RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Università della Svizzera italiana

MODELLING STRATEGIES AND THEIR IMPLICATIONS

Lugano, 19th of October, 2020

Benedetto Lepori

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n° 824091

This presentation

RISIS

- An example of the different results depending on the modelling strategies adopted for multilevel settings
  - Comparing simple regressions, dummy variables and multilevel models
- Generated data on a realistic setting in research policy
  - Understanding researchers' productivity
  - Allows comparing the true solution with model results

Università della Svizzera italiana

Researcher's productivity

RISIS

- Determinants of the productivity of individuals
  - As a function of personal characteristics, i.e. past grant history and past mobility
- Typically individuals are nested within a university
  - You might assume that productivity also depends on some characteristics of the university
  - University reputation and funding
- How can we deal with these dependencies?

Università della Svizzera italiana

Dataset

RISIS

- 425 individuals nested within 20 universities
- Individual-level variables
  - Past funding history (X_{ij} mean 10, stdvev 10)
  - Mobility (Y_{ij} 0 /1, mean 0.23)
  - Error term (ϵ_{ij} mean 0, stdev 2)
- University-level variables
  - Funding levels (Z_j mean 5, stdev 1)
  - Reputation (W_j mean 10, stdev 10)
  - Error term ($\mu_j$mean 0, stdev 1)
- Productivity as predicted by these variables through a linear expression (the TRUE solution), *including the error term*

Università della Svizzera italiana

Steps

RISIS

- Create the dataset
- Generate variables and error terms with a normal distribution generator

  https://www.socscistatistics.com/utilities/normaldistribution/default.aspx
- Attribute variables to cases
  - With some 'sorting' to generate multilevel effects
- Compute productivity as:

$$P_{ij} = 5 + 0.3 * X_{ij} + 0.5 * Y_{ij} + 0.3 * Z_j + 0.3 * W_j + \mu_j + \epsilon_{ij}$$

- All this can be done in excel
  - Importing in Stata for the analysis

Analyzing the dataset

RISIS

- It is highly unbalanced
  - N per university ranges between 5 and 50
  - More than half of the sample in just 5 universities
- 95% of the variance at the university level
- Rather large error terms
  - Both at the individual and at the university level
- Strong sorting
    - Mobile individuals concentrated in selected universities

A typical case for the use of multilevel models

Università della Svizzera italiana

Anaylzing sample

RISIS

University	N(Funding~l)	mean(Fund~l)	mean(Past~y)	mean(Mobi~y)	mean(Repu~n)
1	5	5.71	21.942	0	21.71
2	5	6.19	10.454	1	13.79
3	5	3.98	23.794	0	24.28
4	5	5.27	11.512	1	14.47
5	5	3.45	-12.208	1	-8.67
6	10	6.68	9.717	1	10.17
7	10	7.7	33.648	0	31.96
8	10	4.94	22.977	0	21.77
9	10	4.42	10.816	1	14.39
10	10	5.58	10.179	1	12.16
11	20	5.56	9.1105	1	8.92
12	20	6.16	18.1765	.2	18.78
13	20	5.17	16.0595	0	16.6
14	20	4.67	26.3425	0	27.34
15	20	4.84	20.379	0	20.41
16	50	6.03	1.0958	0	1.85
17	50	5.41	7.1832	0	8.54
18	50	5.55	-4.7188	.66000003	-.75
19	50	4.8	4.1574	0	3.35
20	50	6.4	13.2786	0	15.9

Università della Svizzera italiana

Analyzing variance

RISIS

anova Productivity University

```
. anova Productivity University
```

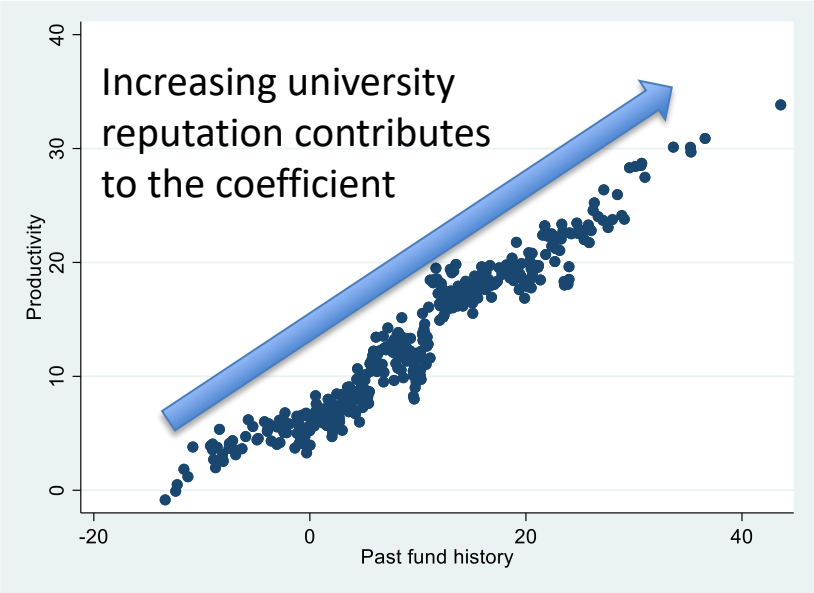
Number of obs =	425	R-squared	=	0.9716	
Root MSE =	1.11777	Adj R-squared	=	0.9703	

