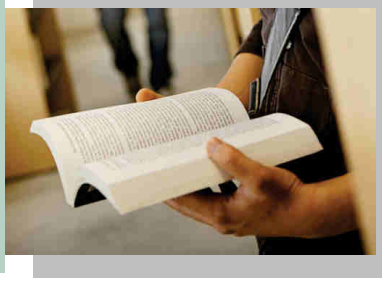
Università della Svizzera italiana

Facoltà di scienze economiche

CORe



Multi-level models for research policy and higher education studies

Benedetto Lepori

USI

Topics

- Why and when are multi-level models useful for studies of research policy and higher education?
- What are the drawbacks and potential limitations of these models?
- Are there alternative strategies? How they compare to multi-level models?
- How to explain multilevel models to referees (particularly in economics journals)?

Goal of the presentation

- Foster critical thinking about choices concerning empirical design and empirical strategies
- Provide a few criteria to assess whether multi-level models are useful in terms of the added value they provide to research questions
- Suggest robust strategies comparing different types of models (including single-level and multi-level) exploiting their complementarities

Statistical inference is a matter of *compromises*, rarely there is a best model which fulfils all the criteria.

Multi-level settings

Multi-level settings are very widespread in research policy.

A few examples:

- Research organizations are nested within countries.
- Research units are nested within larger organizations like universities.
- Individuals belong to research units.

There is evidence that higher-level factors influence activities and performance at the lower level.

That there are nested levels is a necessary condition for multi-level, but it is not sufficient:

- In many instances simpler single-level models might be preferable

Criteria

The research question

- multi-level effects are considered phenomena of substantive (policy or theoretical interest)
- If they are just disturbances there are usually more efficient methods

The theoretical model

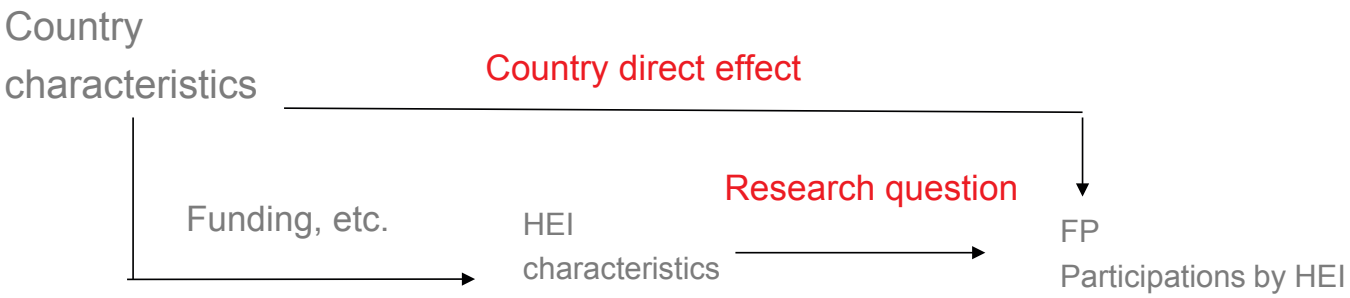
- We can identify a mechanism for multi-level effects
- Which allows modeling them and developing predictions

The empirical data

- Sufficient number of observations (particularly 2° level units)
- There is sufficient variance both at level-1 and level-2 to provide reliable estimates

