MODELAGEM MULTINÍVEL (OU HIERÁRQUICA)

ASPECTOS TEÓRICOS E IMPLEMENTAÇÃO NO R

EDNALDO RIBEIRO

ESTRUTURAS HIERÁRQUICAS (OU MULTINÍVEL) DE DADOS

- **Muitas pesquisas em Ciências Sociais envolvem dados estruturados de forma hierárquica**

  - ✓ Estudos organizacionais sobre a influência de características do local de trabalho sobre a produtividade dos trabalhadores

    - ✓ **Unidades de Análises:** empresas e trabalhadores
    - ✓ **Dois Níveis:** micro e macro

- ✓ Demógrafos podem estar interessados em identificar como diferenças nos níveis de desenvolvimento econômico de diferentes países interagem com indicadores de escolarização para influenciar taxas de fertilidade.

  - ✓ **Nível Macro:** indicadores econômicos nacionais
  - ✓ **Nível Micro:** informações sobre escolaridade e fertilidade coletadas no nível individual

- ➢ **Dados Agrupados:** em ambos os casos podemos identificar que as unidades de análise que compõem o nível micro estão de alguma forma agrupadas no nível macro.

- Estrutura semelhante pode ser verificada quando múltiplas observações são realizadas sobre um mesmo grupo de pessoas ao longo do tempo.
- **Trajetória:** essas múltiplas medidas oferecem informações sobre as trajetórias de cada indivíduo.
    - Psicologia, (estudos de personalidade), Ciência Política (estudos sobre cultura política).
- **Hierarquia:** nestes casos as diferentes observações se agrupam nos indivíduos, que passam a ser as unidades de nível macro.

- **Estruturas de Três Níveis:** em pesquisas na área educacional são comuns estruturas de dados hierarquizadas em três níveis.
  - **Ex. 1:** Identificação dos efeitos de uma nova técnica didático-pedagógica sobre o rendimento dos alunos ao longo do tempo.
    - Nível micro: dados sobre o rendimento dos alunos coletados trimestralmente durante três anos.
    - Nível intermediário: as múltiplas observações dos rendimentos estão agrupadas no nível dos alunos, que passam a ser um nível hierarquicamente superior nesse caso.
    - Nível superior: as escolas

- **Ex. 2:** Identificação dos efeitos de uma nova técnica didático-pedagógica sobre o rendimento dos alunos com apenas uma medida e grupo de controle.
    - Nível micro: dados sobre o rendimento dos alunos submetidos à nova técnica e também dos que compõem o grupo de controle.
    - Nível intermediário: as turmas, com seus distintos professores e dinâmicas de grupo.
    - Nível superior: as escolas

PROBLEMA NA ANÁLISE DE ESTRUTURAS HIERÁRQUICAS

- Apesar dessa recorrência em pesquisas sociais até recentemente os pesquisadores não produziam análises adequadas a essa estrutura hierarquizada de dados.
  - Limitações das técnicas convencionais de estimação de modelos lineares com estruturas agrupadas, gerando nas Ciências Sociais problemas de viés de agregação, erros de precisão e confusões envolvendo as unidades de análises.
  - **Empobrecimento:** essa limitação tem conduzido a uma postura reducionista na formulação de hipóteses, que se concentram apenas em um nível de análise quando a realidade empírica exige claramente uma abordagem multinível.

BREVE HISTÓRICO

- **Diferentes Denominações:**
  - Modelos Lineares Multiníveis na Sociologia
  - Mixed-effects Models na Biometria
  - Random-coefficient Regression Models na Econometria
  - Covariance Componentes Models na Estatística

  - Modelos Lineares Hierárquicos: adotado por Raudenbush e Brick por focalizar a forma de estruturação dos dados.
- O termo foi criado por Lindley e Smith (1972) como parte de uma importante contribuição sobre estimação Bayesiana de modelos lineares.
  - Estimação de Componentes de Variância.

- Dempster, Laird e Rubin (1977) desenvolveram um algoritmo que tornou possível a aplicação do conceito de Lindley e Smith à um conjunto mais ampliado de casos.
- Dempster (1981) demonstrou que a técnica de ECV poderia ser aplicada a estruturas de dados hierarquizadas.
- Em 1982 e 1983 os primeiros trabalhos com aplicações dessa técnica foram publicados.
- Em 1988 foram apresentadas as primeiras ferramentas para ajustes desses modelos:
    - VARCL em 1988
    - MIXOR em 1996
    - SAS Proc Mixer em 1996
    - HLM 2000

FUNDAMENTOS DOS MODELOS MULTINÍVEIS (HIERÁRQUICOS) LINEARES

Regressão e Análise de Variância: a exposição dos fundamentos dessa técnica indica que se trata de algo derivado dos princípios presentes nas técnicas de regressão e de análise de variância.

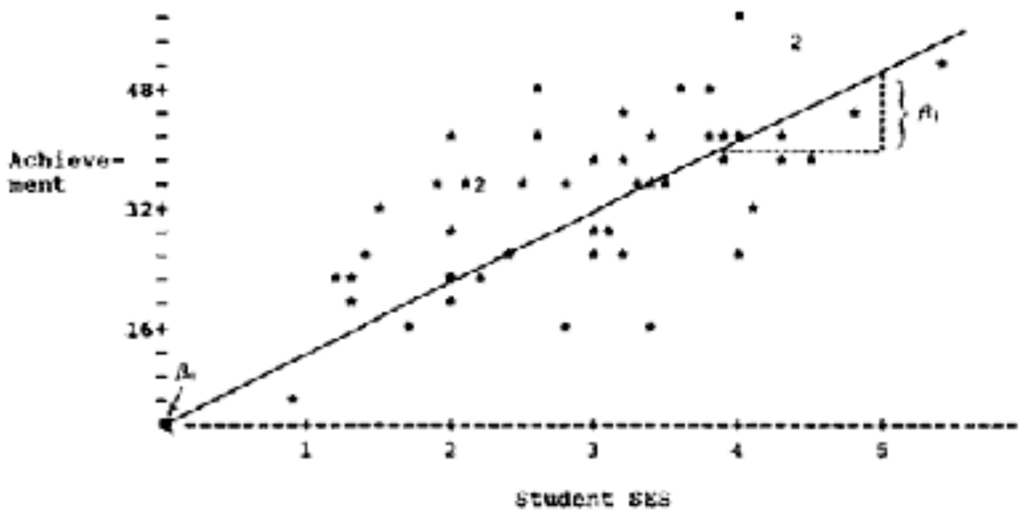
- **Revisão desses princípios (Regressão e ANOVA)**

Estudo Educacional: vamos seguir os passos propostos por Raudenbush e Brik (2002) e usar como exemplo um estudo sobre a relação entre Situação Socioeconômica (SS) e Nota em Matemática (NM).

- <u>Variável Dependente:</u> Nota de Matemática (NM)
- <u>Único Preditor:</u> Situação (Status) Socioeconômica (SS)
- <u>Nível 1:</u> ambas as variáveis se originam na unidade de análise "aluno", referindo-se ao nível micro.

Uma escola: começamos por um modelo que investiga esse problema em apenas uma escola.

Diagrama de Dispersão que descreve o relacionamento entre NM e SS



Nesse diagrama de dispersão podemos verificar esse relacionamento em uma única escola.

