

MBA
USP
ESALQ

*Supervised Machine
Learning:
Modelagem Multinível I*

Prof. Dr. Luiz Paulo Fávero

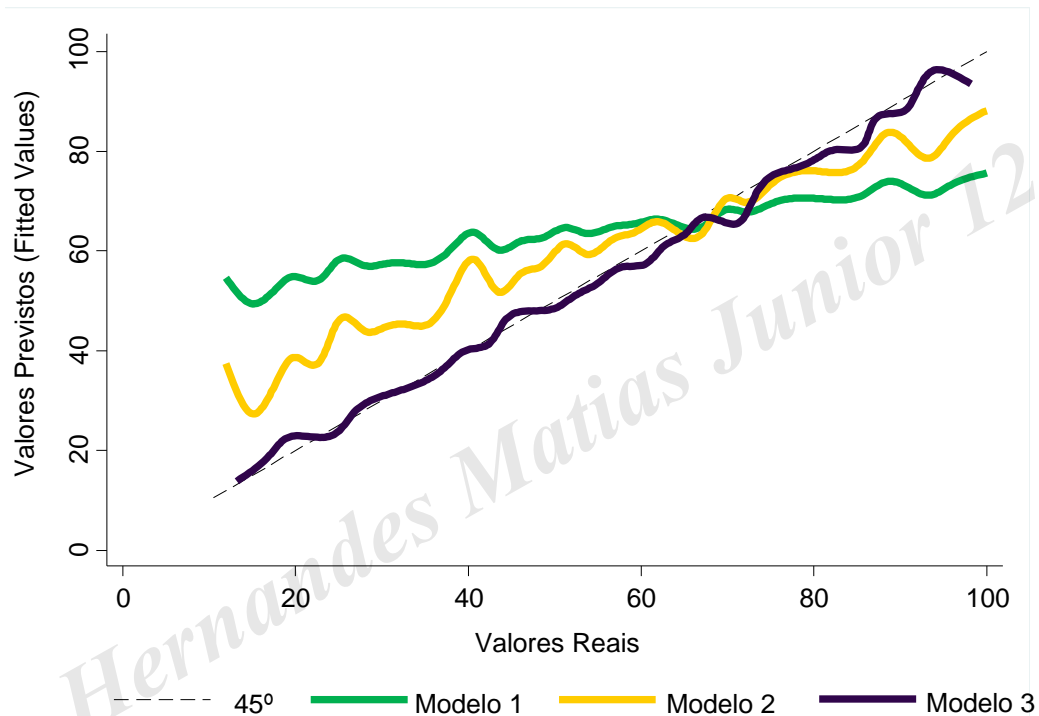
MODELAGEM MULTINÍVEL

Fundamentação teórica dos modelos multinível

Conceitos para a estimação de
modelos multinível

Modelagem multinível no R

Reflexão

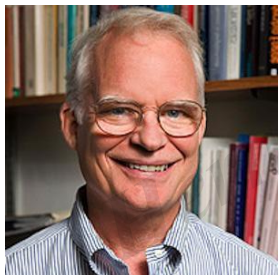


Contexto



O que são Modelos Multinível?

São modelos que reconhecem a existência de estrutura multinível ou hierárquica nos dados.



Stephen W. Raudenbush
University of Chicago

Hierarchical linear models: applications and data analysis methods. 2. ed. Thousand Oaks: Sage Publications, 2002.

