Multi Level Tools

Influential cases in multi level modeling

Katja Möhring & Alexander Schmidt

GK SOCLIFE, Universität zu Köln

Presentation at the German Stata User Meeting in Berlin, 1 June 2012

GK·SOCLIFE
Research Training Group – University of Cologne

Multi level tools - overview

- `mltl2scatter`: Scatter plots at upper levels
- `mlt2stage`: Calculates and stores values for two-stage regression and graphs.
- `mltcooksd`: Estimates the influence measures Cook's D and DFBETAs for the second level units in hierarchical mixed models.
- `mltshowm`: Postestimation command for `mltcooksd`, shows the models which caused Cook's D to be above the cuttoff point.
- `mltrsq`: Gives the Boskers/Snijders and the Bryk/Raudenbusch R-squared values for each level.

GK·SOCLIFE
Research Training Group – University of Cologne

Multi level tools - overview

- `mltl2scatter`: Scatter plots at upper levels
- `mlt2stage`: Calculates and stores values for two-stage regression and graphs.
- `mltcooksd`: Estimates the influence measures Cook's D and DFBETAs for the second level units in hierarchical mixed models.
- `mltshowm`: Postestimation command for `mltcooksd`, shows the models which caused Cook's D to be above the cuttoff point.
- `mltrsq`: Gives the Boskers/Snijders and the Bryk/Raudenbusch R-squared values for each level.

Index

1. Introduction
2. A research example from ASR
3. mltcooksd
4. mlt2stage
5. Outlook

1. Introduction

2. A research example from ASR

3. mltcooksd

4. mlt2stage

5. Outlook

GK·SOCLIFE
Research Training Group – University of Cologne

Influential cases in multi level modeling

- Multi level or hierarchical modeling originates from educational research, here typically pupils (level 1) nested in classes (level 2) are analyzed
- Increasingly used in social sciences to compare individuals nested in countries with data of international surveys
  1. Small number of upper level units
  2. No random sample at upper level

→ Problems of influential outliers concerning the direct impact of macro variables as well as their indirect "moderator" effect

1. Introduction
2. A research example from ASR
   - A research example from the American Sociological Review
   - Cook's D and DFBETAS
3. mltcooksd
   - `mltcooksd` description
   - Stata
4. mlt2stage
   - `mlt2stage` description
   - Stata
5. Outlook

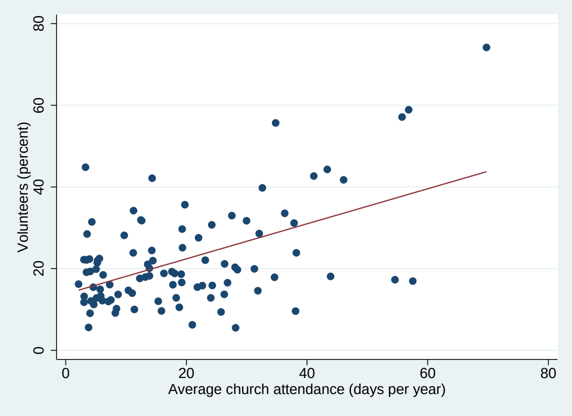
GK·SOCLIFE
Research Training Group – University of Cologne

Why should we consider outliers? A research example

Ruiter and De Graaf (2006): National Context, Religiosity, and Volunteering: Results from 53 Countries. *American Sociological Review*.

- Analysis of World Values Survey data with 53 countries
- Dependent variable *volunteering*
- Independent variable *national religious context*
- Conclusion: Average church attendance is significantly and positively related to volunteering

Van der Meer, Grotenhuis and Pelzer (2010) replicated their results

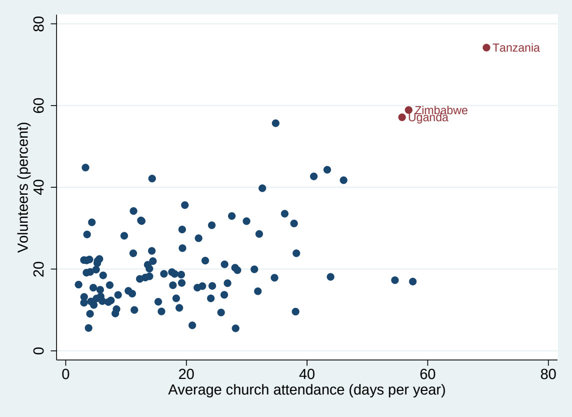


Notes: Data from von der Meer et. al. (2010) - own calculations.