Intercepto e Inclinação: a linha reta representa a equação da regressão e parece representar bem o relacionamento.

$$Y_i = \beta_0 + \beta_1 X_i + r_i.$$

Y= NM

X= *SS*

β_0= NM esperada para um aluno com 0 em SS.

β_1= mudança esperada em NM associada a cada aumento de unidade em SS (na reta isso é representado pela inclinação entre o ponto 4 para 5 do eixo X).

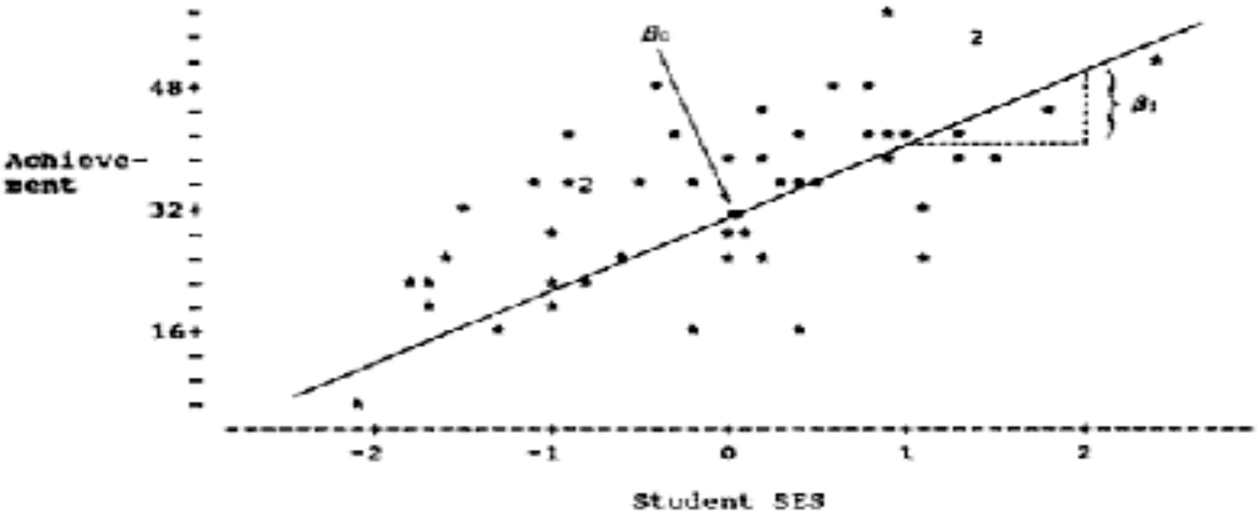
r_i= o termo de erro que representa um único efeito associado com o sujeito *i.* Presume-se que esse erro é normalmente distribuído com média e variância igual a 0.

Centralização de X (SS): nessa equação temos informações úteis sobre o quanto elevações na escala de SS impacta NM, entretanto, podemos tornar o Intercepto mais informativo se centralizarmos a variável independente.

Médias: ao invés da pontuação na escala de SS teremos informação sobre a média de SS na escola e sobre a distância de cada aluno em relação a esse patamar.

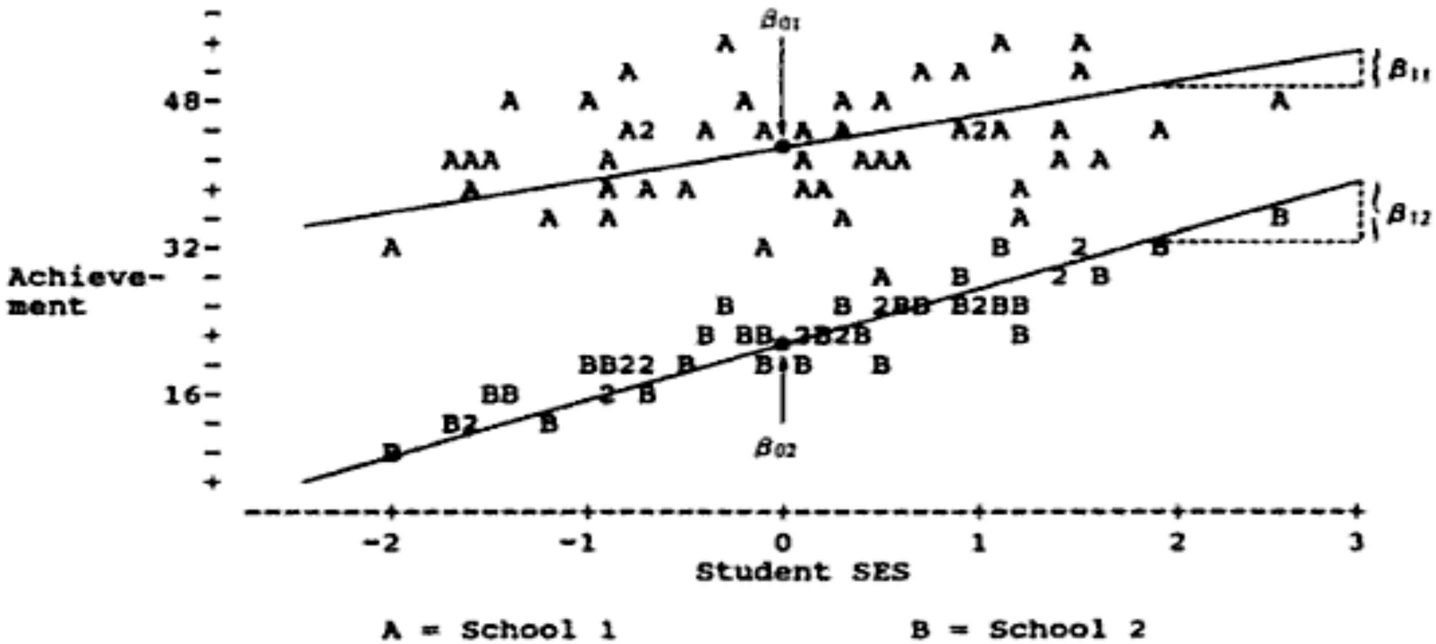
Mais realismo: com isso temos um intercepto que pode ser interpretado de maneira mais realista.

Procedimento: subtração da pontuação em SS de cada aluno pela média de SS da escola.



NM como função de SS centralizado: ao dispormos essa variável centralizada verificamos que o Intercepto (β_0) passa agora a representar a média de SS da escola e a inclinação (β_1) continua a mesma.

Duas Escolas: consideramos agora essa relação entre SS e NM em duas escolas hipotéticas.



Diferenças: podemos identificar que essas duas escolas se distinguem em diferentes aspectos

- ✓ A média de NM na escola A é maior do que na escola B.
- ✓ Essa diferença se reflete nos Interceptos ($\beta_{01} > \beta_{02}$)
- ✓ SS explica menos NM na Escola A do que na B
- ✓ Essa diferença se reflete no coeficiente de regressão ($\beta_{11} < \beta_{12}$)

Efetividade e Igualdade: considerando que os alunos são aleatoriamente distribuídos entre as duas escolas, esses resultados indicam que

✓ Escola A é mais efetiva no ensino, uma vez que possui a melhor média e é mais igualitária, na medida em que o efeito de SS é menor sobre NM.

