Modeling Multilevel Data. The Estimated Dependent Variable Approach

Antonio M. Jaime-Castillo

University of Málaga

October 22, 2015

Spanish Stata Users Group Meeting
IE Business School, Madrid

Contents

Introduction

- Multilevel data have become very popular in the Social Sciences. Several international research projects (e.g., ESS, ISSP and WVS) have produced a large amount of comparative data in recent decades
- The dominant approach to analyze multilevel data uses multilevel models (a mixture of fixed and random effects). Major statistical packages has incorporated routines for estimating mixed models
- This analytical strategy has several advantages over most naïve pooling strategies. However, it also has some drawbacks on both theoretical and practical grounds

The EDV approach

- An alternative to multilevel models is the Estimated Dependent Variable (EDV) approach (Hanusek, 1974; Lewis and Linzer, 2005), which involves two steps
- In the first step we estimate a separate model for individuals nested within each level 2 unit. The estimates of interest are kept for furthter analysis
- In the second step, estimates obtained in the first step become the dependent variable to be explained by a set of aggregate predictors

Advantages

- The statistical theory behind multilevel models is still under development
- The EDV approach allows for complex models at level 1 that are difficult to estimate using multilevel techniques (e.g., matching samples, imputed values)
- The computational burden to estimate non-linear multilevel models, as well as convergence issues, can be challenging in some cases
- The computational burden involved by the EDV approach is much lower

The model

Following Lewis and Linzer (2005), we start with the following model:

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ik} + \epsilon_i \tag{1}$$

However, y_i is not observable. We observe and unbiased estimate y_i^*:

$$y_i^* = y_i + u_i \tag{2}$$

where $E(u_i) = 0$ and $Var(u_i) = \omega_i^2$. By plugging (1) into (2), we get:

$$y_i^* = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ik} + u_i + \epsilon_i \tag{3}$$

Disturbances

It is clear that if $\omega_i \neq \omega_j$ for some i and j, then v_i $(u_i + \epsilon_i)$ is heteroskedastic:

$$E(\mathbf{v}\mathbf{v}') = \mathbf{\Omega} = \begin{bmatrix} \sigma^2 + \omega_1^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 + \omega_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 + \omega_N^2 \end{bmatrix}$$

where $Var(\epsilon) = \sigma^2$. If σ^2 and ω_i^2 were known, we can use WLS to estimate Equation (3). Weights are given by:

$$w_i = \frac{1}{\sqrt{\sigma^2 + \omega_i^2}}$$

Estimation by OLS and WLS

- Equation (3) can be estimated by OLS. However, ω_i must be constant for all observations, which is not usually true. In general, OLS estimators will be inconsistent
- Inconsistent OLS standard errors can be corrected using robust standard errors (Efron, 1982; White, 1980). However, OLS estimators will be inefficient, as they only have partial information about the source of heteroscedasticity
- The WLS approach sets $w_i = 1/\omega_i$, which implies $\sigma^2 = 0$. This amounts to assume that the total residual (v_i) is only due to the sampling error (u_i). In that case, the R^2 for the main regression would be 1! if we could observe y_i instead of y_i^*

Estimation by FGLS

The model originally proposed by Hanusek (1974) exploits the fact that ω_i is usually assumed to be known. Therefore, only an estimate of σ_i^2 is needed to obtain weights for the second stage WLS regression. The expectation of the sum of squared residuals is given by:

$$\begin{aligned} E\left(\sum_i \hat{v}_i^2\right) &= E\left(\mathbf{v}'\mathbf{v}\right) - tr\left(\mathbf{X}'\mathbf{X}^{-\mathbf{1}}\mathbf{X}'\mathbf{\Omega}\mathbf{X}\right) \\ &= N\sigma^2 + \sum_i \omega_i^2 \end{aligned}$$

where $\mathbf{\Omega}$ is the variance-covariance matrix of regression residuals and $\mathbf{\Omega} = \sigma^2\mathbf{I} + \mathbf{G}$, where $\mathbf{G}$ is a diagonal matrix with ω_i^2 as the ith diagonal element

Estimation by FGLS

After some algebra we get:

$$\sigma^2 = \frac{E\left(\sum_i \hat{v}^2\right) - \sum_i \omega^2 + tr\left(\mathbf{X'X^{-1}X'GX}\right)}{N-k}$$

which implies that:

$$\hat{\sigma}^2 = \frac{\sum_i \hat{v}^2 - \sum_i \omega^2 + tr\left(\mathbf{X'X^{-1}X'GX}\right)}{N-k}$$

Now we can use this estimator of σ^2 to compute the weights used to estimate the main regression:

$$w_i = \frac{1}{\sqrt{\omega_i^2 + \hat{\sigma}^2}}$$

Inequality and electoral turnout

- Empirical research has shown that electoral turnout is positively correlated with income at the individual level
- The aggregate relationship between income inequality and electoral turnout is still unclear, as the effect of income on the probability of voting varies substantially across countries
- The relative power theory (Goodin and Dryzek, 1980) predicts that inequality will depress turnout, although there are conflicting empirical results
- Conflict theory (Meltzer and Richard, 1981; Brady, 2004) suggests that the effect of income will increase as party polarization increases
- Mobilization theories (Kumlin and Svallfors, 2007) suggest that the effect of income will decline in well established democracies