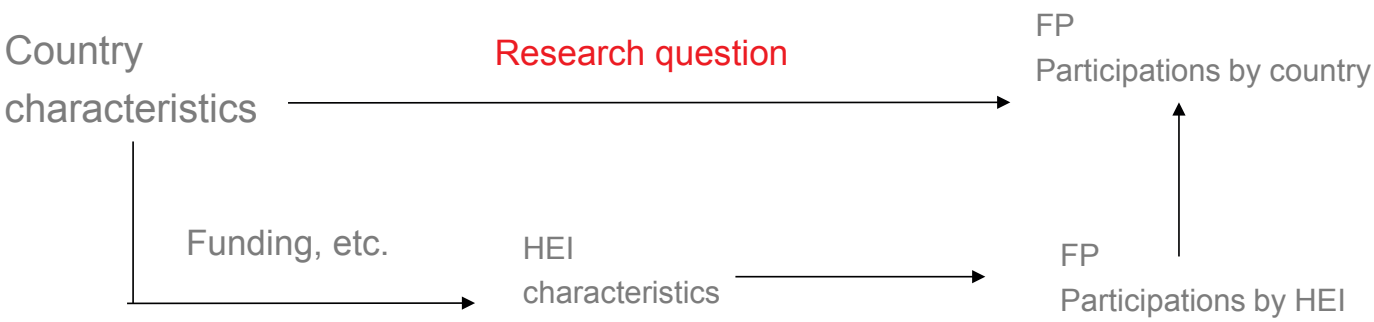
Example: participations to EU FP

- Analyze the factors associated with the number of participation of HEIs to EU-FP
  - For example HEI size, HEI reputation, etc.
  - No need of a multi-level model, but to take into account country differences (which are not modelled through HEI variables): introduce country dummies or cluster SE by country
  - We might want to investigate country effects on participations (for example NMS being favoured for cohesion reasons) > the substantive question includes a multi-level dimension



Same setting, different question

- Analyze the factors associated with the number of participation of countries to HEI-FP
  - Countries do not participate directly, hence a regression with country-level aggregates would be methodologically incorrect
  - To analyze the impact of country characteristics (funding etc.), we need a two-level model



Internationalization of HEIs staff

Migration theories show that the decision to migrate or not is associated with the characteristics of the host country:

- National wealth, employement opportunities
- Which provide estimates of the potential benefit of moving (not yet necessarily realized)

Studies of academics mobility show that academics decide based on the characteristics of the HEI where they are hired

- Reputation, research orientation, fit with own research interests.

There are strong theoretical reasons to assume that both country and HEI characteristics influence internationalization of HEIs

- And the two effects interact
- The relative importance of the two effects is of substantive theoretical and practical relevance

It is «natural» to use a multi-level model and theory suggests specific country and HEI-level variables to be tested.

Policy relevance

Scenario A: internationalization of HEIs depends mostly on their own characteristics, like reputation.

- HEIs hiring strategies might depend on their characteristics, but not on the country.
- Opening policies favour the best HEIs in all Europe
- To internationalize, less good countries need to promote excellence, for example being very selective and having few very good HEIs

Scenario B: internationalization of HEIs depends mostly on the characteristics of their own country:

- Very good HEIs in less attractive countries will not be able to hire foreigners > focus on the best nationals
- Opening policies favour the most attractive countries and penalize good HEIs in less attractive countries
- Less attractive countries need to focus on domestic HR formation and training.

Strategic and policy implications are dramatically different depending on which scenario holds.

Data considerations

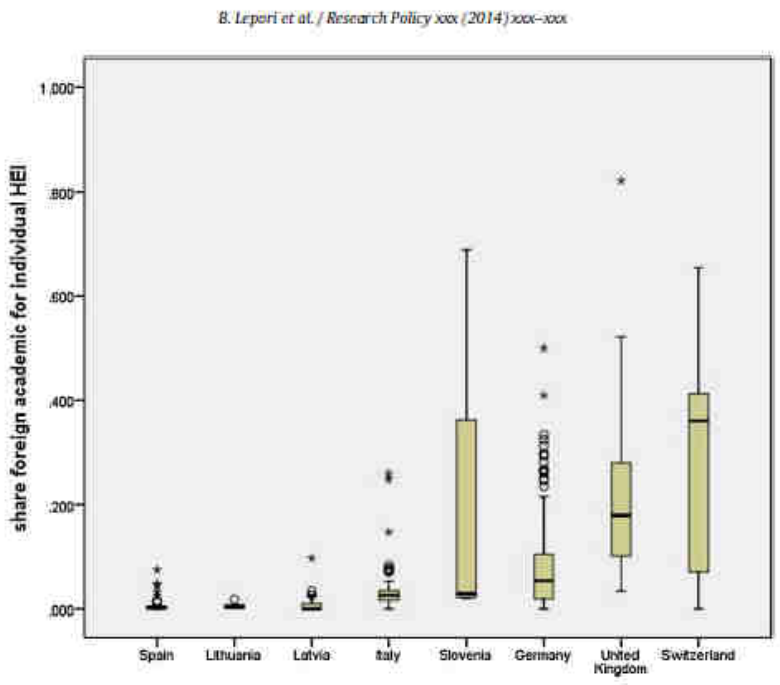
In terms of data, the use of multi-level models is meaningful when:

- There is a sufficient number of units at *both* levels to provide meaningful estimates
  - Applies particularly for level-2 units: with 3 countries, it is highly problematic to estimate a model, better just to introduce country dummies
  - Requirements in terms of number depends on the model: more independent variables and random-slope models require more observations.
- There is enough variance at *both* levels
  - Models can only estimate variations between observations
  - Otherwise models might become unstable and effects not sigificant or robust

Descriptive analysis accross levels (boxplots) and analysis of variance (ANOVA) are basic starting steps for multi-level analysis:

- Also to identify critical issues for model design.