População de Escolas: vejamos como fica a situação quando consideramos a mesma relação em uma população de escolas.

Amostra de Escolas: vamos supor que temos uma amostra aleatória de escolas (J) e que J representa um grande número.

Equação: como não temos condições de rodar um diagrama de dispersão para cada escola, vamos representar essa nova situação com a equação

$$Y_{ij} = \beta_{0j} + \beta_{1j} (X_{ij} - \bar{X}_{.j}) + r_{ij}$$

- ✓ Y_{ij} é a nota do aluno *i* da escola *j*
- ✓ B_{0j} é o intercepto da escola *j*, ou seja, a nota de matemática de um aluno com situação socioeconômica média (lembrando que centralizamos essa variável)
- ✓ B_{1j} é a inclinação na escola *j*, ou seja, o impacto que a alteração de um ponto em SS produz sobre a nota de matemática
- ✓ r_{ij} é o erro na previsão da nota de matemática do aluno *i* da escola *j*
- ✓ *A*ssumimos que r_{ij} é normalmente distribuído com variância homogênea entre as escolas
- ✓ Cada escola tem o seu próprio intercepto (β_0) e sua inclinação (β_1).
- ✓ Podemos entender que Efetividade e Igualdade são descritas pelo pares de valore de β_0 , $\beta_{1.}$

Adicionando termos: como agora lidamos com J e supomos que este segundo nível de análise é relevante para explicar NM, precisamos adicionar na equação termos que se refiram a esse nível hierarquicamente superior.

$E(\beta_{0j}) = \gamma_0$ $\quad\quad$ $\mathrm{Var}(\beta_{0j}) = \tau_{00}$

$E(\beta_{1j}) = \gamma_1$ $\quad\quad$ $\mathrm{Var}(\beta_{1j}) = \tau_{11}$ $\quad\quad$ $\mathrm{Cov}(\beta_{0j,}\beta_{1j,}) = \tau_{01}$

Onde

γ_0 é a média geral das médias de SS considerando a população de escolas (Somatória da média de SS de cada escola seguida pela divisão pelo número de escolas).

τ_{00} é a variância populacional entre as médias das escolas.

γ_1 é a média do coeficiente de regressão (inclinação) para a população de escolas (Somatória dos coeficientes de cada escola seguida pela divisão pelo número de escolas).

τ_{11} é a variância populacional entre os coeficientes de regressão (inclinação)

τ_{01} é a covariância populacional entre inclinação e interceptos.

Covariância: valores positivos de τ_{01} indicam que escolas com altas médias em NM tendem a ter também inclinações positivas.

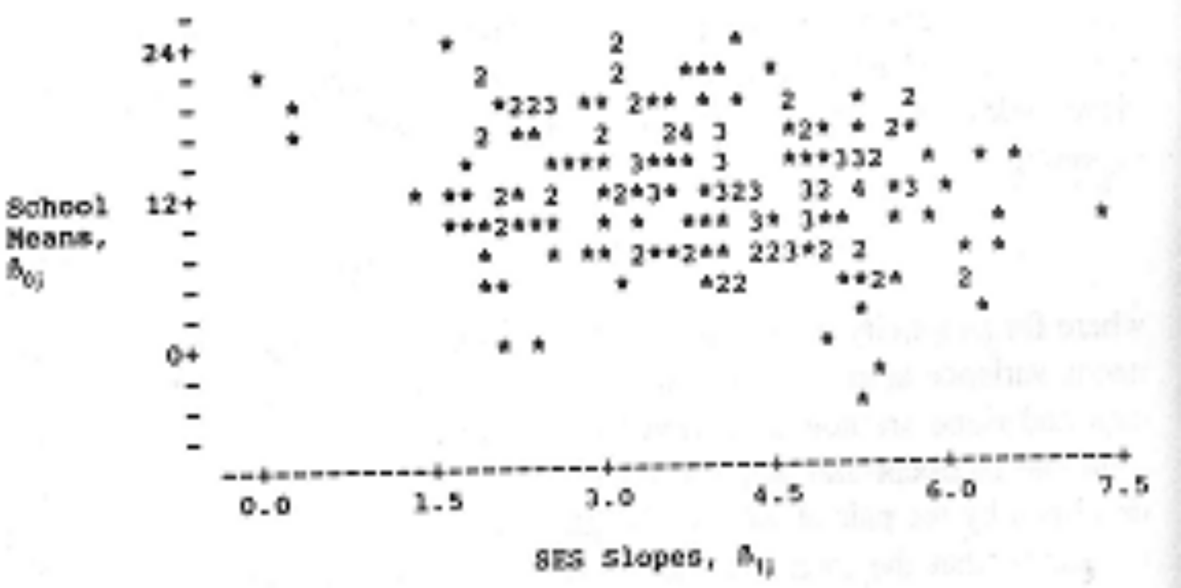
Correlação entre Médias e Inclinações: conhecendo cada variância e a covariância podemos calcular a correlação entre as médias e inclinações com a fórmula

$$\boldsymbol{\rho(\beta_{0j}, \beta_{1j}) = \tau_{01}/(\tau_{00}\,\tau_{11})^{1/2}}$$

Raramente conhecemos todos esses valores populacionais e as médias e inclinações da população de escolas. Na verdade todas essas informações devem ser estimadas com base nos dados coletados.

- Para continuar, entretanto, vamos supor que conhecemos os valores as médias e inclinações de cada escola.

- Esse novo diagrama apresenta a relação entre essas duas informações para uma amostra hipotética de escolas.



- **Informações:**
  - ✓ Maior dispersão entre médias do entre inclinações $(\tau_{00} > \tau_{11})$
  - ✓ Relação negativa entre Média e Inclinação, ou seja, escolas com médias mais altas tendem a apresentar efeito de SS menor.

Variação entre escolas: até agora vimos que as escolas variam em termos de interceptos e inclinações.

Explicação: diante dessa constatação a tarefa que se impõe é buscar explicações para essas variações, ou seja, devemos construir um modelo que possa prever $\boldsymbol{\beta_{0j}}$ **e** $\boldsymbol{\beta_{1j}}$.

Variáveis: como nosso interesse é identificar o que torna uma escola mais efetiva e igualitária, nesse modelo devem ser incorporadas medidas relacionadas a esse segundo nível hierárquico.

- ✓ Orçamento
- ✓ Características organizacionais
- ✓ Nível de envolvimento dos país

- **Exercício:** seguindo o modelo de Bryk vamos incluir como variável preditora algo mais simples
    - **Uma variável binária cujo valor 1 se refere à Escola Católica e 0 à Escola Pública, representada por** $\boldsymbol{W_j}$

Hipóteses

W_j está positivamente relacionada com efetividade e negativamente relacionada com a inclinação (efeito da desigualdade sobre a média de NM).