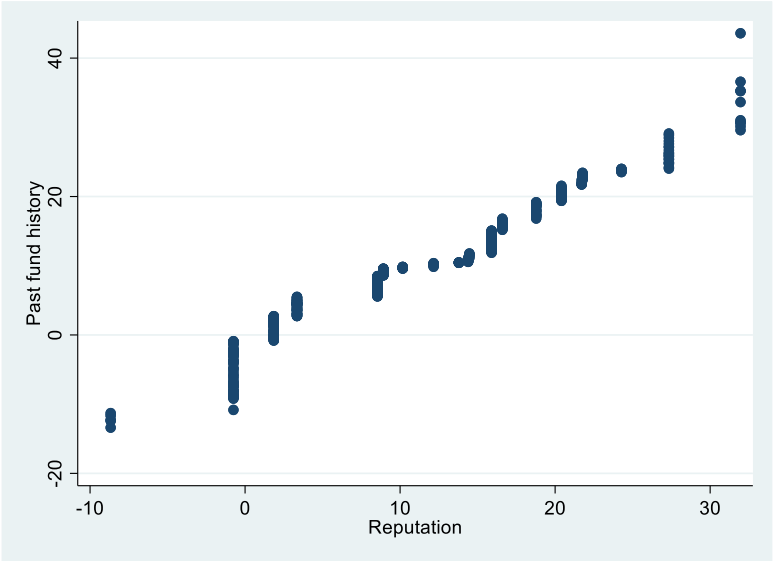
Source	Partial SS	df	MS	F	Prob>F
Model	17332.746	19	912.24981	730.14	0.0000
University	17332.746	19	912.24981	730.14	0.0000
Residual	506.01266	405	1.249414		
Total	17838.759	424	42.072545		