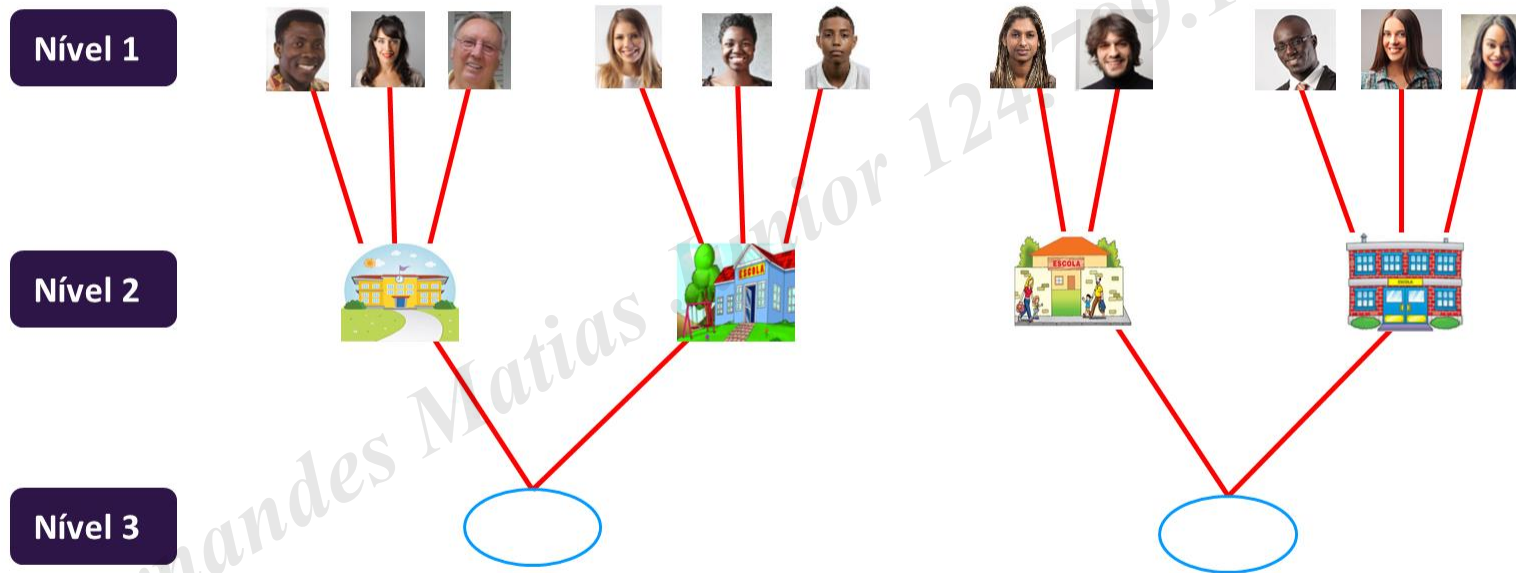


Anthony S. Bryk
Stanford University

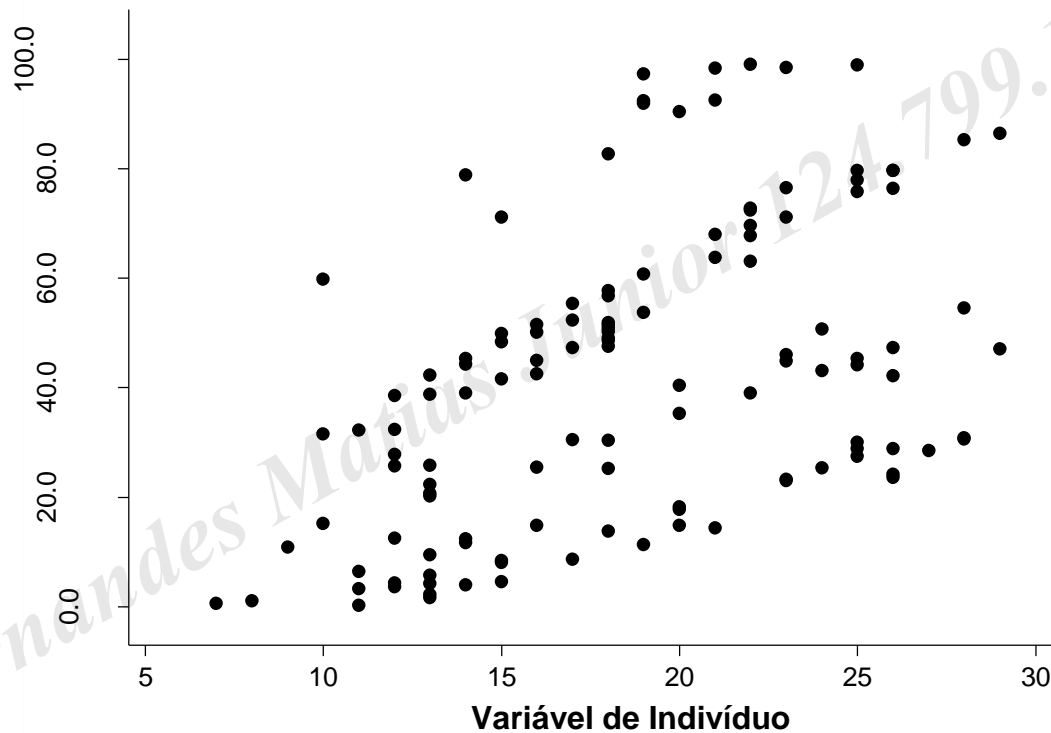
Estrutura Multinível



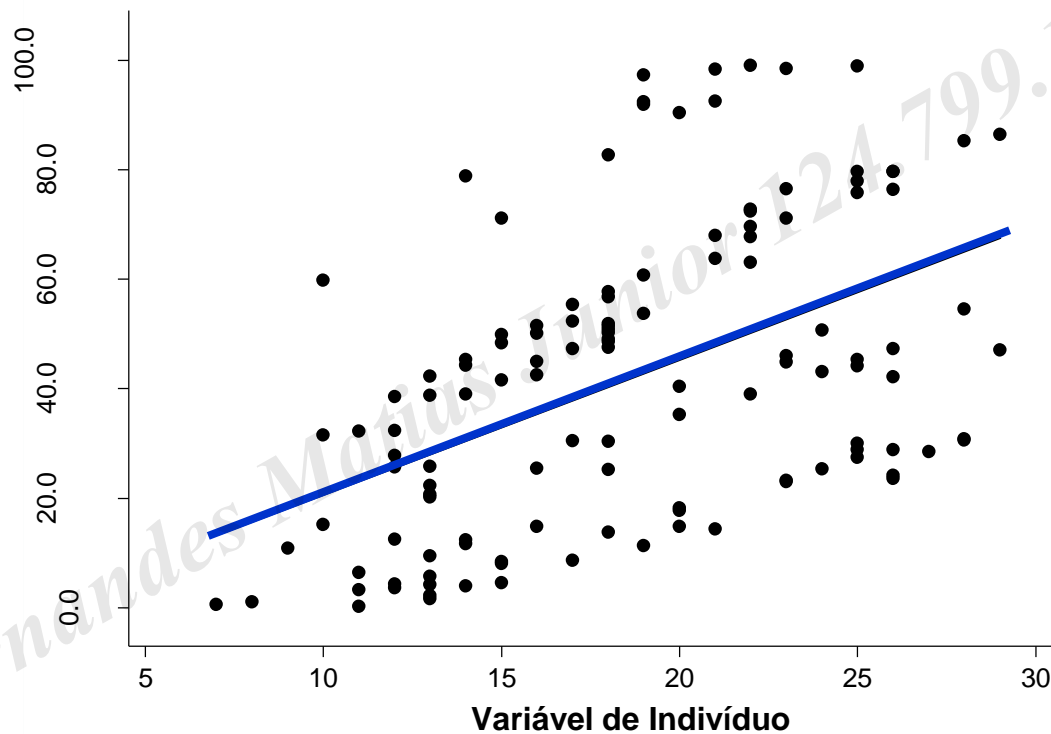
Estrutura Multinível



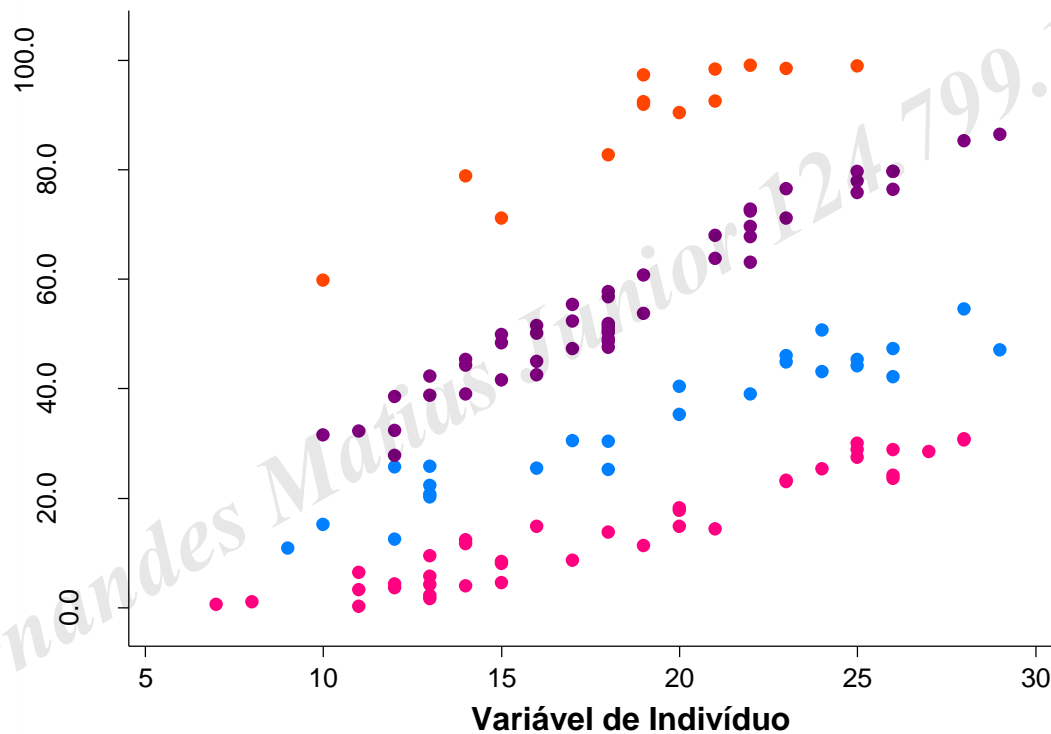
Estrutura Multinível



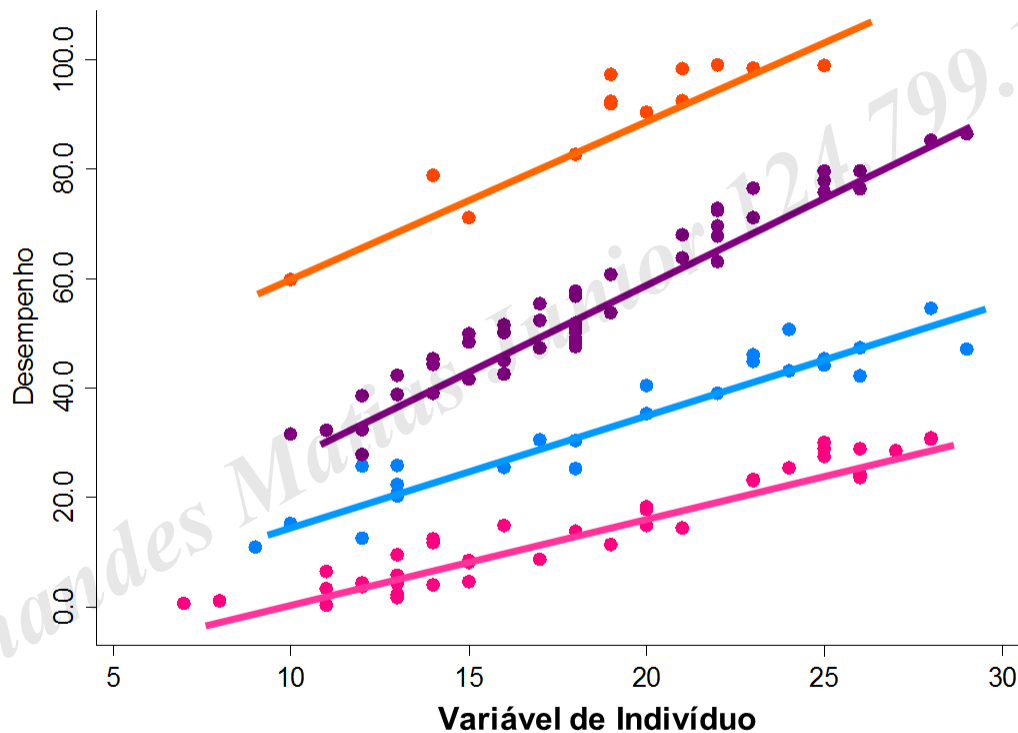
Estrutura Multinível



Estrutura Multinível



Estrutura Multinível



Estrutura Multinível

Escola 1:

$$Y_{i1} = \beta_{01} + \beta_{11} \cdot X_{i1} + \varepsilon_{i1}$$

Escola 2:

$$Y_{i2} = \beta_{02} + \beta_{12} \cdot X_{i2} + \varepsilon_{i2}$$

Escola 3:

$$Y_{i3} = \beta_{03} + \beta_{13} \cdot X_{i3} + \varepsilon_{i3}$$

Escola 4:

$$Y_{i4} = \beta_{04} + \beta_{14} \cdot X_{i4} + \varepsilon_{i4}$$

O Modelo Multinível

Nível 1

$$Y_{ij} = \beta_{0j} + \beta_{1j} \cdot X_{ij} + \varepsilon_{ij}$$

Nível 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot W_j + \nu_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \cdot W_j + \nu_{1j}$$

$$Y_{ij} = \underbrace{\left(\gamma_{00} + \gamma_{01} \cdot W_j + \nu_{0j} \right)}_{\text{intercepto com efeitos aleatórios}} +$$

**intercepto com
efeitos aleatórios**