Podemos expressar essas duas hipóteses em duas equações de regressão:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \qquad \text{E} \qquad \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$$

Onde

γ_{00} é a média de NM para escolas públicas

γ_{01} é a diferença de média em NM entre Católicas e Públicas

γ_{10} é a média de inclinação (efeito de SS) para públicas

γ_{11} é a diferença média de inclinação entre Católicas e Públicas

u_{0j} é o efeito único da escola *j* sobre a média de NM mantendo W_j constante (controlando por W_j)

u_{1j} é o efeito único da escola *j* sobre a inclinação de SS mantendo W_j constante (controlando por W_j)

u_{0j} **e** u_{1j} **:** representam a variância e covariância residual, ou seja, aquilo que permanece após controlar pelos efeitos de W_j (Católica/Pública)

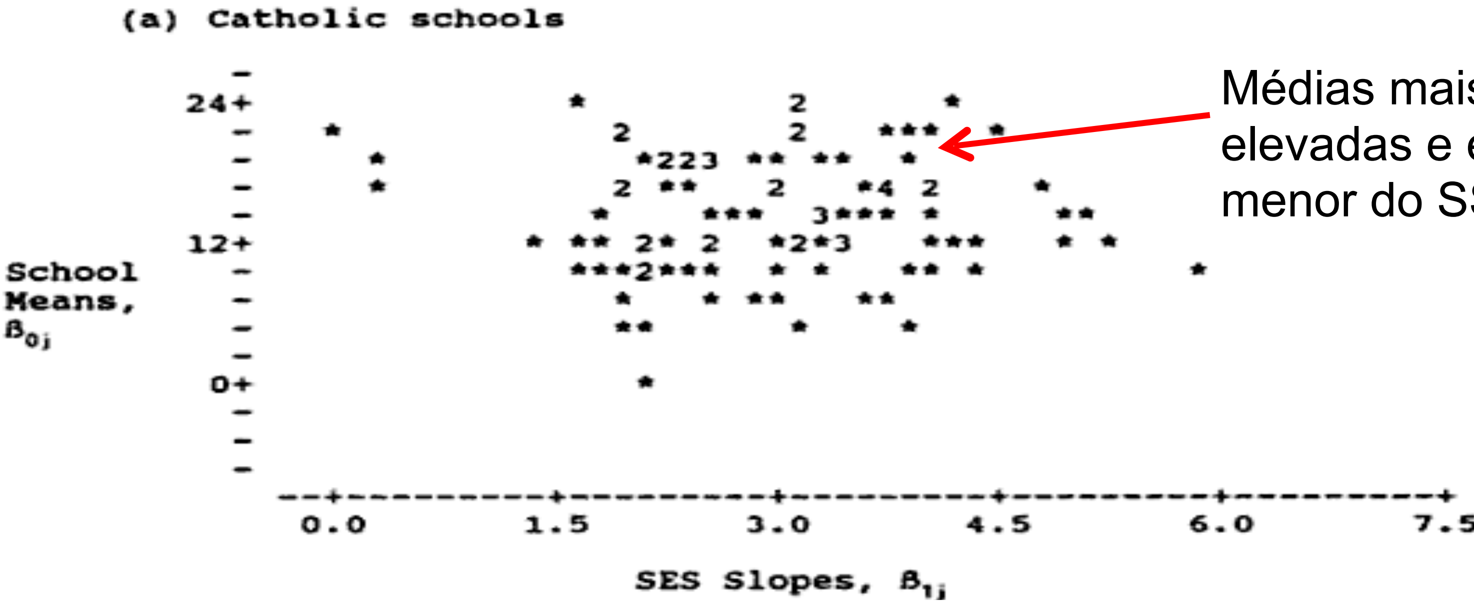
Médias e Inclinações não observáveis: não é possível estimarmos as médias e inclinações das escolas, pois essas informações não são observáveis.

NM e SS nos alunos: só podemos observar as notas dos alunos e as suas medidas socioeconômicas, ou seja, variáveis de nível micro.

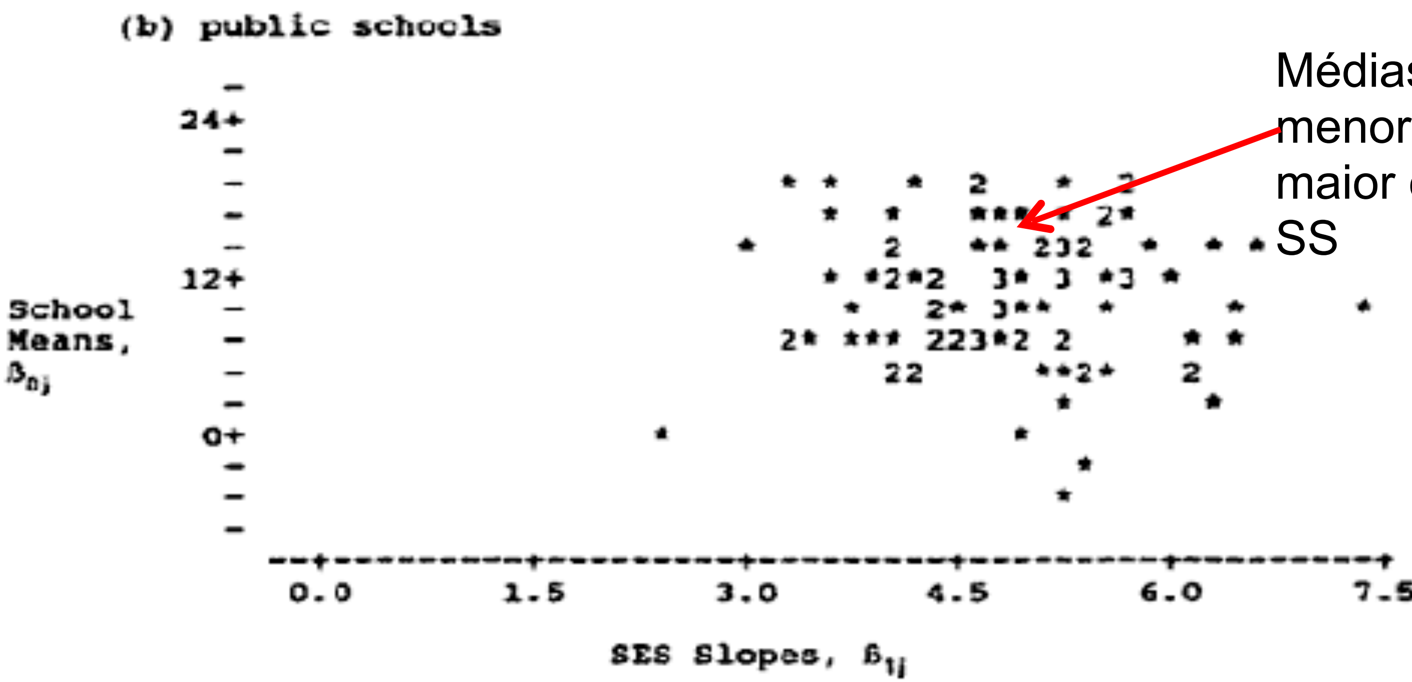
Hierarquia: pretendemos então avaliar o impacto de uma variável de nível 2, sobre interceptos e inclinações no nível 2, porém tomando como fonte de informações primárias variáveis de nível 1.