Università della Svizzera italiana

Linear fit

RISIS



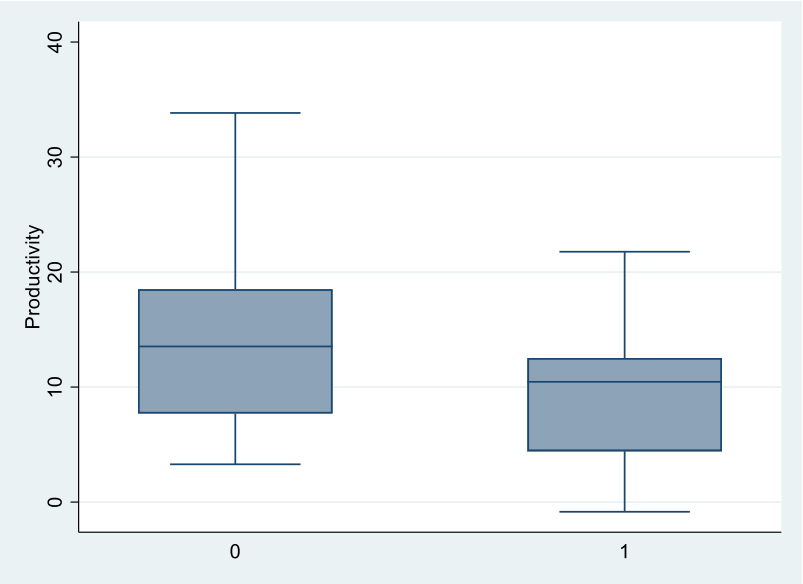


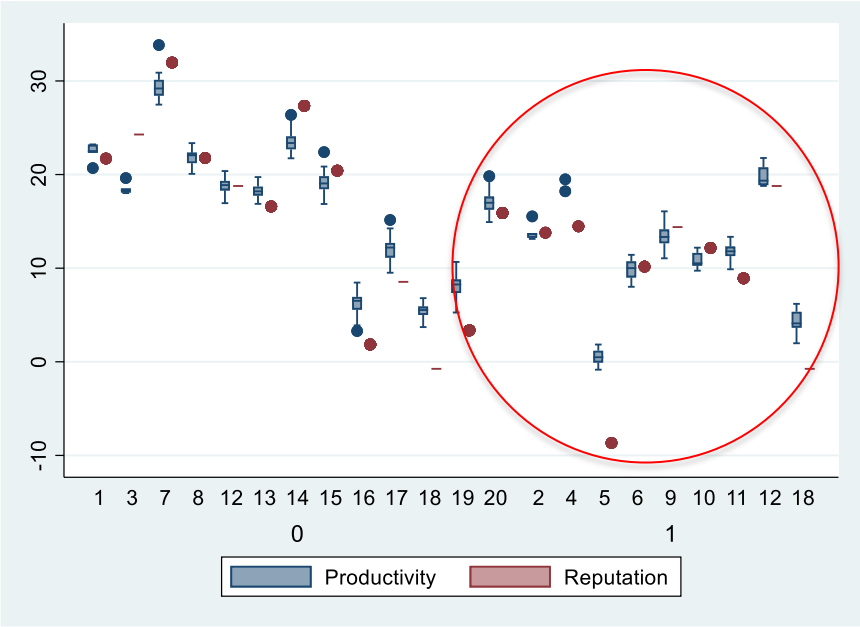
- Coef. 0.65, R2=0.93. Two times the 'real' coefficient!!
- Due to the fact that past funding history is systematically correlated with the reputation of the host university
- You might think there is a *direct* and an *indirect* effect of past grant history, all depends on what you want to look at.

Università della Svizzera italiana

Mobility

RISIS





- Past mobility history has a negative effect on productivity!
- This is generated by the structure of data (*ecological fallacy*)
- Simply the mobile people are concentrated in the less reputed universities generating the effect

Università della Svizzera italiana

Linear regression

RISIS

```
. regress Productivity i.Mobility Pastfundhistory
```

Source	SS	df	MS
Model	16564.8221	2	8282.41107
Residual	1273.93691	422	3.01880785
Total	17838.759	424	42.0725449

Number of obs	= 425
F(2, 422)	= 2743.60
Prob > F	= 0.0000
R-squared	= 0.9286
Adj R-squared	= 0.9282
Root MSE	= 1.7375

True coeffi cients

Productivity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.Mobility	.1232828	.2080895	0.59	0.554	-.2857382	.5323038
Pastfundhistory	.6522931	.009261	70.43	0.000	.6340895	.6704966
_cons	6.698176	.1413005	47.40	0.000	6.420435	6.975917

0.5
0.3

Despite high R2 results are clearly way out from the true coefficient. Disregarding contextual information leads to problematic results when individual observations are not sorted randomly into groups.

Università della Svizzera italiana

Fixed effects model

RISIS

- We introduce a dummy for each university
  - To take out the university effects

$$P_{ij} = \alpha + \beta * X_{ij} + \sum Z_m * D_m + \epsilon_{ij}$$

Where D_m= 1 if j=m, 0 otherwise

  - Puts all individuals on the same 'footing'

- Similar to FE in panel regressions
  - Useful if we focus only on the effect of individual characteristics irrespectively of where individuals are located
  - Or if we don't have information on university characteristics

Università della Svizzera italiana

Fixed effects

RISIS

```
. xtreg Productivity Mobility Pastfundhistory, fe

Fixed-effects (within) regression               Number of obs      =       425
Group variable: University                      Number of groups   =        20

R-sq:                                           Obs per group:
     within  = 0.2224                                         min =         5
     between = 0.9268                                         avg =      21.3
     overall = 0.9249                                         max =        50

                                                F(2,403)           =     57.64
corr(u_i, Xb)  = 0.8595                         Prob > F           =    0.0000
```

Productivity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Mobility	.6549561	.2883325	2.27	0.024	.0881324	1.22178
Pastfundhistory	.3959866	.037897	10.45	0.000	.3214862	.4704871
_cons	8.984082	.3946009	22.77	0.000	8.208349	9.759816
sigma_u	3.3438491					
sigma_e	.98808599					
rho	.91969545	(fraction of variance due to u_i)				

F test that all u_i=0: F(19, 403) = 47.47 Prob > F = 0.0000

.

True coeffi cients

0.5
0.3

Results are more precise, but we have no idea of university effects and how large they are.

Università della Svizzera italiana

Problems

RISIS

- When universities explain most of the differences between individuals
  - Our results are simply uninformative and might be also not very robust as we eliminate most of the variance
  - Look to the intra-class correlation coefficient!
- When university-level effects are of substantive interest
  - For example for decisions on concentrating resources in few universities
- When there are *interactions* between individual-level and university level effects
  - For example past mobility might be less determinant for productivity in top-quality universities

Università della Svizzera italiana

University-level covariates

RISIS

- We introduce university covariates to model university effects

$$P_{ij} = \alpha + \beta * X_{ij} + \gamma * Z_j + \epsilon_{ij}$$

Where Z_j is a vector of characteristics of the university to which the individual belongs to.