Internationalization of HEIs



- 8 countries and 601 HEIs
- Clear descriptive evidence of both country and HEI effects
- Country effects will be problematic to model

Modeling strategies

How to select a reasonable model specification for our problem?

Usually there is no sigle best choice, different specifications might have different pros and cons

Robustness is a key issue: do different specifications provide the same substantive results?

Finally: we need to take into account that research communities (and their journals) might have preferences and know only some models.

Università della Svizzera italiana

Facoltà di scienze economiche

Internationalization paper: referees comments

What happens if the authors ran a simple tobit regression with standard errors clustered by country? (1st round)

However, now that I understand the modeling (something that was not possible with the first version), I do not believe the author(s) have selected an appropriate econometric model. (2nd round)

Now my very serious question is: How is it that this dependent variable is appropriate for binary logistic regression? The entire point of binary logistic regression is that there is an underlying latent probability distribution that is unobserved. The values of the dependent variable can only be 0 or 1. In short, unless I have made many, many mistakes of my own in logistic analysis, this is not the appropriate method for the dependent variable described in this analysis. (2nd round)

Eureka! This is a nicely done revision, and very good paper. The authors may want to consider preparing their methodological response to reviewers for a general audience--I certainly learned a lot about the various models, and it seems that an approach like this would be very useful for people like me who are stats users but not necessarily stats experts. (3rd round)

Between versions, the model design has not really changed, but:

- It has been explained much more clearly.
- The paper is more overt in why this approach has been chosen.
- Alternative specifications have been tested and provide the same results.

Modeling strategies

The simplest possible model is a single-level OLS with both country and HEI variables

$$share\_foreign_{ij} = \alpha + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

Pros: very simple and robust model, efficient estimator.

Problems:

- Dependent is bounded between 0 and 1 and is heteroskedastic (variance goes to 0 at the extremes)
- Cases within the same country are treated as independent
- HEIs have very different numbers of staff > it makes a big difference if we compute a proportion on 10 individuals or on 1000
- Theory suggest a multi-level structure which is not embedded in the model.

Good for a first test, but absolutely not robust. You should always start with this type of model.

Introducing variables

This is a critical issue in the specification of statistical models

- Too many variables or correlated variables render models less stable and make interpretation more complex

Two critical issues

- We have only 8 countries, so we can afford only one country variable
- Some HEI variables are strongly correlated, like international reputation and researh intensity (n. of PhDs)

Constructing variables through factor analysis proves to be a good solution:

- Country attractiveness as a composition of 4 country variables which are all correlated
- Research intensity and teaching orientation as composition of 4 HEI variables

Factor analysis

	Research orientation	Teaching orientation
research_intensity	.662	-.461
teaching_load	-.112	.948
Reputation	.760	-.286
Type_university	.862	.049

From 4 variables which are highly correlated we extract two ortogonal variables referring to two meaningful dimensions of HEI activity

- A more stable solution
- Easy to interpret in substantive terms

Remedies

Econometrics provides remedies for these problems without fundamentally altering the modeling strategy

Bounded variable: Tobit regression (censored at 0 and 1)

Within-countries correlations: clustered standard errors by country.

Number of staff: weights for the estimation.

Proportion modeling: use fractional logit instead of OLS which is a much better approximation (Papke and Wooldridge):

$$logit(share\_foreign_{ij}) = \alpha + \beta_j + \gamma_{ij}$$

The model remains not exactly specified, but it is robust and rather simple to estimate:

- Note that the fractional logistics has no separate parameter for variance and therefore it is less flexible

A simple multilevel specification

(1) $share\_foreign_{ij} = \alpha + \beta_j + \gamma_{ij} + + u_j + \epsilon_{ij}$

By introducing a country-level random intercept we deal with the multi-level structure and different numbers of HEIs by country

- All other problems remain
- We could also use a two-level fractional logistics which is a better model

(2) $logit(share\_foreign_{ij}) = \alpha + \beta_j + \gamma_{ij} + u_j$

Equations (1) and (2) are better specified, but estimating country-level variance u_j with only 8 countries proves to be problematic.