Equação

$$\boldsymbol{Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + \gamma_{11} W_j (X_{ij} - \bar{X}_{.j}) + u_{oj}}$$
$$\boldsymbol{+ u_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}}.$$



Médias mais elevadas e efeito menor do SS



Médias menores e maior efeito do SS

BASES MICRO E MACRO SOCIOLÓGICAS DA CULTURA POLÍTICA

Modelo: para ilustrar uma aplicação prática da modelagem Multinível vamos analisar os condicionantes do que chamamos de Orientação Cognitiva para a Política em perspectiva internacional, uma combinação do interesse por política com a eficácia política subjetiva.

Bases Micro e Macro: a literatura que investiga os condicionantes ou determinantes da cultura política tem identificado variáveis desses dois níveis.

Rara integração: todavia, são raras as tentativas de integração desses dois níveis em um único modelo.

Micro fundamentos: de um lado podemos verificar um grupo de trabalhos que buscam identificar que tipo de atributos individuais favorecem o envolvimento político (escolaridade, sexo, renda, posicionamento ideológico, etc).

Macro fundamentos: de outro lado encontramos pesquisadores que enfatizam dimensões estruturais do fenômeno como o nível de desenvolvimento econômico dos países (estrutura econômica) e o grau de liberalização das instituições políticas (estrutura política).

Regressão Linear no Nível 1: o primeiro grupo em geral se vale de modelos multivariados de regressão linear para identificar os possíveis preditores, trabalhando assim apenas com variáveis de nível 1.

Regressão Linear no Nível 2: o segundo em geral toma o percentual de participantes de cada país como variável dependente e introduz, também em modelos lineares, variáveis independentes como PIB per capita, Gini e indicadores sobre a estrutura política.

Integração necessária: essas soluções são precárias por vários motivos, mas o fundamental é que desprezam a natureza hierárquica da estrutura dos dados relativos ao fenômenos estudado.

Dados Micro: Latinobarômetro (18 países) e European Social Survey (36 países), ondas de 1995 a 2015.

Dados Macro: diversas fontes como o DataBank do World Bank, Freedom House, etc.

Objetivo: identificar os efeitos de preditores individuais e nacionais da variável dependente e suas possíveis interações.

Implementação no R!

UM EXEMPLO DE MODELAGEM MULTINÍVEL NO R

Banco micro: combinando Latinobarometro e ESS construímos um banco de dados com variáveis comuns.

	YearC	Year	Int	Ef	TParl	TPart	SDem	Gndr	Age	Ed	W	Att	Att10
1	12005	2005	1.0000000	0	0.6666667	0.3333333	0.6666667	1	70	0	0.9665447	1.0000000	10.000000
2	12005	2005	0.0000000	0	0.0000000	0.0000000	0.0000000	0	36	0	1.1394527	0.0000000	0.000000
3	12005	2005	0.6666667	1	0.0000000	0.0000000	0.0000000	0	20	0	0.8384388	1.6666667	16.666667
4	12005	2005	0.6666667	1	0.0000000	0.0000000	0.0000000	1	72	0	1.1394527	1.6666667	16.666667
5	12005	2005	0.6666667	0	0.0000000	0.0000000	0.6666667	1	74	0	1.1394527	0.6666667	6.666667

Variável Dependente: podemos ver a variável de Interesse (Int) e a de Eficácia (Ef), que foram combinadas para produzir uma medida de Att (Attentiveness for Politics, posteriormente batizada de Cognitive Orientation for Politics) padronizada para variar de 1 a 10.

Preditores Individuais: sexo, idade e escolaridade.

Banco macro: atributos econômicos e políticos dos países.

	CNTRY Countries	CNTRY_N Country number	YEAR Years	YEARCNTRY Year and Country	GDP GDP anual growth	UNMPL Unemployment	HDI HDI	GINI Gini Index	LAW Rule of Law - control of criminality	CORRUPTION Control of Corruption - Corru
1	Albania	19	2012	192012	1.42000000	13.438	750	28.95	-0.54	
2	Argentina	1	2005	12005	8.85165992	11.305	782	49.27	-0.58	
3	Argentina	1	2007	12007	9.00765088	8.415	792	47.37	-0.63	
4	Argentina	1	2010	12010	10.12559816	7.714	816	44.50	-0.57	
5	Argentina	1	2013	12013	2.40532370	7.100	825	42.28	-0.73	
6	Austria	20	2002	202002	1.65385494	4.849	837	NaN	1.86	
7	Austria	20	2004	202004	2.70574409	5.844	848	29.87	1.85	

Preditores Nacionais: Crescimento do PIB, IDH, Freedom House, etc.

Junção dos Bancos: usando o pacote *multilevel* do R devemos primeiramente fazer a junção desses dois bancos.

Variável Agrupadora: para isso é muito importante criar uma variável que possa conectar os casos individuais aos respectivos países.

- Como temos ondas diferentes em cada país, criamos uma variável que combina o ano da pesquisa com o código numérico do país
- **Exemplos:**
  - se um país X é codificado como 10 e teve uma onda em 2005, seu código na variável agrupado será 102005
  - um outro país que tem código 2 e teve uma onda em 2015 tem codificação final 22015

	YearC
1	12005
2	12005
3	12005
4	12005
5	12005
6	12005
7	12005
8	12005

```
LatinEuroMMfinal <- merge(LatinEuro, MacroLevel, by = c("YC"))
```

Banco Multinível: vamos pular essas etapas e trabalhar com o banco já estruturado de forma Multinível *LatinEuroMMfinal*

CORRELAÇÃO INTRA CLASSE (ICC)

Primeiro passo: como vimos no início, a modelagem multinível envolve a previsão de variância em diferentes níveis, logo, deve começar pela verificação dos níveis em que tais variações são estatisticamente significativas.

Variação entre os grupos: de forma bastante direta, se não existem evidências de que existe variação entre os grupos (no nosso caso países) não se justifica esse tipo de modelagem.

Preditores individuais: neste caso a variação se deve apenas à preditores de nível individual, então um modelo de regressão MQO seria suficiente.

Nível dos grupos: primeira examinamos as propriedades da vari[avel dependente no nível dos grupos e depois determinamos se a variância do intercepto é significativamente maior do que 0.

Modelo Nulo: esses dois aspectos são avaliados por meio de um modelo incondicional, também chamado de nulo.

- Sem preditores, mas com um termo de variância aleatório de intercepto para os grupos.
- Estima assim quanta variabilidade existe na média dos valores da variável dependente (variabilidade de intercepto) em relação a variabilidade total.

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}.$$

Descrição: esses modelo estabelece que

- A variável dependente Y (indivíduo i no grupo j) é uma função de um intercepto comum y_{00} e dois termos de erros
  - O termo de erro entre os grupos (u_{0j})
  - O termo de erro no interior dos grupos (r_{ij})

Qualquer valor de Y, portanto, pode ser descrito como uma média geral, envolvendo todos os indivíduos de todos os grupos, mais algum erro associado ao pertencimento aos grupos e outro erro de natureza individual.

Duas estimativas de variância: uma indicando o intercepto de cada grupo varia do intercepto geral e outra refletindo o quanto cada indivíduo se diferencia dos seus “colegas" de grupo.