- Allows modelling directly university effects based on known characteristics

Università della Svizzera italiana

Linear regression

RISIS

```
. regress Productivity Fundinglevel Reputation i.Mobility Pastfundhistory
```

Source	SS	df	MS		
Model	16936.2912	4	4234.07281	Number of obs	= 425
Residual	902.467823	420	2.14873291	F(4, 420)	= 1970.50
				Prob > F	= 0.0000
Total	17838.759	424	42.0725449	R-squared	= 0.9494
				Adj R-squared	= 0.9489
				Root MSE	= 1.4659

True coeffi cients

Productivity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]	
Fundinglevel	.2550451	.0992281	2.57	0.011	.0599996	.4500907	0.3
Reputation	.4948967	.0398051	12.43	0.000	.4166547	.5731388	0.3
1.Mobility	-.659835	.1864966	-3.54	0.000	-1.026418	-.293252	0.5
Pastfundhistory	.1907594	.0376637	5.06	0.000	.1167266	.2647923	0.3
_cons	4.558756	.5499008	8.29	0.000	3.477856	5.639657	

Despite high R2 results are incorrect and, for mobility, even the sign of the mobility coefficient is wrong (but significant!).

The model does not account correctly for the fact that observations are nested and errors are correlated.

Università della Svizzera italiana

Multilevel model

RISIS

- We replace the university dummies (uninformative)

$$P_{ij} = \alpha + \beta * X_{ij} + \sum Z_m * D_m + \epsilon_{ij}$$

- With a fixed part + a random university-level intercept

$$P_{ij} = \alpha + \beta * X_{ij} + \gamma * Z_j + \mu_j + \epsilon_{ij}$$

So we decompose the university effect into an observable part and an error.

- The simplest possible multilevel model
  - See later in this course for more complex models

Multilevel-model

RISIS

```
. xtreg Productivity Mobility Pastfundhistory Fundinglevel Reputation , re

Random-effects GLS regression                   Number of obs     =       425
Group variable: University                      Number of groups  =        20

R-sq:                                           Obs per group:
     within  = 0.2223                                         min =         5
     between = 0.9358                                         avg =      21.3
     overall = 0.9428                                         max =        50

                                                Wald chi2(4)      =    363.36
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =    0.0000
```

Productivity	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Mobility	.5653821	.2790793	2.03	0.043	.0183967	1.112368
Pastfundhistory	.3895176	.037611	10.36	0.000	.3158013	.4632339
Fundinglevel	.2155802	.4712858	0.46	0.647	-.7081229	1.139283
Reputation	.2822516	.059943	4.71	0.000	.1647655	.3997377
_cons	4.722307	2.500941	1.89	0.059	-.1794477	9.624062
sigma_u	1.8872415					
sigma_e	.98808599					
rho	.78485784	(fraction of variance due to u_i)				

True coefficients

0.3
0.3
0.5
0.3

Università della Svizzera italiana

What you get more

RISIS

- The model provides reasonable estimates of the individual-level effects
  - Similar to the FE model
  - But at the same time allows also estimating the effects of the university-level variables
- However: more complex models, not necessarily better results
  - Estimates are more complex and may become very time-consuming
  - Linear regression as the simplest estimator

Università della Svizzera italiana

Discussion

RISIS

- Multi-level/nested structures are highly frequent in research policy / higher education studies
  - Individuals within universities
  - Universities within countries
  - Individuals within universities within countries
- Two basic ways to deal with them
  - Dummy variables (fe): when the interest is only at the micro-level and interactions do not matter
  - Multi-level models: when the interest is at both levels and there are lots of interactions
- The best approach depends
  - On your substantive interest
  - On the structure of the data

Q&A

RISIS

Università della Svizzera italiana

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

THANK YOU !

WWW.RISIS2.EU

ZENODO.ORG/COMMUNITIES/RISIS

CONTACT@RISIS2.EU

@RISIS_EU

FACEBOOK.COM/RISIS.EU

RISIS2 EU PROJECT

Université Gustave Eiffel · AIT Austrian Institute of Technology · Politecnico Milano 1863 · Consiglio Nazionale delle Ricerche · Universiteit Leiden · The University Of Sheffield. · Fraunhofer ISI · Università della Svizzera italiana · JOANNEUM RESEARCH

University of Sussex · DZHW Deutsches Zentrum für Hochschul- und Wissenschaftsforschung · University of Strathclyde Glasgow · Sapienza Università di Roma · NIFU Nordisk institutt for studier av innovasjon, forskning og utdanning · CSIC Consejo Superior de Investigaciones Científicas · Samuel Neaman Institute for National Policy Research · Technology Centre ASCR · ATHENA Research & Innovation Information Technologies