$$+ \underbrace{\left(\gamma_{10} + \gamma_{11} \cdot W_j + \nu_{1j} \right) \cdot X_{ij}}_{\text{inclinação com efeitos aleatórios}} + \varepsilon_{ij}$$

**inclinação com
efeitos aleatórios**

O Modelo Multinível

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} \cdot X_{ij} + \gamma_{01} \cdot W_j + \gamma_{11} \cdot W_j \cdot X_{ij}}_{\text{Efeitos Fixos}} + \underbrace{v_{0j} + v_{1j} \cdot X_{ij} + \varepsilon_{ij}}_{\text{Efeitos Aleatórios}}$$

- Os modelos tradicionais de regressão ignoram as interações entre variáveis no componente de efeitos fixos e as interações entre termos de erro e variáveis no componente de efeitos aleatórios.

Multilevel statistical models. 4. ed. Chichester: John Wiley & Sons, 2011.

Harvey Goldstein
Centre for Multilevel Modelling
University of Bristol



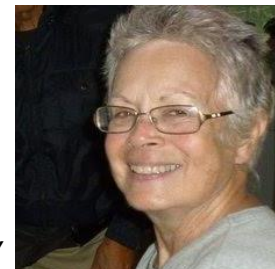
Variância dos Termos Aleatórios

Se as variâncias dos termos aleatórios ν_{0j} e ν_{1j} forem estatisticamente diferentes de zero, procedimentos tradicionais de estimação dos parâmetros do modelo, como mínimos quadrados ordinários, não serão adequados.



Barbara G. Tabachnick
California State University

Using multivariate statistics. 6. ed. Boston: Pearson, 2013.



Linda S. Fidell
California State University

Dummies?

Apenas a inserção de **dummies de grupo** não capturaria os **efeitos contextuais**, visto que não permitiria que se separassem os efeitos observáveis dos não observáveis sobre a variável dependente.



Sophia Rabe-Hesketh
U. C. Berkeley

Multilevel and longitudinal modeling using Stata. 3. ed.
College Station: Stata Press, 2012.

Anders Skrondal
*Norwegian Institute of Public Health
University of Oslo
U. C. Berkeley*



Por que Utilizar?

Os modelos multinível permitem, portanto, o desenvolvimento de novos e mais bem elaborados constructos para predição e tomada de decisão.

“Dentro de uma estrutura de modelo com equação única, parece não haver uma conexão entre indivíduos e a sociedade em que vivem. Neste sentido, o uso de equações em níveis permite que o pesquisador ‘pule’ de uma ciência a outra: alunos e escolas, famílias e bairros, firmas e países. Ignorar esta relação significa elaborar análises incorretas sobre o comportamento dos indivíduos e, igualmente, sobre os comportamentos dos grupos. Somente o reconhecimento destas recíprocas influências permite a análise correta dos fenômenos.”

Methodology and epistemology of multilevel analysis.

London: Kluwer Academic Publishers, 2003.