Ativação do pacote: para estimar esses valores precisamos ativar o pacote com o comando *library(multilevel).*

Modelo

Null.Model <- lme(PolAtt10~1, random = ~1|YC, data = LatinEuroMMfinal, control = list(opt="optim"), na.action=na.omit)

- *Null.Model* é o nome que queremos atribuir ao output do modelo
- O <- indica que queremos que o nome seja atribuído ao que vem na sequência
- *lme* é o comando que indica o tipo de modelo, neste caso um linear multilevel
- Dentro do ( ) a primeira variável deve ser a dependente, seguida do ~, que neste caso significa “função de”
- Como se trata de um modelo nulo, não são inseridos preditores
- Após a, deve ser inserido o argumento de que a variável dependente deve variar entre os grupos, neste caso os países, por isso ~1|YC
- Na sequência precisamos indicar qual a fonte dos dados
- Para otimizar o tempo de análise inserimos use o comando *control = list(opt="optim"*
- Finalize solicitando a omissão de todos os casos com dados faltantes com na.action=na.omit)

Após executarmos o modelo, o output é exibido no console ao chamamos o objeto, mas mostra apenas a variância do intercepto (entre os grupos

```
> Null.Model
Linear mixed-effects model fit by REML
  Data: LatinEuroMMfinal
  Log-restricted-likelihood: -486646.7
  Fixed: PolAtt10 ~ 1
(Intercept)
   4.578273

Random effects:
 Formula: ~1 | YC
        (Intercept) Residual
StdDev:   0.9281079 3.133865

Number of Observations: 189856
Number of Groups: 168
```

```
> VarCorr(Null.Model)
YC = pdLogChol(1)
            Variance  StdDev
(Intercept) 0.8613843 0.9281079
Residual    9.8211077 3.1338647
```

Variação entre os países de 0,861

Variação nos grupos de 9,821

Para obter a estimativa da variância entre os indivíduos precisamos executar o comando *VarCorr(Null.Model)*

Cálculo ICC: com esses dois valores podemos calcular o índice de correlação intercalasse que é justamente uma medida do quanto da variação da medida dependente se localiza no nível dos grupos.

- ICC = Var. Intercepto/(Var. Intercepto + Var. Residual)

```
> 0.8613843/(0.8613843+9.8211077)
[1] 0.08063515
```

Var. Entre os países ICC Var. Nos grupos

Esse resultado indica que 8% da variação em nossa variável dependente se deve a fatores que se localizam no nível dos países, o que justifica o emprego da modelagem multinível.

Comparação: para determinar se a variância do intercepto é significativamente diferente de 0, podemos comparar os valores do log da probabilidade -2 desse modelo com outro sem essa randomização.

Novo modelo: para isso estimamos primeiro esse modelo alternativo com

```
> Null.gls<-gls(PolAtt10 ~1,data=LatinEuroMMfinal, control=list(opt="optim")
,na.action=na.omit)
```

Log das probabilidades: na sequência calculamos o logaritmo das probabilidades -2 para os dois modelos e, finalmente, subtraímos.

```
> logLik(Null.gls)*-2
'log Lik.' 988025.4 (df=2)
> logLik(Null.Model)*-2
'log Lik.' 973293.5 (df=3)
> 988025.4-973293.5
[1] 14731.9
```

14731.9 é a diferença entre os dois valores, que é significativamente em uma distribuição de qui-quadrado com 1 grau de liberdade, ou seja, indica significativa variação do intercepto.

ANOVA: adicionalmente podemos fazer uma análise de variância para verificar se os dois modelos são realmente diferentes em termos da sua capacidade explicativa.

```
> anova(Null.gls, Null.Model)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Null.gls       1  2 988029.4 988049.7 -494012.7
Null.Model     2  3 973299.5 973330.0 -486646.7 1 vs 2 14731.95  <.0001
```

Em um nível bastante elevado de exigência (0,001), as diferenças entre os modelos é significativa.

EXPLICANDO AS VARIAÇÕES DE NÍVEL 1 E 2

Duas fontes de variação: o modelo anterior mostrou que variações em nossa medida de Atenção Cognitiva para a Política dependem de fatores individuais (intra grupo) e nacionais (entre grupos).

Sexo, Idade e Escolaridade: no nível individual vamos propor que essa medida de cultura política seja influenciada pelo sexo, idade e escolaridade dos entrevistados.

- Sexo - em razão da sabida desigualdade entre homens e mulheres, supomos que os primeiros apresentem pontuações maiores (Mulher=1)
- Idade - quanto mais velho, maior a pontuação (4 Faixas)
- Escolaridade - quanto mais escolarizado, maior a pontuação (Superior=1)

Estrutura Econômica e Política: ao mesmo tempo pressupomos que a medida de cultura policia esteja relacionada a indicadores que sintetizam as diferentes estruturas econômicas e policias dos países.

- **Crescimento do PIB:** percetual de crescimento anual
- **IDH:** índice de GINI dividido por 100 para gerar uma escala de 0-10
- **FREEDOM HOUSE:** de 1 a 5, sendo que o 5 indica menor liberdade
- **ENEP:** número efetivo de partidos no parlamento
- **Sistema Político:** se parlamentarista (0) ou presidencialista (1)

Para além do efeito individual: isso quer dizer que se esperam efeitos dessas medidas nacionais acima dos efeitos das medidas individuais, em um modelo incremental.

Modelo

Model1.1Att <- lme(PolAtt10 ~ Gndr+ FxAge+ Ed+ GDP+ HDI100+ FREEDOM + ENEP+ POLSYS_rec, random = ~1|YC, data = LatinEuroMMfinal, na.action=na.omit, control = list(opt=“optim"))

Onde,

- O *Model1.1Att* <- atribui esse nome ao modelo especificado pelo restante da linha
- O *lme* indica o uso da função para modelos lineares multiníveis
- O ~ indica função de… e a lista de variáveis que vem a seguir são os preditores individuais e nacionais adicionados
- O *random* = *~1|YC* determina que os interceptos podem variar em razão dos países
- *O data*= indica a fonte dos dados
- O *na.action* determina que os dados faltantes sejam excluídos da análise
- O *control*=*…* apenas otimiza o processamento

Output: para exibir o resultado completo use o comando *summary(Model1.1Att).*

```
> summary(Model1.1Att)
Linear mixed-effects model fit by REML
 Data: LatinEuroWWfinal
       AIC    BIC    logLik
  840128.8 840239 -420053.4

Random effects:
 Formula: ~1 | YC
        (Intercept) Residual
StdDev:   0.8039551 3.067135

Fixed effects: PolAtt10 ~ Gndr + FxAge + Ed + GDP + HDI100 + FREEDOM + ENEP +      POLSYS_rec
                 Value Std.Error     DF   t-value p-value
(Intercept) -1.0488246 0.9050413 165122  -1.15887  0.2465
Gndr        -0.5700544 0.0151677 165122 -37.58349  0.0000
FxAge        0.0144823 0.0081013 165122   1.78765  0.0738
Ed           1.6623230 0.0269359 165122  61.71411  0.0000
GDP          0.0304398 0.0181019    146   1.68158  0.0948
HDI100       0.6874022 0.1008097    146   6.81881  0.0000
FREEDOM      0.1672774 0.0734420    146   2.27768  0.0242
ENEP        -0.0059118 0.0250775    146  -0.23574  0.8140
POLSYS_rec  -0.3390365 0.1516823    146  -2.23518  0.0269
 Correlation:
           (Intr) Gndr   FxAge  Ed     GDP    HDI100 FREEDO ENEP
Gndr       -0.009
FxAge      -0.007 -0.026
Ed          0.016  0.005  0.030
GDP        -0.239 -0.001 -0.001  0.001
HDI100     -0.983  0.001 -0.014 -0.020  0.199
FREEDOM    -0.573 -0.002 -0.002 -0.013  0.133  0.503
ENEP        0.074  0.001  0.003 -0.006 -0.139 -0.207 -0.063
POLSYS_rec -0.445  0.005  0.007 -0.005 -0.235  0.468 -0.191 -0.011

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.3174871 -0.8666812 -0.2571816  0.9218132  2.4676312

Number of Observations: 165277
Number of Groups: 152
```

