Twostep multilevel analysis using Stata

Johannes Giesecke [1] Ulrich Kohler [2]

[1] Humboldt University Berlin
Department of Social Sciences

[2] University of Potsdam
Faculty of Economic and Social Sciences

2021 German Stata Conference
June 25th 2021
Hosted by the University of Potsdam

Contents

Aim of presentation

- Introducing Stata command `twostep`.
- `twostep` is a bundle of programs to ease multilevel analyses with the “twostep approach”.
- Main purpose: Convenient interactive commands to be used as compagnion for `mixed`.

Statistical Background

Step 1 Estimate a parameter of interest separately for each category of a cluster level identifier.

Step 2 Analyse the estimates in cluster level data.

The second step may (or should) respect that the outcome is a an estimate.

Typical usages

1. Superior to the one-step approach if the numbers of observation on the cluster level is small (e.g. international comparisons; see Achen, 2005; Heisig et al., 2017).
2. To check assumptions of the one-step approach.
3. Exploratory data analysis of multilevel data.

`twostep` has in mind the second and third usage: Similar short commands for various related methods.

Overview

`twostep` is both, a prefix command (Syntax 1), and a standalone command (Syntax 2). The main purpose is the prefix command.
A simplified syntax diagram is:

```
twostep cluster_id:  cmd-1 || cmd-2
```

`cluster_id` is the identifier for the cluster level.
`cmd-1` is an estimation command for the unit level data, or `unitcpr`.
`cmd-2` is one of the cluster level commands `clustercpr`, `dot`, `edv`, `mk2nd`, `unitregby` or an arbitrary Stata command with standard syntax ("fallback mode").

Running example

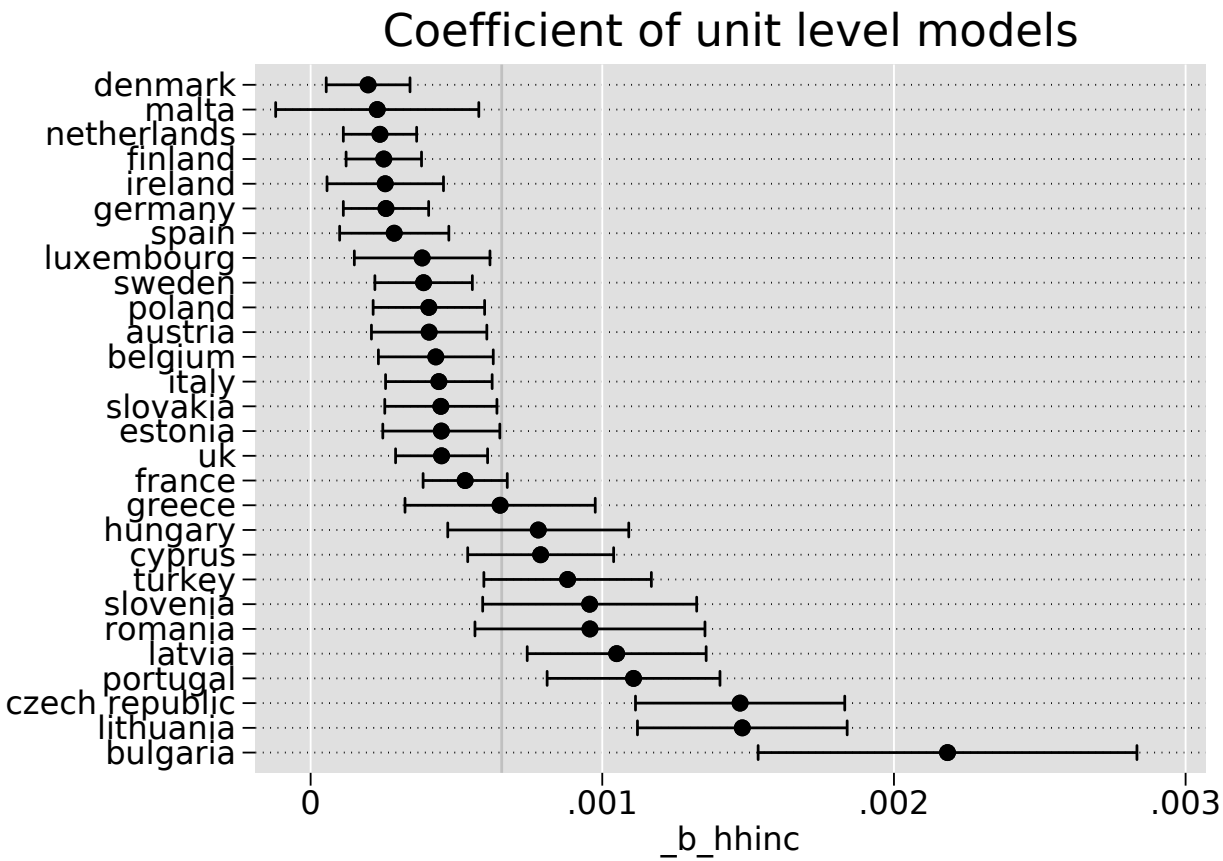
- Data: European Quality of Life Survey (EQLS), Round 4 from 2016
- Step 1: Regression of life satisfaction on household income and gender.
- Step 2: Analyse estimated coefficient of household income on country characteristics.

