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
Multi level tools - overview

- `mlt12scatter`: 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 cutoff point.
- `mltrsq`: Gives the Boskers/Snijders and the Bryk/Raudenbusch R-squared values for each level.
Multi level tools - overview

- **mlt12scatter**: 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 mltcooksdl

4 mlt2stage

5 Outlook
1 Introduction

2 A research example from ASR

3 mltcooksd

4 mlt2stage

5 Outlook
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
Why should we consider outliers? A research example


- 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.

**Figure:** Volunteering and Church Attendance
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.

![Volunteering and Church Attendance - Revisited I](image)

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

**Figure:** Volunteering and Church Attendance - Revisited I
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:

\[
C_j^F = \frac{1}{r} (\hat{\beta} - \hat{\beta}_{(-j)})' \hat{S}_{F(-j)^{-1}} \hat{\beta} - \hat{\beta}_{(-j)})
\]

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

Random part of the model:

\[
C_j^R = \frac{1}{p} (\hat{\eta} - \hat{\eta}_{(-j)})' \hat{S}_{R(-j)^{-1}} (\hat{\eta} - \hat{\eta}_{(-j)})
\]

, with \( p \) = number of random parameters.

Overall:

\[
C_j = \frac{1}{r + p} (r C_j^F + p C_j^R)
\]
Introduction

A research example from ASR

A research example from the American Sociological Review

Cook’s D and DFBETAS

mltcooksd

mltcooksd description

Stata

mlt2stage

mlt2stage description

Stata

Outlook
The `mltcooks`d 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
### Syntax

```
mltcooksdf [,
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
Group variable: Country

Number of obs = 21498
Number of groups = 22

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

Wald chi2(4) = 948.65
Prob > chi2 = 0.0000

Log likelihood = -28233.225

| 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
<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
</table>
| 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
. 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 |
+---------------------------------+-----------------+-----------------+-----------------+
Level-two units with DFBETAs above cut off value:

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

![Box plots of DFBETAs](image)

**Figure:** Distribution of DFBETAS

Notes: Data from the ISSP - output of the `mltcooksd` graph option.
What’s wrong with Chile and Korea?

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

Figure: Social Spending and Support for Redistribution
Chile and Korea excluded:

Mixed-effects ML regression
Group variable: Country

Number of obs = 19433
Number of groups = 20

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

Wald chi2(4) = 984.00
Prob > chi2 = 0.0000

Log likelihood = -25784.384

| 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   |
| respincerc | -.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 **mltcooks**d ado - an example

![Social Spending and Support for Redistribution](image)

Notes: Data from the ISSP - plot produced with mlt2scatter.

**Figure:** Social Spending and Support for Redistribution
Introduction

A research example from ASR

A research example from the American Sociological Review
Cook’s D and DFBETAS

mltcooks
d description
Stata

mlt2stage
d description
Stata

Outlook
The two-stage approach

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

First-stage regression specification is:

$$y_j = X_j \beta_j + u_j \ (j = 1, \ldots, m) \quad (1)$$

Second-stage regression specification is:

$$\beta^1 = z \gamma + \nu \quad (2)$$

- Two-stage graphs to examine the moderator effect of a macrovariable and detect potentially influential cases
The `mlt2stage` command

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

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

A research example from ASR

Stata

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

-----------------------------------------
<table>
<thead>
<tr>
<th>Country</th>
<th>mean(coef_g^-c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>-.0787574</td>
</tr>
<tr>
<td>Canada</td>
<td>-.1056875</td>
</tr>
<tr>
<td>Chile</td>
<td>-.0109568</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>-.0546495</td>
</tr>
<tr>
<td>Denmark</td>
<td>-.0809449</td>
</tr>
<tr>
<td>Finland</td>
<td>-.0801003</td>
</tr>
<tr>
<td>France</td>
<td>-.0712633</td>
</tr>
<tr>
<td>Hungary</td>
<td>-.0470008</td>
</tr>
<tr>
<td>Ireland</td>
<td>-.0365443</td>
</tr>
<tr>
<td>Israel</td>
<td>-.0550879</td>
</tr>
<tr>
<td>Japan</td>
<td>-.0374054</td>
</tr>
<tr>
<td>Republic of Korea</td>
<td>-.0245050</td>
</tr>
<tr>
<td>Latvia</td>
<td>-.0239054</td>
</tr>
<tr>
<td>Netherlands</td>
<td>-.1280941</td>
</tr>
<tr>
<td>New Zealand</td>
<td>-.105004</td>
</tr>
<tr>
<td>(....)</td>
<td></td>
</tr>
</tbody>
</table>
-----------------------------------------
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

![Graph showing the distribution of country coefficients over social spending.](image)

**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 mltcooksds

4 mlt2stage

5 Outlook
more multi level tools...

mltl2scatter, mlt2stage, mltcooksdl, mltshowm, mltcooksdl, 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
Comments & questions welcome!

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