Income and voter turnout

Table 1: Voter turnout by income quintile (selected countries)

	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$	$Q5 - Q1$
Denmark (2007)	97.2	96.3	97.3	99.2	98.2	1.0
Austria (2007)	97.4	98.6	98.8	99.3	99.7	2.3
Mexico (2009)	75.8	74.3	79.2	78.2	78.7	2.9
France (2007)	81.6	80.5	80.6	86.3	85.0	3.3
Canada (2008)	85.9	88.0	89.8	89.6	91.2	5.4
Turkey (2011)	90.0	96.3	93.2	94.9	95.5	5.5
Netherlands (2006)	90.2	91.9	93.2	94.5	95.8	5.6
Spain (2008)	77.8	85.9	81.5	87.2	89.5	11.8
Estonia (2011)	71.2	79.9	84.2	84.0	85.1	14.0
Norway (2009)	81.6	89.7	91.9	92.5	96.8	15.2
Portugal (2009)	69.2	70.1	76.0	79.3	84.6	15.4
Finland (2007)	75.0	83.1	83.0	82.9	93.3	18.3
Switzerland (2010)	59.9	74.2	69.8	75.1	81.4	21.5
Poland (2005)	41.3	49.7	54.8	57.8	63.3	22.0

Source: Comparative Study of the Electoral Systems (2013)

First step

- Individual logistic regression for each country
- Dependent variable: Cast a vote in the last national election
- Explanatory variables: income and controls for gender, age, marital status, education level and work status (employed, unemployed and not in the labor force)
- Data: Comparative Study of the Electoral Systems (2013), Module 3
- Sample: 80,000 individuals within 41 countries

Second step

- Dependent variable: Marginal effect of income
- Explanatory variables:
  - Market inequality: Gini index
  - Party Polarization: average distance between parties in policy positions (weighted by vote share) (Jansen et al. 2013)
  - Democracy stock: average level of democracy (1945-)
- Data: Solt (2013), Manifesto Project (Volkens et al., 2014) and Polity IV (2014)
- Estimation techniques: OLS, WLS and FGLS

Estimated marginal effects

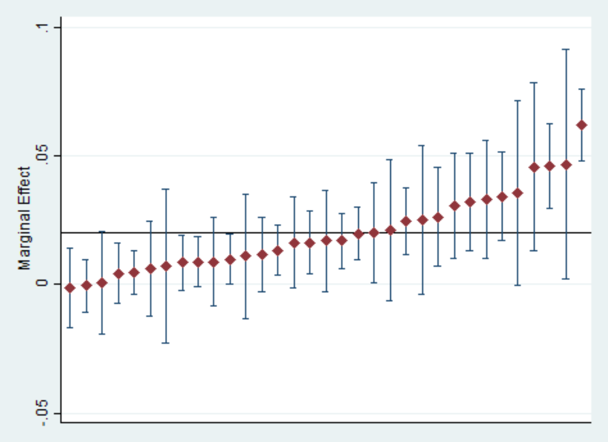

Figure 1: Marginal effects of income

FGLS estimation

```
*generating residuals
reg mfxinc mktgini polariz demst45
predict resid, residuals
gen residsq = resid^2
quietly sum residsq
local sumresidsq = r(sum)

*getting omega
gen omegasq = se_mfxinc^2
mkmat omegasq, matrix(omegasq)
matrix G = diag(omegasq)
quietly sum omegasq
local sumomegasq = r(sum)

*generating matrices
gen ones = 1
mkmat mktgini polariz demst45 ones, matrix(X)
local N = rowsof(X)
local k = colsof(X)
matrix S = inv(X'*X)*X'*G*X
```

FGLS estimation

```
*computing sigma and weights
local tr_S = trace(S)
local sigmahatsq = (`sumresidsq' - `sumomegasq' + `tr_S')/(`N' - `k')
gen weight = 1/(sqrt(omegasq + `sigmahatsq'))

*second step regression
reg mfxinc mktgini polariz demst45 [pweight = weight]
display "sigmahat " sqrt(`sigmahatsq')
quietly sum se_mfxinc
display "omega(average) " r(mean)
```

Second step estimates

Table 2: Cross-national variation in the marginal effect of income

	OLS[1]	WLS	FGLS
Market inequality	0.004	0.042	0.014
	(0.050)	(0.031)	(0.055)
Party Polarization	0.018**	0.017***	0.018**
	(0.008)	(0.003)	(0.008)
Democracy Stock	-0.002***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)
Intercept	0.015	-0.006	0.008
	(0.022)	(0.014)	(0.024)
R^2	0.301		
$\hat{\sigma}$	0.002		0.008
Average ω			0.009
N	33	33	33

Sources: CSES (2013), Polity IV (2014), Solt (2013) and Volkens et al. (2014)
Notes: $^{***}p < 0.01$, $^{**}p < 0.05$ and $^{*}p < 0.10$
[1] Robust standard errors (Efron, 1982)

Conclusions

Main findings

- The effect of income on the probability of voting increases with party polarization
- Differences in electoral participation by income decrease in older democracies

Methodological issues

- The EDV approach allows to estimate the impact of aggregate covariates on estimates obtained at lower levels of analysis
- The EDV approach is computationally very efficient as compared to standard multilevel techniques

Thank you. Comments are welcome!!