Daniel Courgeau

*Institut National D'Études
Démographiques*



The background is a teal-colored collage of financial data. It includes various stock indices such as '深圳證券交易所創業板' (Shenzhen Stock Exchange ChiNext), '上證綜指' (Shanghai Composite Index), '創業板指' (ChiNext Index), '上證180指數' (Shanghai 180 Index), '深圳A股指數' (Shenzhen A-share Index), '深證成份指數' (Shenzhen Component Index), '恒生指數' (Hang Seng Index), '滬深300能源指數' (Shanghai-Shenzhen 300 Energy Index), and '滬深300指數' (Shanghai-Shenzhen 300 Index). There are also numerical values like '+2.74', '+0.97', '+3.02', '+0.57', '+1.25', '+1.46', '+0.23%', '+0.01%', '+1.03%', '+1.13%', '+0.54%', '+0.11%', '+0.61%', '+0.27%', '+0.65%', '+0.29%', '+0.52%', '+0.79%', '+3.85%', '+0.48%', '+0.11%', '+0.27%', '+0.61%', '+1.25%', '+1.37%', '+0.61%', and '+0.27%'. A large, semi-transparent watermark '124.799.116-40' is diagonally across the center, and the word 'Junior' is partially visible below it.

APLICAÇÕES DE MODELAGEM MULTINÍVEL

Aplicações

Business, Economics & Management

Periódico	Índice h5 (Google Scholar)	% / Modelos Supervisionados
American Economic Review	158	10,78%
Journal of Business Research	140	12,71%
Tourism Management	118	14,04%
Journal of Business Ethics	117	12,15%
Journal of Financial Economics	116	11,83%
The Quarterly Journal of Economics	110	3,75%
The Review of Financial Studies	108	6,88%
Technological Forecasting and Social Change	106	4,59%
International Journal of Information Management	105	8,15%
Management Science	103	8,57%
		9,26%

Aplicações

Engineering & Computer Science

Periódico	Índice h5 (Google Scholar)	% / Modelos Supervisionados
IEEE/CVF Conference on Computer Vision and Pattern Recognition	356	4,78%
Advanced Materials	294	3,56%
International Conference on Learning Representations	253	6,22%
Neural Information Processing Systems	245	4,83%
Renewable and Sustainable Energy Reviews	225	3,42%
Advanced Energy Materials	206	2,53%
International Conference on Machine Learning	204	8,24%
Energy & Environmental Science	202	3,54%
ACS Nano	202	2,89%
European Conference on Computer Vision	197	3,38%
		4,28%

Aplicações

Health & Medical Sciences

Periódico	Índice h5 (Google Scholar)	% / Modelos Supervisionados
The New England Journal of Medicine	410	1,97%
The Lancet	345	2,34%
Cell	288	2,41%
Journal of the American Medical Association	253	2,75%
Proceedings of the National Academy of Sciences	245	0,98%
Journal of Clinical Oncology	213	1,75%
Nature Medicine	205	0,73%
The Lancet Oncology	196	0,45%
PLoS ONE	185	0,43%
Nature Genetics	184	2,34%
		1,70%

Aplicações

Social Sciences

Periódico	Índice h5 (Google Scholar)	% / Modelos Supervisionados
Journal of Business Ethics	117	0,97%
Computers & Education	109	0,34%
Research Policy	95	0,41%
New Media & Society	93	0,75%
American Journal of Public Health	90	0,98%
Global Environmental Change	86	0,75%
Nature Human Behaviour	84	0,73%
Health Affairs	84	0,45%
Social Science & Medicine	83	0,43%
Teaching and Teacher Education	83	0,97%
		0,64%

Pouca Utilização: Qual a Razão?

- Estrutura dos dados.
- Não consideração de natureza multinível nos dados.
- Capacidade computacional por vezes insuficiente, principalmente quando da existência de interações profundas.



Emmanuel Lazega
Institut d'Études Politiques de Paris

Multilevel network analysis for the social sciences: theory, methods and applications. New York: Springer, 2016.



Tom Snijders
University of Oxford

Aplicação

RAJAN, R.G.; ZINGALES, L.

What do we know about capital structure? Some evidence from international data.
Journal of Finance, v. 50-5, p. 1421-1460, 1995.

- *Compustat Global e MSCI*;

- 4.557 empresas;

- 7 países;

- período: 1987-1991.

Country	Local Market Index	Number of Firms
United States	S&P 500	2.583
Japan	Nikkei 500	514
Germany	FAZ Share Index	191
France	CAC General Index	225
Italy	MIB Current Index	118
United Kingdom	FT 500	608
Canada	TSE 300	318

$$\begin{aligned} \text{Leverage}_i = & \beta_0 + \beta_1 \cdot (\text{Tangible Assets})_i + \beta_2 \cdot (\text{Market to Book})_i \\ & + \beta_3 \cdot (\text{Log Sales})_i + \beta_4 \cdot (\text{ROA})_i + \varepsilon_i \end{aligned}$$

Aplicação

$$\text{Leverage}_i = \beta_0 + \beta_1 \cdot (\text{Tangible Assets})_i + \beta_2 \cdot (\text{Market to Book})_i \\ + \beta_3 \cdot (\text{Log Sales})_i + \beta_4 \cdot (\text{ROA})_i + \varepsilon_i$$

Nível 1

$$\text{Leverage}_{ij} = \beta_{0j} + \beta_{1j} \cdot (\text{Tangible Assets})_{ij} + \beta_{2j} \cdot (\text{Market to Book})_{ij} \\ + \beta_{3j} \cdot (\text{Log Sales})_{ij} + \beta_{4j} \cdot (\text{ROA})_{ij} + \varepsilon_{ij}$$

