.644185     .15783      2.352253    2.972348
-----------------------------+------------------------------------------------
               var(Residual) |   4.091342   .0704253      3.955613    4.231728
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 2231.46       Prob >= chibar2 = 0.0000
```

Modell B

```
mixed y X_ijkl Z_jkl Z_l Q_k  || country: , mle var nolog


Mixed-effects ML regression                     Number of obs     =      7,500
Group variable: country                         Number of groups  =         30

                                                Obs per group:
                                                              min =        250
                                                              avg =      250.0
                                                              max =        250

                                                Wald chi2(4)      =    3190.50
Log likelihood = -17415.258                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9766817   .0280912    34.77   0.000     .9216241    1.031739
       Z_jkl |  -.9902477   .0294647   -33.61   0.000    -1.047998   -.9324979
         Z_l |   .8701323   .2013426     4.32   0.000     .4755081    1.264756
         Q_k |    .985453   .0340085    28.98   0.000     .9187976    1.052108
       _cons |   1.190044   .1604699     7.42   0.000     .8755286    1.504559
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
country: Identity            |
                  var(_cons) |   .7325184   .1953436      .4343355    1.235412
-----------------------------+------------------------------------------------
               var(Residual) |   6.003793   .0982382      5.814304    6.199457
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 759.02        Prob >= chibar2 = 0.0000
```

Modell C

```
mixed y X_ijkl Z_jkl Z_l Q_k  || year: || cyear: , mle var nolog

Mixed-effects ML regression                     Number of obs     =      7,500

------------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------
           year |         25        300      300.0        300
          cyear |        750         10       10.0         10
------------------------------------------------------------------

                                                Wald chi2(4)      =    2133.93
Log likelihood =  -16568.83                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9783854   .0240774    40.63   0.000     .9311945    1.025576
       Z_jkl |  -.9311254   .0541839   -17.18   0.000    -1.037324   -.8249268
         Z_l |   .8716799   .0667698    13.06   0.000     .7408136    1.002546
         Q_k |     .98588   .2456911     4.01   0.000     .5043343    1.467426
       _cons |   1.191458   .2049392     5.81   0.000     .7897843    1.593131
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
year: Identity               |
                  var(_cons) |   .9752408   .2954866      .5385274    1.766103
-----------------------------+------------------------------------------------
cyear: Identity              |
                  var(_cons) |     1.6699   .1094239      1.468633    1.898748
-----------------------------+------------------------------------------------
               var(Residual) |   4.091343   .0704254      3.955614    4.231729
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 2451.88                   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Modell D

```
mixed y X_ijkl Z_jkl Z_l Q_k  || country: || cyear: , mle var nolog

Mixed-effects ML regression                     Number of obs     =      7,500

------------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------
        country |         30        250      250.0        250
          cyear |        750         10       10.0         10
------------------------------------------------------------------

                                                Wald chi2(4)      =    2153.08
Log likelihood = -16618.714                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9769283   .0241118    40.52   0.000     .9296701    1.024187
       Z_jkl |  -.9873321   .0586654   -16.83   0.000    -1.102314   -.8723501
         Z_l |   .8702112   .2013263     4.32   0.000      .475619    1.264803
         Q_k |   .9854737   .0679007    14.51   0.000     .8523908    1.118557
       _cons |   1.190116   .1604863     7.42   0.000     .8755683    1.504663
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
country: Identity            |
                  var(_cons) |   .6606474   .1953918      .3700156    1.179558
-----------------------------+------------------------------------------------
cyear: Identity              |
                  var(_cons) |   1.984184   .1263362      1.751397    2.247912
-----------------------------+------------------------------------------------
               var(Residual) |   4.091341   .0704253      3.955612    4.231727
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 2352.11                   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Modell E

```
mixed y X_ijkl Z_jkl Z_l Q_k  || _all: R.year || country: , mle var nolog

Mixed-effects ML regression                     Number of obs     =      7,500

------------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------
           _all |          1      7,500    7,500.0      7,500
       country |         30        250      250.0        250
------------------------------------------------------------------