Contents

Dot-chart of unit level estimates

The cluster level command `dot` is being used to create horizontally labeled dot charts of coefficients with confidence intervals as shown by Bowers and Drake (2005, Fig. 1):

```
. twostep cntry: reg lsat hhinc i.sex || dot _b_hhinc
```



Variants

- Order of clusters can be controlled by a cluster variable:

```
. twostep cntry: reg lsat hhinc i.sex || dot _b_hhinc eu15d hdirank
```

- Other stats of unit level models can be used:

```
. twostep cntry, stats(r2) : reg lsat hhinc i.sex || dot _stat_r2
```

- Unit level model may be changed:

```
. twostep cntry: logit lsatd hhinc i.sex || dot _b_hhinc
```

- Standard graph options can be used to control look & feel of the graph:

```
. twostep cntry: reg lsat hhinc i.sex || dot _b_hhinc,
  scopts(mcolor(red) ms(S)) ciopts(lwidth(0))
```

Contents

Estimated Dependent Variable Regression

The cluster level command `edv` fits the "Estimated Dependent Variable Model" (EDV model) as described by Lewis and Linzer (2005).

```
. twostep cntry: reg lsat hhinc i.sex || edv _b_hhinc hdirank
(sum of wgt is 178,357,816.02346)
```

Source	SS	df	MS	Number of obs	=	28
				F(1, 26)	=	11.54
Model	1.5475e-06	1	1.5475e-06	Prob > F	=	0.0022
Residual	3.4864e-06	26	1.3409e-07	R-squared	=	0.3074
				Adj R-squared	=	0.2808
Total	5.0339e-06	27	1.8644e-07	Root MSE	=	.00037

_b_hhinc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hdirank	.0000116	3.41e-06	3.40	0.002	4.58e-06	.0000186
_cons	.0002944	.0001172	2.51	0.019	.0000534	.0005353

```
Sampling Variance Proportion =   .8
```

Variants

The EDV model weights the cluster level regression by an inverse of the uncertainty of the model estimates. One can use any of the following `methods()` for the weighting:

- `ols` (no weights),
- `wls`,
- `borjas`,
- `fgls1` (default),
- `fgls2`.

```
. twostep cntry: reg lsat hhinc i.sex || edv _b_hhinc hdirank, method(wls)
. twostep cntry: reg lsat hhinc i.sex || edv _b_hhinc hdirank, method(ols)
. twostep cntry: reg lsat hhinc i.sex || edv _b_hhinc hdirank, method(borjas)
```