Nível 2

$$\beta_{0j} = \gamma_{00} + v_{0j}$$

$$\beta_{1j} = \gamma_{10} + v_{1j}$$

$$\beta_{2j} = \gamma_{20} + v_{2j}$$

$$\beta_{3j} = \gamma_{30} + v_{3j}$$

$$\beta_{4j} = \gamma_{40} + v_{4j}$$

leverage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
tang_assets	.3462677	.049087	7.05	0.000	.2500589	.4424765
market_book	-.0641481	.0143289	-4.48	0.000	-.0922322	-.036064
logsale	.0353799	.0098784	3.58	0.000	.0160185	.0547413
roa	-.7729998	.2071899	-3.73	0.000	-1.179085	-.366915
_cons	-.6153343	.795045	-0.77	0.439	-2.173594	.9429252

Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
country:					
var(tang_a~s)		4.33e-13	33%	1.83e-28	1021.785
var(market~k)		.0087904		.002262	.0341598
var(logsale)		7.67e-06		1.96e-06	.00003
var(roa)		10.84124		3.37015	34.87456
var(_cons)		3.811897		1.036019	14.02538
var(Residual)		29.18282	67%	28.01925	30.39471

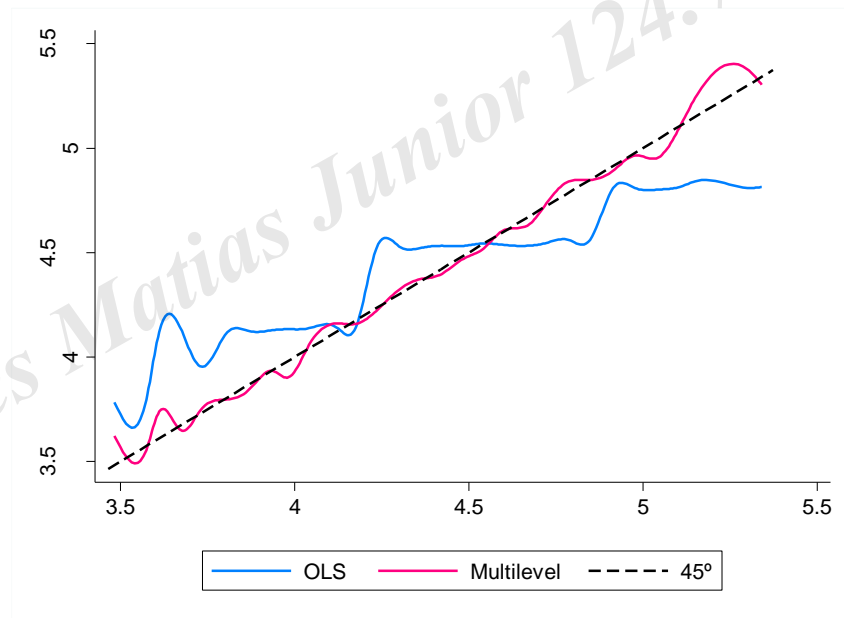
LR test vs. linear regression: chi2(5) = 1047.68 Prob > chi2 = 0.0000

Aplicação

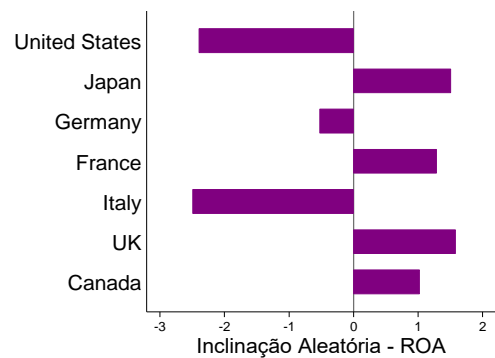
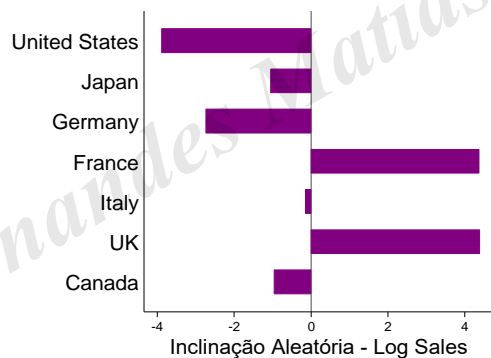
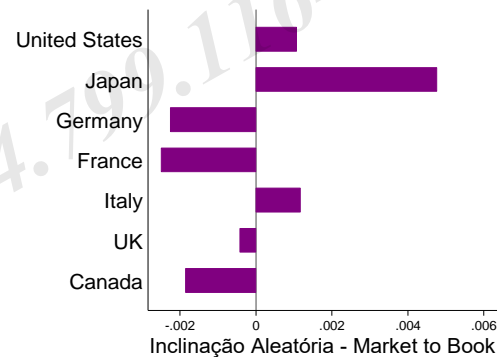
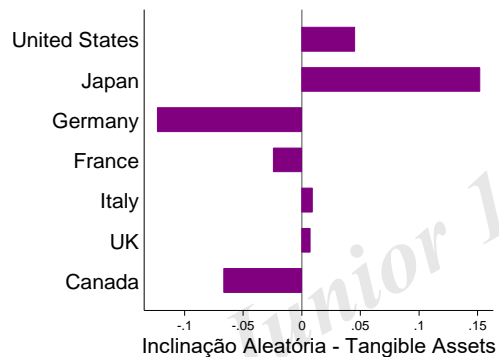
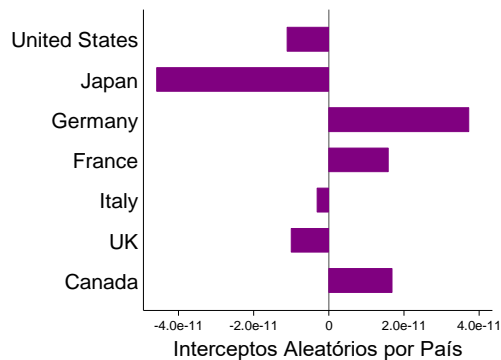
RAJAN, R.G.; ZINGALES, L.