Nível Micro:

- **Sexo:** podemos perceber que a primeira variável com efeito significativo foi o sexo, com valor negativo indicando que mulheres (=1) tem pontuação 0,57 menor do que os homens.
- **Faixa Etária:** não atingiu significância estatística ao nível de 0,05.
- **Educação:** se mostrou estatisticamente relevante com efeito positivo, elevando em 1,66 a pontuação.

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-1.0488246	0.9050413	165122	-1.15887	0.2465
Gndr	-0.5700544	0.0151677	165122	-37.58349	0.0000
FxAge	0.0144823	0.0081013	165122	1.78765	0.0738
Ed	1.6623230	0.0269359	165122	61.71411	0.0000
GDP	0.0304398	0.0181019	146	1.68158	0.0948
HDI100	0.6874022	0.1008097	146	6.81881	0.0000
FREEDOM	0.1672774	0.0734420	146	2.27768	0.0242
ENEP	-0.0059118	0.0250775	146	-0.23574	0.8140
POLSYS_rec	-0.3390365	0.1516821	146	-2.23518	0.0269

Nível Macro:

- **Crescimento do PIB:** não se mostrou significativo
- **IDH:** efeito positivo relevante e positivo de 0,69 a cada ponto adicional
- **FREEDOM:** efeito positivo de 0,17
- **ENEP:** sem efeito
- **Sistema Político:** efeito negativo indicando que em países presidencialistas a medida de atenção é menor

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-1.0488246	0.9050413	165122	-1.15887	0.2465
Gndr	-0.5700544	0.0151677	165122	-37.58349	0.0000
FxAge	0.0144823	0.0081013	165122	1.78765	0.0738
Ed	1.6623230	0.0269359	165122	61.71411	0.0000
GDP	0.0304398	0.0181019	146	1.68158	0.0948
HDI100	0.6874022	0.1008097	146	6.81881	0.0000
FREEDOM	0.1672774	0.0734420	146	2.27768	0.0242
ENEP	-0.0059118	0.0250775	146	-0.23574	0.8140
POLSYS_rec	-0.3390365	0.1516821	146	-2.23518	0.0269

Modelo incremental: Hofmann e Gavin (1998) atribuem essa terminologia por entender que as variáveis de nível 2 predizem a variância depois dos controles das variáveis de nível, então

- Uma mulher em um país presidencialista tem pontuação ainda mais agravada (-0,57 + -0,24)
- Um homem (0,57), em um país com HDI de 9 (9x0,69=6,21) tem pontuação de 6,78, bem acima da média geral de 4,59

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-1.0488246	0.9050413	165122	-1.15887	0.2465
Gndr	-0.5700544	0.0151677	165122	-37.58349	0.0000
FxAge	0.0144823	0.0081013	165122	1.78765	0.0738
Ed	1.6623230	0.0269359	165122	61.71411	0.0000
GDP	0.0304398	0.0181019	146	1.68158	0.0948
HDI100	0.6874022	0.1008097	146	6.81881	0.0000
FREEDOM	0.1672774	0.0734420	146	2.27768	0.0242
ENEP	-0.0059118	0.0250775	146	-0.23574	0.8140
POLSYS_rec	-0.3390365	0.1516821	146	-2.23518	0.0269

Variância explicada: é importante avaliar qual porção de variância passa a ser explicada com esse modelo, em comparação com a registrada no modelo nulo.

Nulo: intra grupos 9,82 e 0,86 entre os grupos.

Para estimar essas variâncias no novo modelo basta usar a função *VarCorr*

Intra: 9,41

Entre: 0,36

Em ambos os casos podemos perceber reduções, o que indica que a variância diminuiu na presença dos preditores.

Podemos calcular esse ganho com

Variância Explicada = 1-(Var com Preditores/Va sem preditores)

Intra: 1-(9,41/9,82)=0,04 (4%)

Entre: 1-(0,86-0,36)=0,5 (50%)

```
> VarCorr(Model1.1Att)
YC = pdLogChol(1)
            Variance  StdDev
(Intercept) 0.3647618 0.6039551
Residual    9.4073181 3.0671352
```

VARIÂNCIA DE INCLINAÇÕES

Terceiro passo: continuando nosso exercício de análise pode ser interessante avaliar se existem diferenças significativas nas efeitos das variáveis de nível 1 nos diferentes contextos (nível 2).

Modelo com inclinações variáveis: para isso precisamos estimar modelos que permitem a variação das inclinações dos preditores nos diferentes grupos (no nosso caso, países).

Escolaridade: vamos avaliar se a escolaridade tem efeitos diferentes entre os países e para isso basta adicionar essa variável na parte da equação relativa ao intercepto

Model.2.AttEd<-lme(PolAtt10 ~ Gndr + FxAge + Ed + GDP + HDI100 + FREEDOM + ENEP+POLSYS_rec,random= **~Ed|YC**, data = LatinEuroMMfinal, na.action=na.omit, control=list(opt="optim"))

Modelo com Inclinações Fixas

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-1.0595412	0.8896527	165122	-1.19096	0.2337
Gndr	-0.5674928	0.0151713	165122	-37.40576	0.0000
FxAge	0.0144009	0.0081031	[illegible]	[illegible]	[illegible]
Ed	1.6063196	0.0530143	165122	[illegible]	0.0000
GDP	0.0259045	0.0177966	146	1.45559	0.1477
HDI100	0.6907041	0.0991479	146	6.96640	0.0000
FREEDOM	0.1483627	0.0717534	146	2.06767	0.0404
ENEP	0.0000663	0.0245910	146	0.00270	0.9979
POLSYS_rec	-0.3276525	0.1494663	146	-2.19215	0.0300

interaction.plot(x.factor, trace.facto
c(1L:9, 0, letters), xpd = NULL, leg.b

Modelo com Inclinações Variáveis

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-1.0488246	0.9050413	165122	-1.15887	0.2465
Gndr	-0.5700544	0.0151677	165122	-37.58349	0.0000
FxAge	0.0144823	0.0081013	165122	1.78765	0.0738
Ed	1.6623230	0.0269359	165122	61.71411	0.0000
GDP	0.0304398	0.0181019	146	1.68158	0.0948
HDI100	0.6874022	0.1008097	146	6.81881	0.0000
FREEDOM	0.1672774	0.0734420	146	2.27768	0.0242
ENEP	-0.0059118	0.0250775	146	-0.23574	0.8140
POLSYS_rec	-0.3390365	0.1516821	146	-2.23518	0.0269

Comparação: para saber se a variação nos efeitos da educação sobre a cultura política nos diferentes países é estatisticamente significativa precisamos comparar esse modelo com interceptos variáveis com o anterior, com efeitos fixos.