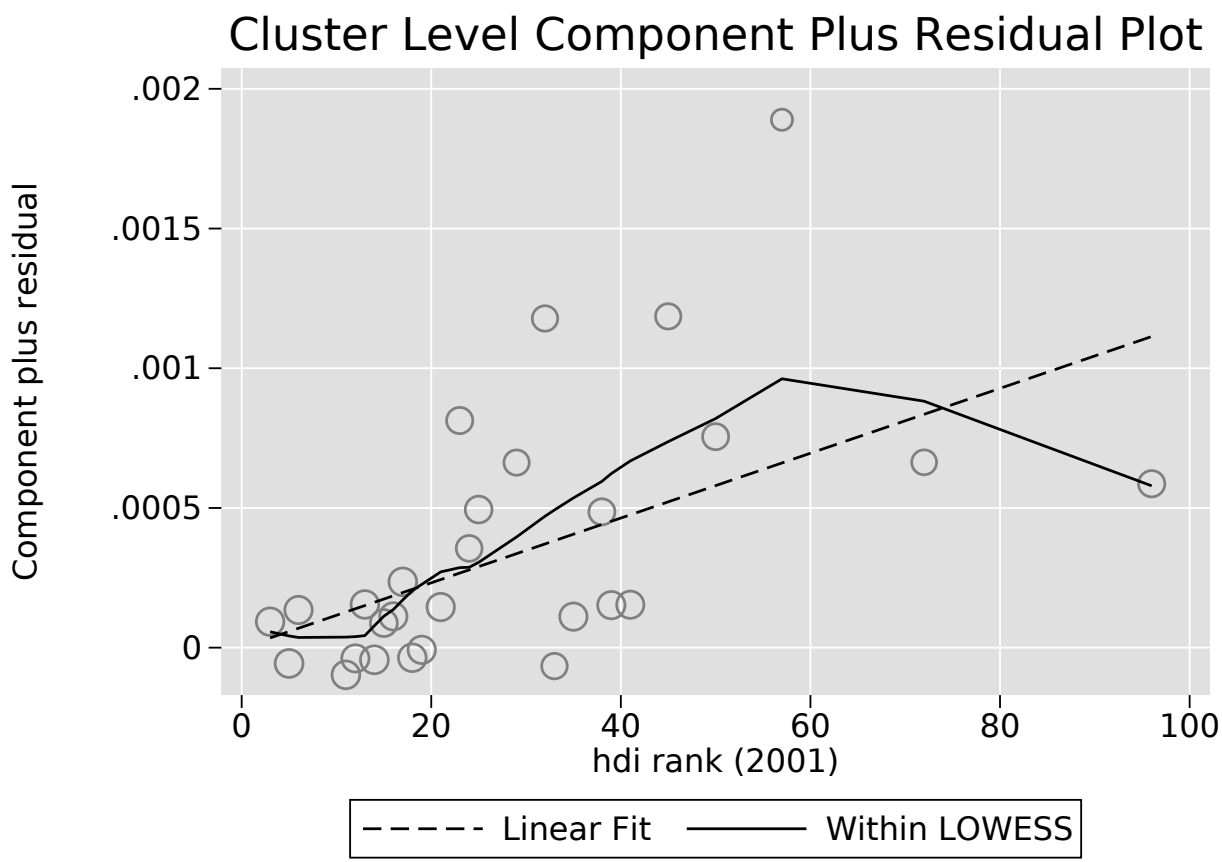
See Lewis and Linzer (2005) and Borjas and Sueyoshi (1994) for a discussion of these methods.