An exact specification

We model our dependent variable as the average of a binary variable which specifies whether an individual position is occupied by a foreigner.

- So we have individuals nested within HEIs nested within countries – i.e. a three-level model
- With a logistic regression on a true binary variable, therefore with the correct model for the variance

The model for the probability that position k within HEI i within country j is occupied by a foreigner becomes:

$$logit\,\{\Pr(y_{ijk} = 1 \mid x_{jk}\, x_k\} \;=\; \beta_0 + \beta_1 x_{jk} + \beta_2 x_k + u_k + u_{jk}$$

Since there are no individual-level covariates this gives immediately the proportion of foreigners at the HEI level.

Discussion

Pros

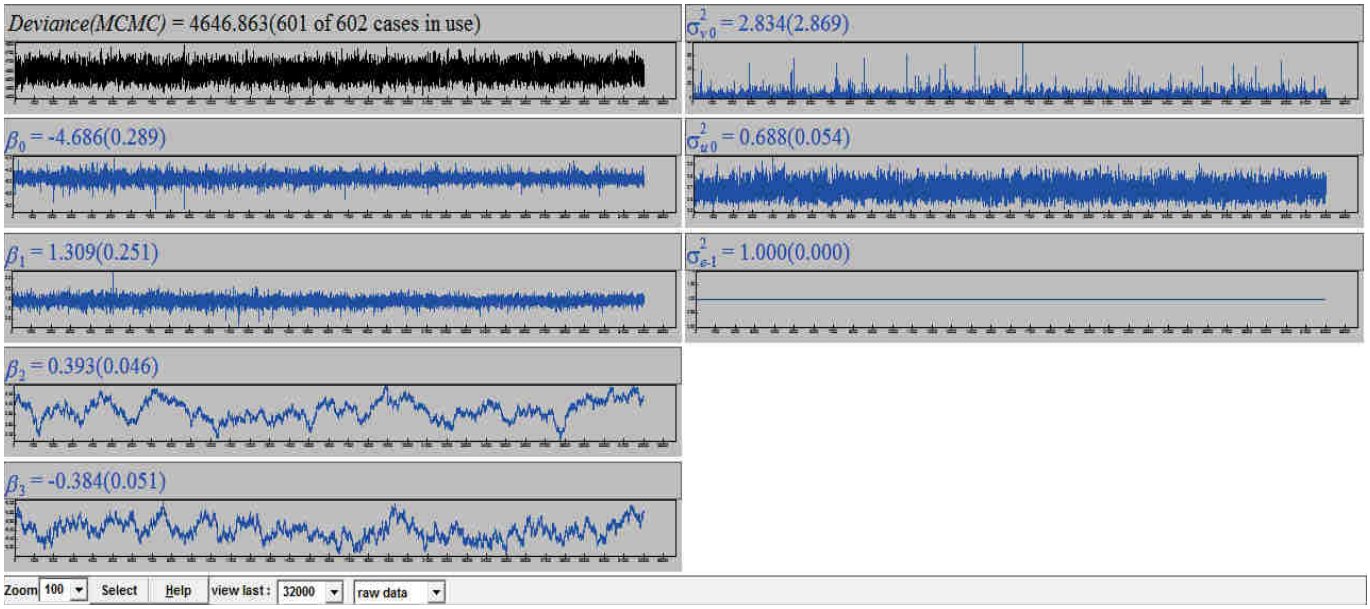
- The model is correctly specified at the micro-level since our proportion is indeed the aggregate of a large number of binary choices
- The number of positions is correctly considered, as well as group-level correlations

Cons

- The model has obvious problems in estimating country-level coefficients and variance
- Standard QMLE techniques are known to be biased for this kind of model
- MCMC methods converge only with difficulties and very slowly

The model is correctly specified, but complex to fit and estimates of country-level parameters change substantially depending on the estimation method.