Figure: Volunteering and Church Attendance

A research example from the American Sociological Review

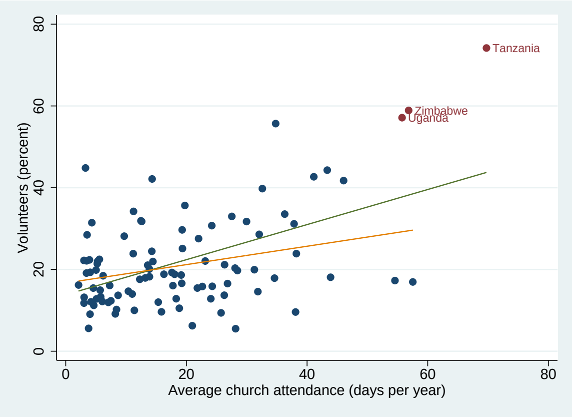
and showed ...



Notes: Data from von der Meer et. al. (2010) - own calculations.

Figure: Volunteering and Church Attendance - Revisited

...that African countries build exceptional and influential cases



Notes: Data from von der Meer et. al. (2010) - own calculations.

Figure: Volunteering and Church Attendance - Revisited I

Cook's D and DFBETAs: diagnostics for influential cases

Cook's D

- Measures the influence of one single (level-two) unit on all model parameters or a subset of parameters
- After non-hierarchical linear regressions it can be estimated from the hat matrix. Not possible after hierarchical mixed models
- However, we can estimate Cook's D empirically (Snijders and Berkhof 2008: 157ff.)

DFBETAs

- Measures the influence of one single level-two unit on a single parameter
- Again, we can only estimate this statistic empirically

DFBETAs

DFBETAS can be interpreted as the standardized difference in the estimated slope with and without unit j.

$$DFBETAS_{jZ} = \frac{\hat{\beta}_Z - \hat{\beta}_{(-j)Z}}{se(\hat{\beta}_{(-j)Z})}$$

, where $\hat{\beta}_Z - \hat{\beta}_{(-j)Z}$ is the difference between the estimated slopes of predictor Z. $\hat{\beta}_Z$ is the estimate in the full sample and $\hat{\beta}_{(-j)Z}$ is the estimated slope when unit j is excluded.

Cook's D

Fixed part of the model:

$$\underline{C}_j^F = \frac{1}{r}(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(-j)})'\underline{\hat{S}}_{F(-j)}^{-1}(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(-j)})$$

, with $r =$number of fixed parameters. $\underline{\hat{S}}_{F(-j)}$ is the variance-covariance matrix after unit j has been excluded.

Random part of the model:

$$\underline{C}_j^R = \frac{1}{p}(\underline{\hat{\eta}} - \underline{\hat{\eta}}_{(-j)})'\underline{\hat{S}}_{R(-j)}^{-1}(\underline{\hat{\eta}} - \underline{\hat{\eta}}_{(-j)})$$

, with $p =$number of random parameters.

Overall:

$$\underline{C}_j = \frac{1}{r+p}(r\underline{C}_j^F + p\underline{C}_j^R)$$

1. Introduction

2. A research example from ASR
   - A research example from the American Sociological Review
   - Cook's D and DFBETAS

3. mltcooksd
   - `mltcooksd` description
   - Stata

4. mlt2stage
   - `mlt2stage` description
   - Stata

5. Outlook

the mltcooksd ado

The mltcooksd command

- Calculates Cook's D after hierarchical mixed models (xtmixed and xtmelogit)
    - for the fixed part (C_j^F)
    - for the random part (C_j^R)
    - for the whole model (C_j)
- Gives DFBETAs for each fixed parameter in the model
- Compares the estimated values of Cook's D and DFBETAs to cutoff values proposed by Belsley et. al (1980) and reports those cases that have been detected as influential

mltcooksd syntax

Syntax

`mltcooksd [,`	
`fixed`	show estimates of C_j^F
`random`	show estimates of C_j^R
`keepvar(prefix)`	keep estimates in the data set
`counter`	estimate and show computing time
`graph`	show DFBETAs in box plot
`slabel]`	suppress labels in the output

the mltcooksd ado - an example

```
Mixed-effects ML regression                     Number of obs      =     21498
Group variable: Country                         Number of groups   =        22

                                                Obs per group: min =       441
                                                               avg =     977.2
                                                               max =      2345

                                                Wald chi2(4)       =    948.65
Log likelihood = -28233.225                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
  gr_incdiff |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.0329264   .0128818    -2.56   0.011    -.0581742   -.0076786
         age |   .0031901    .000379     8.42   0.000     .0024472     .003933
 respincperc |  -.0605727    .002245   -26.98   0.000    -.0649728   -.0561726
    socspend |   .0076906   .0121715     0.63   0.527    -.0161651    .0315463
       _cons |   3.086072   .2506038    12.31   0.000     2.594897    3.577246
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
Country: Identity            |
                  var(_cons) |   .0809317   .0246771      .0445222    .1471162
-----------------------------+------------------------------------------------
               var(Residual) |   .8058066   .0077762      .7907088    .8211928
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =  1984.18 Prob >= chibar2 = 0.0000
```