What do we know about capital structure? Some evidence from international data.

Journal of Finance, v. 50-5, p. 1421-1460, 1995.



Aplicação



Aplicação

BELFIORE, P.; FÁVERO, L. P.; SERRA, R. G.; SOUZA, R. F.

Operational efficiency in Brazilian airports: an analysis through the hierarchical linear modeling with repeated measures. **Int. Journal of Logistics Systems and Management**. No prelo.

- ANAC e INFRAERO;

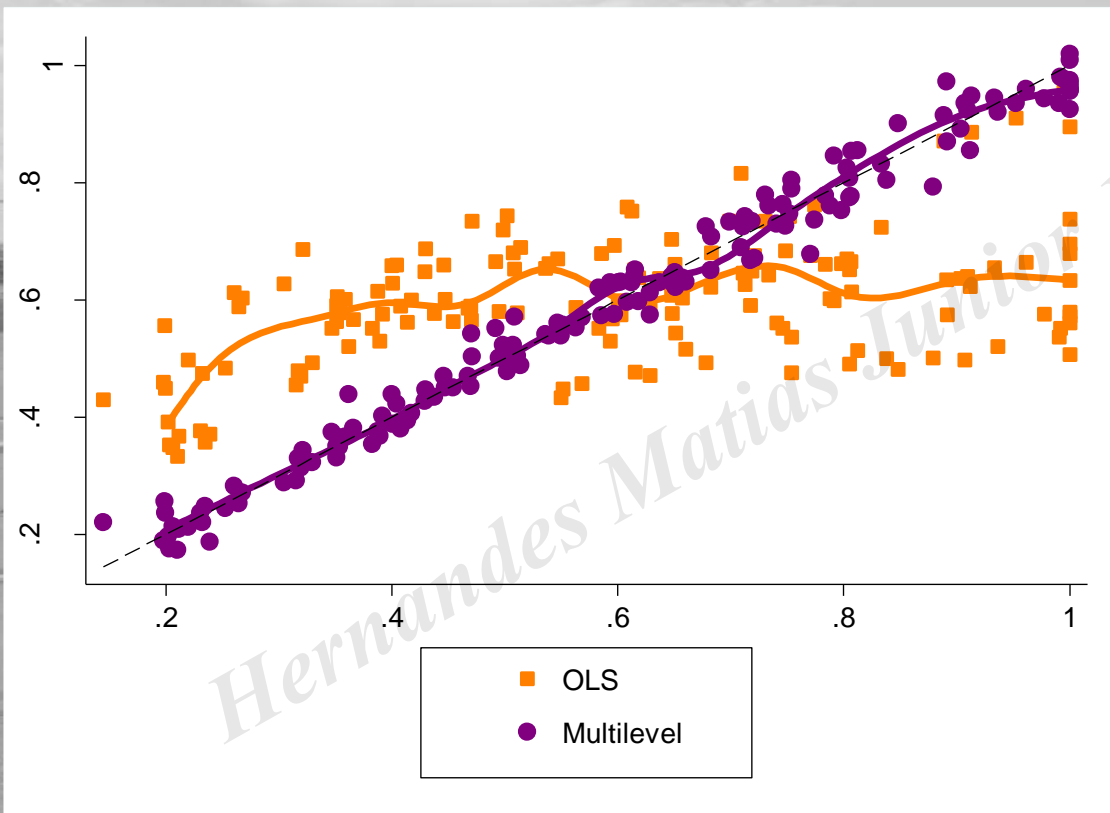
- 30 maiores aeroportos por tráfego;

- 23 UFs;

- período: 2014-2018.

Determinant Factor	Variable	Unit	Label
Governance Structure	Property	Nominal	Public Private Mixed
Airport Operational Characteristics	Size	m ²	Size
	Number of commercial establishments		Commerce
	Number of aircraft parking positions		Positions
	Number of vehicle parking lots		Parking lots
	Airport years of experience		Experience
Service Strategy	Number of airlines		Airlines
Location	Airport location	State	Location

Aplicação



Desafios em Modelagem Multinível

**Interações Profundas e
Capacidade de Processamento**

**Métodos de Estimação dos
Parâmetros**

Clusterização da Amostra

**Estimação de
modelos com a
melhor aderência
possível entre os
valores reais e
previstos**



Andrew Gelman

Multilevel Conference, 31 Out 2015, Columbia University, NYC.