                                                Wald chi2(4)      =    2732.56
Log likelihood = -16762.877                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9870953   .0255995    38.56   0.000     .9369213    1.037269
       Z_jkl |  -.9571266   .0272905   -35.07   0.000    -1.010615   -.9036382
         Z_l |   .8711505    .203214     4.29   0.000     .4728583    1.269443
         Q_k |   .9856729    .247362     3.98   0.000     .5008522    1.470494
       _cons |   1.190952   .2609175     4.56   0.000     .6795628    1.702341
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
_all: Identity               |
                 var(R.year) |   1.042221   .3003204      .5924921    1.833316
-----------------------------+------------------------------------------------
country: Identity            |
                  var(_cons) |   .7507702   .2002802       .445077    1.266423
-----------------------------+------------------------------------------------
               var(Residual) |   4.973711   .0815144      4.816484     5.13607
------------------------------------------------------------------------------
LR test vs. linear model: chi2(2) = 2063.78                   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Modell F

```
mixed y X_ijkl Z_jkl Z_l Q_k  || _all: R.year || country: || cyear: , mle var nolog

Mixed-effects ML regression                     Number of obs     =      7,500

------------------------------------------------------------------
                |     No. of       Observations per Group
 Group Variable |     Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------
           _all |          1      7,500    7,500.0      7,500
       country |         30        250      250.0        250
         cyear |        750         10       10.0         10
------------------------------------------------------------------

                                                Wald chi2(4)      =    2153.71
Log likelihood = -16452.063                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      X_ijkl |   .9801526   .0239675    40.90   0.000     .9331771    1.027128
       Z_jkl |  -.9566861   .0449323   -21.29   0.000    -1.044752   -.8686205
         Z_l |   .8710508   .2031855     4.29   0.000     .4728146    1.269287
         Q_k |   .9856903   .2472953     3.99   0.000     .5010004     1.47038
       _cons |   1.190877   .2586632     4.60   0.000     .6839066    1.697848
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
_all: Identity               |
                var(R.year) |   1.013121   .3001321      .5668876    1.810613
-----------------------------+------------------------------------------------
country: Identity            |
                 var(_cons) |   .7163046   .2002246      .4141581     1.23888
-----------------------------+------------------------------------------------
cyear: Identity              |
                 var(_cons) |   .9440757   .0728857      .8115052    1.098303
-----------------------------+------------------------------------------------
              var(Residual) |   4.091351   .0704256      3.955621    4.231737
------------------------------------------------------------------------------
LR test vs. linear model: chi2(3) = 2685.41                   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Empfehlung Schmidt-Catran

- Modelle mit drei Ebenen und cross-classification auf der obersten Ebene
- CY-Variablen (z.B. Arbeitslosigkeit, Zuwanderung . . .) nach Möglichkeit in between- und within-Komponente aufteilen (Zentrieren am Länder-Mittelwert)
- Macht bisher fast keiner

Probleme der Simulationsstudie

- Zahl der Respondenten pro CY *sehr* klein - vermutlich irrelevant
- Daten idealisiert (Variablen unkorreliert) - Probleme in Realität evtl. noch größer?
- Zahl der Länder in Simulation und Realität relativ klein (30 oder weniger)
- Zahl der Jahre in Realität (und den meisten Simulationen) *sehr* klein (5-7)
- Modell F ist sehr komplex: Schätzung dauert lange + Konvergenzprobleme

"How many countries for multilevel modeling?" (Stegmueller 2013)

- "individual-level estimates are robust to small country-level sample sizes"
- "maximum likelihood estimates are sharply biased upward when the number of countries is fewer than 20"
- Standardfehler zu klein
- Cross-level interaction und Varianzkomponenten noch problematischer
- Bayesianische Methoden besser/konservativer

Random effect für Jahre?

- Fünf oder sieben Einheiten *extrem wenig* für random effect
- Evtl. besser fixed effects (dummies für Befragungsjahre)

Was nehmen wir mit?

Was nehmen wir mit?

- Mehr-Ebenen-Modellierung ist komplex, man kann sich leicht in den Fuß schießen
- Modelle sollten nicht unnötig kompliziert sein
- Immer mit einfachen Modellen anfangen, theoriegleitet ausbauen
- In der Politikwissenschaft/Soziologie ist die Methode (relativ) neu, Standards sind noch im Fluß