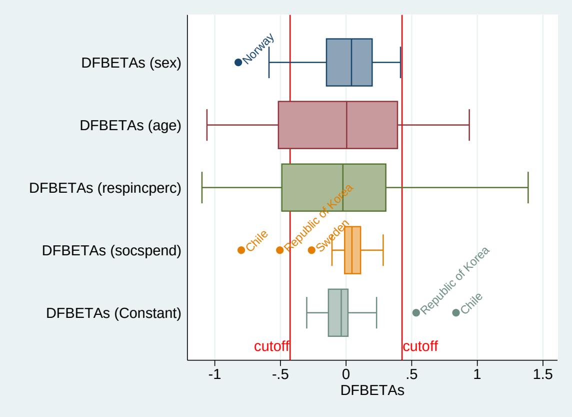
the mltcooksd ado - an example

```
.    mltcooksd, fixed random graph
Level 2 variable is Country

Calculating DFBETAs for the parameters of
 sex age respincperc socspend  _cons

Cutoff value for DFBETAs is
0.4264
Cutoff value for Cook's D is
0.1818

Level-two units with Cook's D above the cut off value:
  +------------------------------------------------------------+
  |                     L2ID    CooksD_f    CooksD_r     CooksD |
  |------------------------------------------------------------|
  |                 Portugal    .6616195    3.098742    1.35794 |
  |                Australia    .1800247    3.848549   1.228174 |
  |                    Chile    .6308343     2.56214   1.182636 |
  | United States of America    .1445634    1.989775   .6717668 |
...
  |           Czech Republic    .1419795    .3855572   .2115731 |
  |        Republic of Korea    .2438738    .0923624   .2005848 |
  |                  Hungary    .0475411    .5732102   .1977323 |
  +------------------------------------------------------------+
```

the mltcooksd ado - an example

```
Level-two units with DFBETAs above cut off value:

+-------------------------------------------------------------------------------+
|                    L2ID   DFB_sex   DFB_age   DFB_re~c   DFB_so~d   DFB_cons |
|-------------------------------------------------------------------------------|
|                Portugal    0.0335   -0.9608     1.3871     0.1956    -0.1090 |
|               Australia    0.0871   -0.5155    -0.7639     0.1678    -0.1420 |
|                   Chile   -0.0699   -0.5185     1.3678    -0.7983     0.8374 |
| United States of America    0.2718   -0.5825    -0.4614     0.2827    -0.2996 |
|                   Spain   -0.0439   -1.0599     1.3739     0.0566     0.0000 |
|             New Zealand   -0.0606   -0.2903    -0.9856     0.0943    -0.1344 |
|             Netherlands    0.2113    0.8566    -1.0978    -0.0106    -0.0187 |
|                   Japan    0.2648    0.3343     0.5692     0.0468    -0.1422 |
|                  France    0.0492    0.9389    -0.2171     0.0426    -0.0908 |
|                  Sweden   -0.1991    0.5152    -0.9410    -0.2625     0.2324 |
|                  Norway   -0.8209    0.5698    -0.4893    -0.0012     0.0144 |
|                  Canada    0.4149    0.1610    -0.8004     0.0782    -0.0931 |
|          Czech Republic    0.1199    0.7360     0.1394     0.0545    -0.2036 |
|       Republic of Korea    0.0035   -0.5778     0.7074    -0.5044     0.5339 |
|                 Finland   -0.5870    0.3408    -0.3167     0.0270    -0.0152 |
+-------------------------------------------------------------------------------+
```

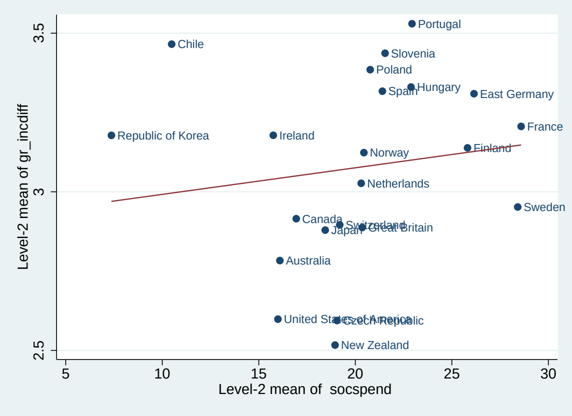
the `mltcooksd` ado - an example



Notes: Data from the ISSP - output of the `mltcooksd graph` option.