Contents

Cluster Level CPR Plot

The cluster level command `clustercpr` shows component plus residual plots for the EDV regression models:

```
. twostep cntry: reg lsat hhinc i.sex || clustercpr _b_hhinc hdirank
```



Variants

- More than just one cluster level covariate

```
. twostep cntry: reg lsat hhinc i.sex
  || clustercpr _b_hhinc hdirank corrupt
```

- Any of `method()` of EDV model

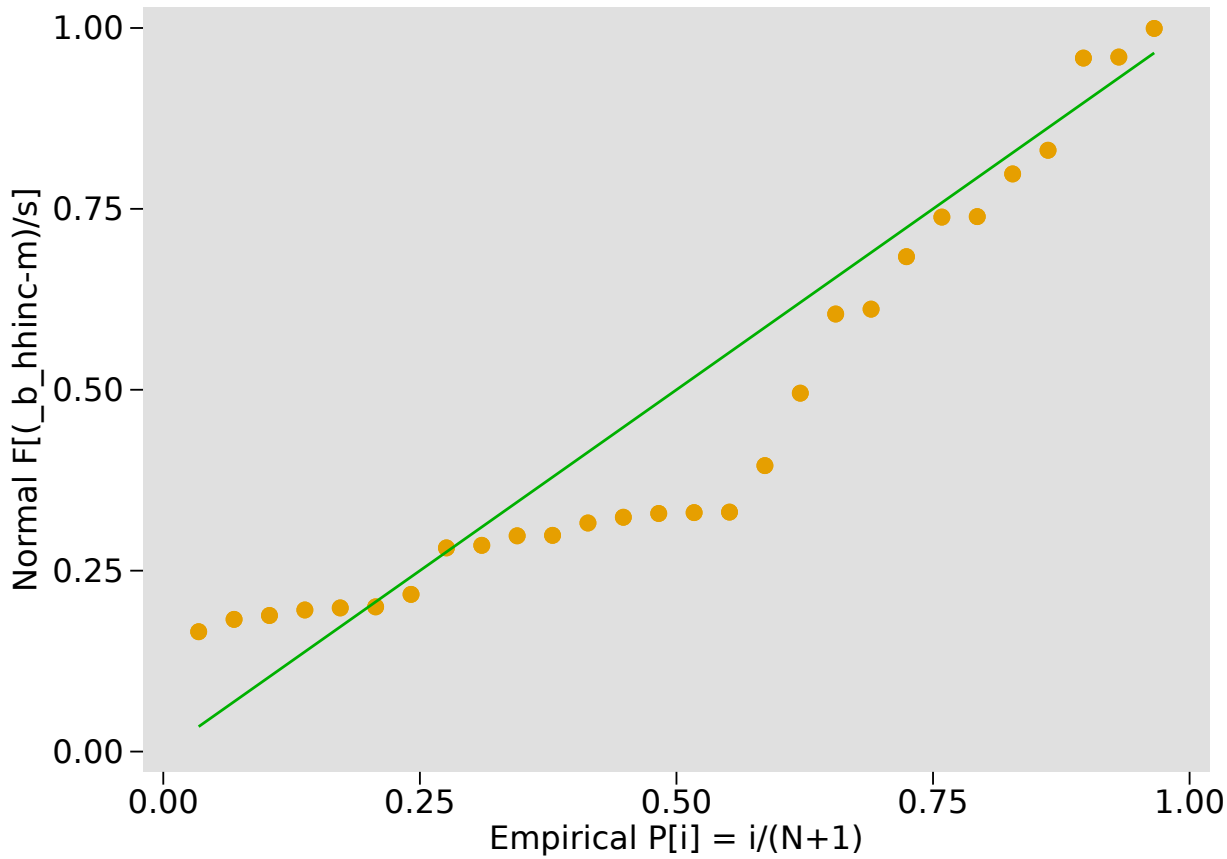
```
. twostep cntry: reg lsat hhinc i.sex
  || clustercpr _b_hhinc hdirank, method(ols)
```

Contents

Distributional diagnostic plots

All plots described in `help diagnostic plots` may be invoked as `cmd-2` (except `qqplot`):

```
. twostep cntry: reg lsat hhinc i.sex || pnorm _b_hhinc
```



Variants

The distributional diagnostic plots are already part of `twostep`'s fallback mode. The fallback mode allows to invoke arbitrary Stata commands:

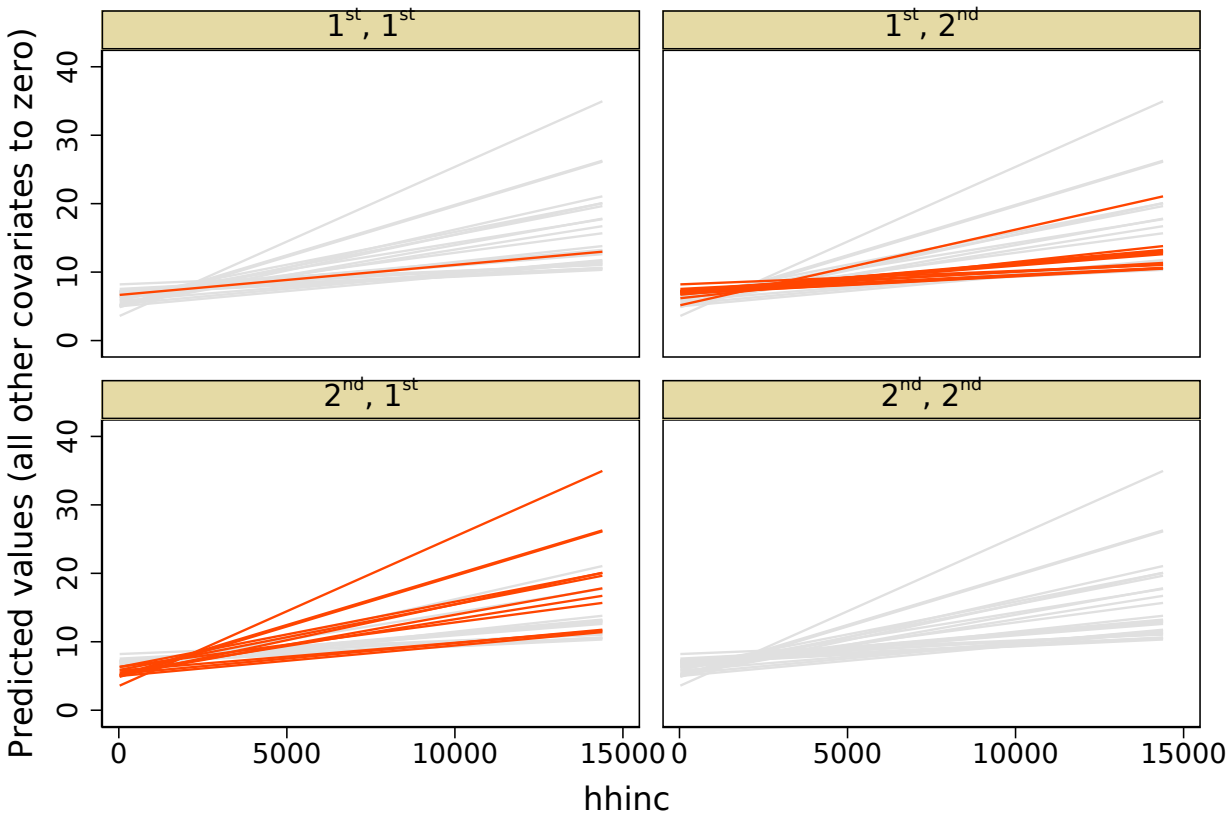
```
. twostep cntry: reg lsat hhinc i.sex || kdensity _b_hhinc
. twostep cntry, stats(r2): reg lsat hhinc i.sex || kdensity _stat_r2
. twostep cntry: reg lsat hhinc i.sex || scatter _b_hhinc hdirank
. twostep cntry: reg lsat hhinc i.sex || reg _b_hhinc hdirank
```

Contents

Unitregby plot

The cluster level command `unitregby` creates a plot of unit level regression slopes by groups defined using cluster level variables. The graph was proposed by Bowers and Drake (2005).

```
. twostep cntry: reg lsat hhinc i.sex
  || unitregby _b_hhinc hdirank corrupt, scheme(s1color)
```