MODELAGEM MULTINÍVEL NO



Modelagem HLM2 com Dados Agrupados



Modelagem HLM2

Modelo Nulo

Nível 1

$$desempenho_{ij} = \beta_{0j} + \varepsilon_{ij}$$

Nível 2

$$\beta_{0j} = \gamma_{00} + v_{0j}$$

Substituindo...

$$desempenho_{ij} = \gamma_{00} + v_{0j} + \varepsilon_{ij}$$



LINHA 275
DO SCRIPT

Modelagem HLM2

Modelo com Interceptos Aleatórios

Nível 1

$$desempenho_{ij} = \beta_{0j} + \beta_{1j} \cdot horas_{ij} + \varepsilon_{ij}$$

Nível 2

$$\begin{cases} \beta_{0j} = \gamma_{00} + v_{0j} \\ \beta_{1j} = \gamma_{10} \end{cases}$$

Substituindo...

$$desempenho_{ij} = \gamma_{00} + \gamma_{10} \cdot horas_{ij} + v_{0j} + \varepsilon_{ij}$$

LINHA 326
DO SCRIPT

Modelagem HLM2

Modelo com Interceptos e Inclinações Aleatórios

Nível 1

$$desempenho_{ij} = \beta_{0j} + \beta_{1j} \cdot horas_{ij} + \varepsilon_{ij}$$

Nível 2

$$\begin{cases} \beta_{0j} = \gamma_{00} + v_{0j} \\ \beta_{1j} = \gamma_{10} + v_{1j} \end{cases}$$

Substituindo...

$$desempenho_{ij} = \gamma_{00} + \gamma_{10} \cdot horas_{ij} + v_{0j} + v_{1j} \cdot horas_{ij} + \varepsilon_{ij}$$

LINHA 365
DO SCRIPT

Modelagem HLM2

Modelo Final HLM2

Nível 1

$$desempenho_{ij} = \beta_{0j} + \beta_{1j} \cdot horas_{ij} + \varepsilon_{ij}$$

Nível 2

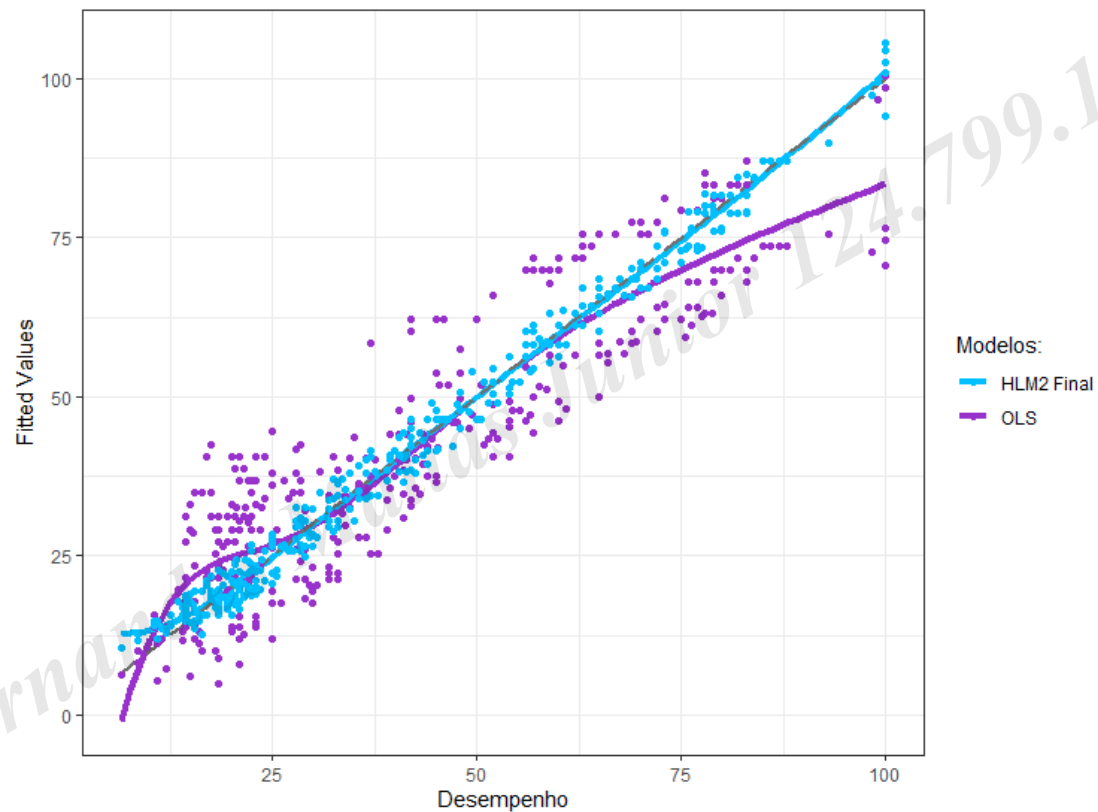
$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} \cdot texp_j + v_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} \cdot texp_j + v_{1j} \end{cases}$$

Substituindo...

$$desempenho_{ij} = \gamma_{00} + \gamma_{10} \cdot horas_{ij} + \gamma_{01} \cdot texp_j + \gamma_{11} \cdot texp_j \cdot horas_{ij} + v_{0j} + v_{1j} \cdot horas_{ij} + \varepsilon_{ij}$$

LINHA 406
DO SCRIPT

HLM2 x OLS



MUITO OBRIGADO!

Prof. Dr. Luiz Paulo Fávero

Professor Titular de Data Science & Analytics da USP