Figure: Distribution of DFBETAS

What’s wrong with Chile and Korea?



Notes: Data from the ISSP - plot produced with `mltl2scatter`.

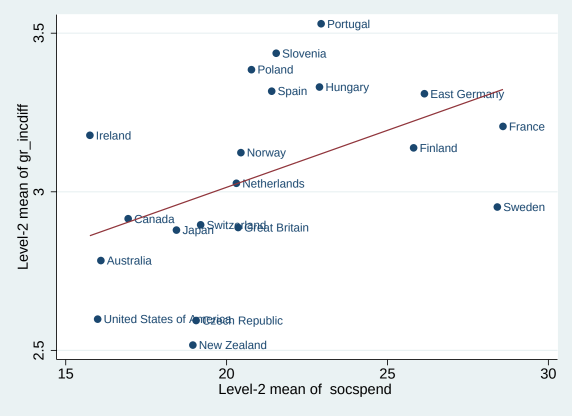
Figure: Social Spending and Support for Redistribution

the mltcooksd ado - an example

Chile and Korea excluded:

```
Mixed-effects ML regression                     Number of obs      =     19433
Group variable: Country                         Number of groups   =        20

                                                Obs per group: min =       441
                                                               avg =     971.6
                                                               max =      2345

                                                Wald chi2(4)       =    984.00
Log likelihood = -25784.384                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
  gr_incdiff |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.0321106   .0137022    -2.34   0.019    -.0589663   -.0052548
         age |   .0036384   .0004015     9.06   0.000     .0028515    .0044254
 respincperc |  -.0656586   .0024008   -27.35   0.000     -.070364   -.0609531
    socspend |   .0356661   .0150762     2.37   0.018     .0061173    .0652149
       _cons |   2.468119   .3224328     7.65   0.000     1.836162    3.100076
------------------------------------------------------------------------------

* Random part omitted
```

the `mltcooksd` ado - an example



Notes: Data from the ISSP - plot produced with `mltl2scatter`.

Figure: Social Spending and Support for Redistribution

1. Introduction
2. A research example from ASR
   - A research example from the American Sociological Review
   - Cook's D and DFBETAS
3. mltcooksd
   - mltcooksd description
   - Stata
4. mlt2stage
   - mlt2stage description
   - Stata
5. Outlook

GK·SOCLIFE
Research Training Group – University of Cologne

The two-stage approach

- Two-stage approach to model cross-level interactions in multi level data (Achen 2005; Gelman 2005)
- Coefficients from single country regressions are used for macro level estimations, e.g. two-stage regression

First-stage regression specification is:

$$y_j = X_j \beta_j + u_j \; (j = 1, ..., m) \tag{1}$$

Second-stage regression specification is:

$$\beta^1 = z\gamma + \nu \tag{2}$$

- Two-stage graphs to examine the moderator effect of a macro variable and detect potentially influential cases

the mlt2stage ado

The mlt2stage command

- Calculates and stores the coefficients of country separate linear and logistic regressions
- Plots the estimated values against a macro level indicator

mlt2stage syntax

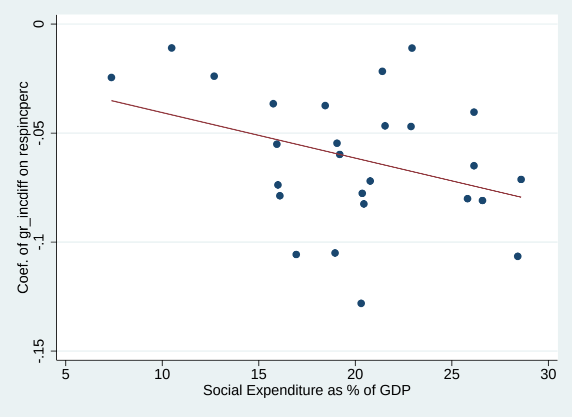
Syntax

`mlt2stage ,`	
`l2id(varname)`	define level 2 identifier
`[vname(prefix)`	define variable name for estimates in the data set
`logit`	calculate logistic model
`graph(varname)`	plot level 1 coefficients over level 2 variable
`all]`	store coefficients for all variables in the model