Variants

- Grouping of cluster level variables can be controlled by option `nquantiles(#)` (default 2):

```
. twostep cntry: reg lsat hhinc i.sex
  || unitregby _b_hhinc hdirank, nq(4)
```

- Cluster level variables can be declared discrete with option `discrete()`:

```
. twostep cntry: reg lsat hhinc i.sex
  || unitregby _b_hhinc hdirank eu15d, discrete(eu15d)
```

- Groups of unit level variables can be added with option `unitby()`:

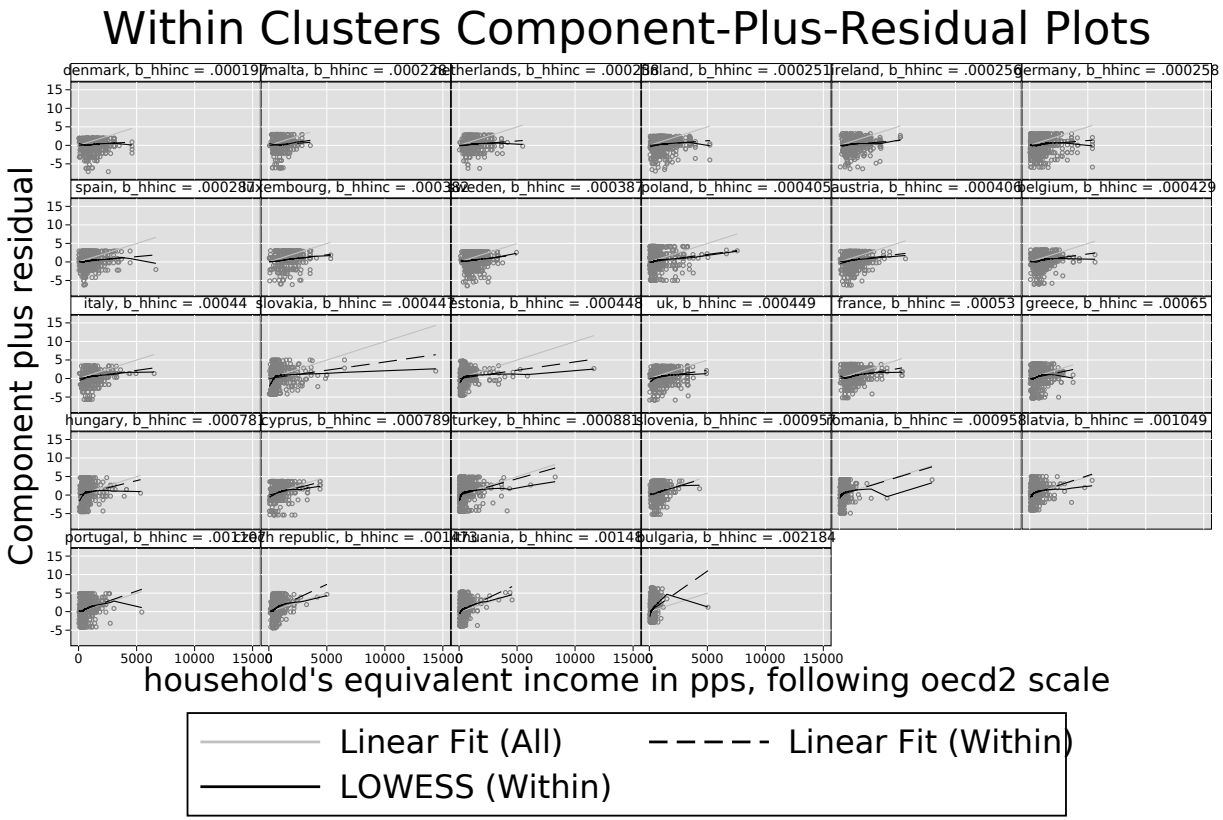
```
. twostep cntry: reg lsat hhinc
  || unitregby _b_hhinc hdirank, unitby(sex)
```

Contents

Unit level CPR plot

The unit level command `unitcprplot` shows the CPR plot of a selected unit level covariate for all the unit level models.

```
. twostep cntry: unitcpr lsat hhinc i.sex || _b_hhinc
```



Note: With `unitcprplot` there is no cluster level command!