MCMC estimates



MCMC estimation runs successive iterations of a model using the previous model parameters as starting points

- Do not produce point estimates of parameters, but distributions of them
- The model converges reasonably well only after 500'000 iterations

Comparing estimation methods

	RIGLS		MCMC		Diff (%)
Fixed Part					
cons	-3.239	0.326	-4.207	0.265	30%
country_attractiveness	0.883	0.341	0.959	0.317	9%
research_orientation	0.328	0.056	0.374	0.050	14%
teaching_orientation	-0.393	0.055	-0.376	0.050	-4%
Border_HEI_1	1.049	0.250	1.075	0.239	2%
urban_centrality	0.411	0.200	0.564	0.175	37%
natural_technical_HEI_1	-0.065	0.115	-0.081	0.102	25%
business_HEI_1	0.998	0.232	1.219	0.212	22%
private	0.037	0.151	-0.139	0.142	-476%
(total_staff_1000-gm)	-0.029	0.021	-0.011	0.019	-62%
Country level variance	0.731	0.390	2.582	2.693	253%
HEI level variance	0.770	0.051	0.593	0.048	-23%
Individual level variance	1	0	1.000	0.000	0%

Università della Svizzera italiana

Facoltà di scienze economiche

Comparing models

	Multilevel model															Fractional logistics			Double-censored Tobi		
	Full			*Excluding DE*			*Excluding UK*			*Excluding CH*			*Excluding outliers*								
		S.E.	Sig.		S.E.	Sig.		S.E.	Sig.		S.E.	Sig.		S.E.	Sig.		S.E.	Sig.		S.E.	Sig.
cons	-4.207	0.265	***	-3.799	0.558	***	-3.462	0.924	***	-3.949	0.428	***	-3.724	0.585		-2.911	0.241	***	0.038	0.017	**
country_attractivene	0.959	0.317	*	1.832	0.524	*	1.494	0.570	*	2.057	0.567	*	1.711	0.466	**	0.862	0.238	***	0.103	0.019	***
research orientation	0.374	0.050	***	0.464	0.083	***	0.430	0.063	***	0.334	0.055	***	0.389	0.043	***	0.325	0.113	**	0.041	0.018	*
teaching orientation	-0.376	0.050	***	-0.201	0.075	**	-0.448	0.069	***	-0.396	0.05	***	-0.352	0.043	***	-0.256	0.166		-0.031	0.006	***
Border HEI	1.075	0.239	**	1.103	0.265	***	1.123	0.273	***	1.543	0.323	***	0.917	0.19	***	0.576	0.316		0.121	0.052	**
urban_centrality	0.564	0.175	*	0.503	0.202	*	1.962	0.494	***	0.540	0.174	**	0.356	0.146	*	0.657	0.230	**	0.099	0.019	***
Technical HEI	-0.081	0.102		-0.123	0.199		-0.116	0.124		-0.127	0.107		0.013	0.088		-0.026	0.118		0.010	0.015	
Business HEI	1.219	0.212	**	1.241	0.316	***	1.210	0.249	***	1.245	0.22	***	0.903	0.226	***	0.958	0.146	***	0.130	0.034	***
Private HEI	-0.139	0.142		0.437	0.248	*	-0.25	0.16		-0.209	0.142		-0.336	0.137	*	0.038	0.338		-0.013	0.026	
Total staff*	-0.011	0.019		-0.008	0.03		-0.035	0.024		-0.016	0.019		-0.002	0.016		-0.004	0.016		-0.003	0.001	***

The different specifications, as well as models dropping cases, provide the same substantial results for the variables of interest

- With some expected variation in the coefficient of the country attractiveness
- This strongly supports the robustness of the results

Presentational issues

- We include a clear explanation of the model in the methodological section of the paper since this type model is not current in economics journals
- We prefer the multi-level specification in the paper since it is better specified and, would we have more data, would be clearly a superior option (and represents an innovative element of the paper)
- In the response letter we clearly explained the different possible modeling strategies
- We systematically tested for robustness

Conclusion

Multi-level models are a useful addition in the toolkit of statistical analysis:

- for developing models and new research questions *and* for estimating models

There use needs proper justification concerning

- The nature of the research question
- The underlying conceptual model
- Data consideration

You are always advised to compare results with those of simpler models

- Especially when submitting to economics journals where multi-level methods are not always current.

Università della Svizzera italiana

Facoltà di scienze economiche

QUESTIONS AND DISCUSSION

USI