the mlt2stage ado - an example

```
. mlt2stage gr_incdiff respincperc age sex, l2id(Country) graph(socspend)

command:regress
graph:socspend
Two stage calculated for the dependent variable gr_incdiff
and the main explanatory variable respincperc
with the independent variables  respincperc age sex

Level 2 variable is Country

-----------------------------------------
Country                  | mean(coef_g~c)
-------------------------+---------------
               Australia |      -.0787574
                  Canada |      -.1056875
                   Chile |      -.0109568
          Czech Republic |      -.0546495
                 Denmark |      -.0809449
                 Finland |      -.0801003
                  France |      -.0712633
                 Hungary |      -.0470008
                 Ireland |      -.0365443
                  Israel |      -.0550879
                   Japan |      -.0374054
       Republic of Korea |       -.024505
                  Latvia |      -.0239054
             Netherlands |      -.1280941
             New Zealand |       -.105004
                  (...)
-----------------------------------------
```

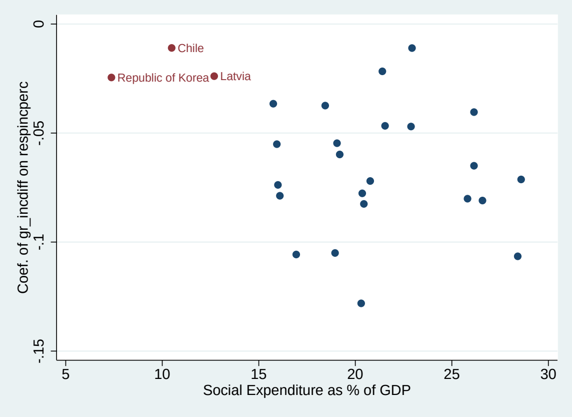
the mlt2stage ado - an example



Notes: Data from the ISSP - output of the `mlt2stage graph` option.

Figure: Distribution of country coefficients over social spending

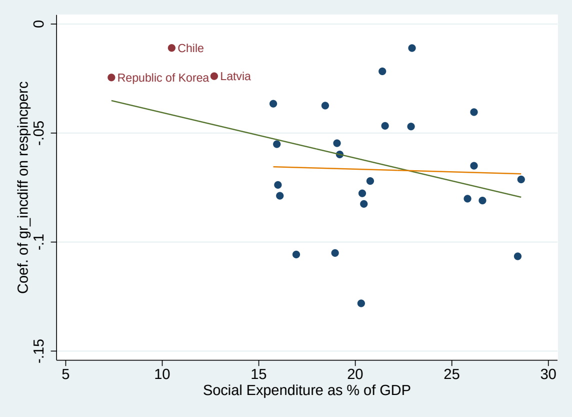
the `mlt2stage` ado - an example



Notes: Data from the ISSP - output of the `mlt2stage graph` option.

Figure: Distribution of country coefficients over social spending

the mlt2stage ado - an example



Notes: Data from the ISSP - output of the `mlt2stage graph` option.

Figure: Distribution of country coefficients over social spending

1. Introduction
2. A research example from ASR
3. mltcooksd
4. mlt2stage
5. Outlook

more multi level tools...

`mltl2scatter`, `mlt2stage`, `mltcooksd`, `mltshowm`, `mltcooksd`, `mltrsq` ...

- Extension of ados for three or more levels
- Ado to compare multi level and country FE results
- Ado to calculate model fit values for logistic multi level models

GK SOCLIFE
Research Training Group – University of Cologne

Comments & questions welcome!

→ moehring@wiso.uni-koeln.de, www.katjamoehring.de

→ alexander.schmidt@wiso.uni-koeln.de, www.alexanderwschmidt.de

GK·SOCLIFE Research Training Group – University of Cologne

References

Achen, Christopher H. (2005): Two-Step Hierarchical Estimation: Beyond Regression Analysis, in: *Political Analysis* 13(4): 447-456, doi:10.1093/pan/mpi033.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. (1980): Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley.

Gelman, Andrew (2005): Two-Stage Regression and Multilevel Modeling: A Commentary, in: *Political Analysis* 13(4): 459-461, doi: 10.1093/pan/mpi032.

Ruiter, Stijn and Nan Dirk De Graaf (2006): National Context, Religiosity, and Volunteering: Results from 53 Countries. *American Sociological Review* 71, pp. 191210.

Snijders, Tom A. B. and Johannes Berkhof (2008): Diagnostic Checks for Multilevel Models. pp. 457514 in Handbook of Multilevel Analysis, edited by J. De Leeuw and E. Meijer. New York: Springer.

Van der Meer, Tom, Manfred Te Grotenhuis and Ben Pelzer (2010): Influential Cases in Multilevel Modeling: A Methodological Comment. *American Sociological Review* 75: pp. 173-178.