Variants

- Order of plots can be controlled by the varlist in the cluster level part:

```
. twostep cntry, stats(r2): unitcpr lsat hhinc i.sex || _b_hhinc _stat_r2
```

- Graph option can be used to show cluster specific regression and LOWESS lines, only:

```
. twostep cntry: unitcpr lsat hhinc i.sex
  || _b_hhinc hdirank, scopts(ms(i)) allopts(lwidth(0))
```

Contents

Create cluster level data

Cluster level command `mk2nd` creates cluster level data. This allows arbitrary follow up analisis on the cluster level data. EDV can be invoked by using `twostep` as standalone command:

```
. twostep cntry: reg lsat hhinc i.sex
  || mk2nd _all hdirank corrupt, clear
. twostep _b_hhinc hdirank corrupt
```

Cluster level data from external file

If the variables for the cluster level are stored in a second file, they will be accessible with `using`:

```
. use ../eqls_4x, clear
. twostep cntry: reg lsat hhinc i.sex
  || edv _b_hhinc hdirank corrupt using aggregates
(sum of wgt is 181,228,599.5834)
```

Source	SS	df	MS	Number of obs	=	26
				F(2, 23)	=	8.19
Model	1.9921e-06	2	9.9606e-07	Prob > F	=	0.0021
Residual	2.7980e-06	23	1.2165e-07	R-squared	=	0.4159
				Adj R-squared	=	0.3651
Total	4.7901e-06	25	1.9161e-07	Root MSE	=	.00035

_b_hhinc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hdirank	4.30e-06	5.23e-06	0.82	0.419	-6.51e-06	.0000151
corrupt	-.0000966	.0000527	-1.83	0.080	-.0002056	.0000124
_cons	.0011075	.0004571	2.42	0.024	.0001619	.0020531

```
Sampling Variance Proportion =  .81
```

Standard features

Standard Stata features are supported to a great extent

- Factor variable notation
- `if`, `in`, `weights`
- Stored estimates for `edv`
- `graph` options, `twoway` options

Acknowledgements

- We wish to thank Lena Hipp and Kekeli Abbey for beta testing. Ulrich Kohler wishes to thank the participants of summer's 2017 and winter's 2020/21 multilevel seminar for commenting on earlier versions of `twostep`.
- Graphs in this presentation are design using Ben Jann's `grstyle` package
- `twostep` with `edv` is based on `edvreg` originally written by Jeffrey Lewis, with contributions by Eduardo Leoni.
- Jeffrey Lewis (UCLA) gave valuable hints on the EDV model.

Bibliograpy I

Achen, C. 2005. Two-Step Hierarchical Estimation: Beyond Regression Analysis. *Political Analysis* 13: 447–456. http://pan.oxfordjournals.org/content/13/4/447.full.pdf+html.
URL http://pan.oxfordjournals.org/content/13/4/447.full.pdf+html

Borjas, G. and G. Sueyoshi. 1994. A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64: 165–182.

Bowers, J. and K. Drake. 2005. EDA for HLM: Visualization when Probabilistic Inference Fails. *Political Analysis* 13: 301–326. http://pan.oxfordjournals.org/content/13/4/301.full.pdf+html.
URL http://pan.oxfordjournals.org/content/13/4/301.full.pdf+html

Bibliograpy II

Heisig, J. P., M. Schaeffer, and J. Giesecke. 2017. The Costs of Simplicity: Why Multilevel Models May Benefit from Accounting for Cross-Cluster Differences in the Effects of Controls. *American Sociological Review* 82(4): 796–827.

Lewis, F. B. and D. A. Linzer. 2005. Estimating Regression Model in Which the Dependent Variable is Based on Estimates. *Political Analysis* 13: 345–364.