Atualização do modelo: não precisamos voltar ao modelo anterior para isso, mas apenas atualizar esse último mudando apenas o ponto relativo às inclinações entre os países.

Model.2.Att.aEd<-update(Model.2.AttEd,random=~1|YC)

Em seguida basta testar com ANOVA a diferença da razão de probabilidade -2 entre os dois modelos

anova(Model.2.Att.aEd,Model.2.AttEd)

```
> Model.2.Att.aEd<-update(Model.2.AttEd,random=~1|YC)
> anova(Model.2.Att.aEd,Model.2.AttEd)
                Model df      AIC    BIC    logLik   Test  L.Ratio p-value
Model.2.Att.aEd     1 11 840128.8 840239 -420053.4
Model.2.AttEd       2 13 839999.8 840130 -419986.9 1 vs 2 133.0618  <.0001
```

Vejam que on logLik do modelo com interceptou variáveis é maior

A diferença na razão de probabilidade é de 133,06

E estatisticamente significativa

Qual contexto importa? Depois de constatar que o efeito da educação varia entre os países, nos resta tentar identificar quais características contextuais importam.

IDH e Escolaridade: uma hipótese possível relaciona os efeitos da educação em diferentes contextos de desenvolvimento humano

- Quanto mais desenvolvido o contexto menor o efeito da escolaridade sobre a cultura política
- Redução do efeito da desigualdade educacional sobre a cultura política em contextos mais favoráveis

Modelo Cross-Level: ou de interação micro-macro

ModelAtt.HDI.Ed<-lme (PolAtt10 ~ Gndr + FxAge +Ed **+Ed:HDI100** +GDP+HDI100+FREEDOM+ENEP+POLSYS_rec,random=**~Ed|YC**, data=LatinEuroMMfinal, na.action=na.omit, control=list(opt="optim"))

Vejam que na primeira linha a interação entre educação e HDI e no intercepto mantemos a variação no intercepto.

```
> summary(ModelAtt.HDI.Ed)
```

```
> summary(ModelAtt.HDI.Ed)
Linear mixed-effects model fit by REML
 Data: LatinEuroMMFinal
       AIC      BIC  logLik
  839989.9 840130.2 -419981

Random effects:
 Formula: ~Ed | YC
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 0.6080716 (Intr)
Ed          0.4073647 -0.256
Residual    3.0647173

Fixed effects: PolAtt10 ~ Gndr + FxAge + Ed + Ed:HDI100 + GDP + HDI100 + FREEDOM +       ENEP + POLSYS_rec
                 Value Std.Error     DF   t-value p-value
(Intercept) -0.6984958 0.8969162 165121  -0.77877  0.4361
Gndr        -0.5669346 0.0151713 165121 -37.36891  0.0000
FxAge        0.0147277 0.0081029 165121   1.81757  0.0691
Ed          -0.3074102 0.4618143 165121  -0.66566  0.5056
GDP          0.0268159 0.0178596    146   1.50149  0.1354
HDI100       0.6458592 0.1000651    146   6.45439  0.0000
FREEDOM      0.1480736 0.0720787    146   2.05433  0.0417
ENEP        -0.0022133 0.0246973    146  -0.08962  0.9287
POLSYS_rec  -0.3215503 0.1499265    146  -2.14472  0.0336
Ed:HDI100    0.2409404 0.0575101 165121   4.18953  0.0000
 Correlation:
            (Intr) Gndr   FxAge  Ed     GDP    HDI100 FREEDO ENEP   POLSYS
Gndr        -0.009
FxAge       -0.007 -0.026
Ed          -0.094 -0.011 -0.008
GDP         -0.238 -0.001 -0.001 -0.004
HDI100      -0.983  0.001 -0.015  0.105  0.198
FREEDOM     -0.567 -0.002 -0.002  0.000  0.134  0.496
ENEP         0.072  0.001  0.003  0.015 -0.138 -0.203 -0.062
POLSYS_rec  -0.445  0.006  0.007 -0.015 -0.235  0.467 -0.191 -0.010
Ed:HDI100    0.096  0.011  0.009 -0.995  0.004 -0.107 -0.001 -0.017  0.012

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.2872361 -0.8705086 -0.2569247  0.9212781  2.4664263

Number of Observations: 165277
```

ED:HDI: podemos verificar que a interação proposta apresenta efeito estatisticamente significativa e positivo, indicando algo diferente do que esperávamos.

- Parte da variação nos efeitos da escolaridade entre os países se deve à variação no IDH entre os mesmos
- Em países com maior IDH o efeito da escolaridade tende a ser potencializado

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.6984958	0.8969162	165121	-0.77877	0.4361
Gndr	-0.5669346	0.0151713	165121	-37.36891	0.0000
FxAge	0.0147277	0.0081029	165121	1.81757	0.0691
Ed	-0.3074102	0.4618143	165121	-0.66566	0.5056
GDP	0.0268159	0.0178596	146	1.50149	0.1354
HDI100	0.6458592	0.1000651	146	6.45439	0.0000
FREEDOM	0.1480736	0.0720787	146	2.05433	0.0417
ENEP	-0.0022133	0.0246973	146	-0.08962	0.9287
POLSYS_rec	-0.3215503	0.1499265	146	-2.14472	0.0336
Ed:HDI100	0.2409404	0.0575101	165121	4.18953	0.0000

Valores preditos: podemos detalhar essa interação propondo uma análise de valores preditos para diferentes perfis de indivíduos

Indivíduo1 = sem educação superior, residente no país com o menor IDH da região

Indivíduo2 = sem educação superior, residente no país com o maior IDH da região

Indivíduo3 = com educação superior, residente no país com o menor IDH da região

Indivíduo4 = com educação superior, residente no país com o maior IDH da região

Todas as demais variáveis são atribuídas igualmente aos quatro.

```
TDAT.Att.HDI.Ed<-data.frame(Ed=c(0,0,1,1),
                            Gndr=c(0,0,0,0),
                            FxAge=c(2,2,2,2),
                            GDP=c(3.021924, 3.021924, 3.021924, 3.021924),
                            HDI100=c(5.70, 9.48, 5.70, 9.48),
                            FREEDOM=c(1.643,1.643,1.643,1.643),
                            ENEP=c(4.89, 4.89, 4.89, 4.89),
                            POLSYS_rec=c(0,0,0,0))
```

Aplicação do modelo: uma vez criado esse pequeno banco de dados com apenas 4 casos, podemos aplicar o modelo que criamos com a interação e solicitar que pontuações para a cultura política desses 4 indivíduos sejam calculadas.

```
> predict(ModelAtt.HDI.Ed,TDAT.Att.HDI.Ed,level=0)
[1] 3.325855 5.767203 4.391805 7.743908
attr(,"label")
[1] "Predicted values"
```

Indivíduos de países com IDH baixo, mas o primeiro com escolaridade superior e o segundo não.

Indivíduos de países com IDH alto, mas o primeiro com escolaridade superior e o segundo não.

A diferença é maior entre os indivíduos em contextos de maior IDH.

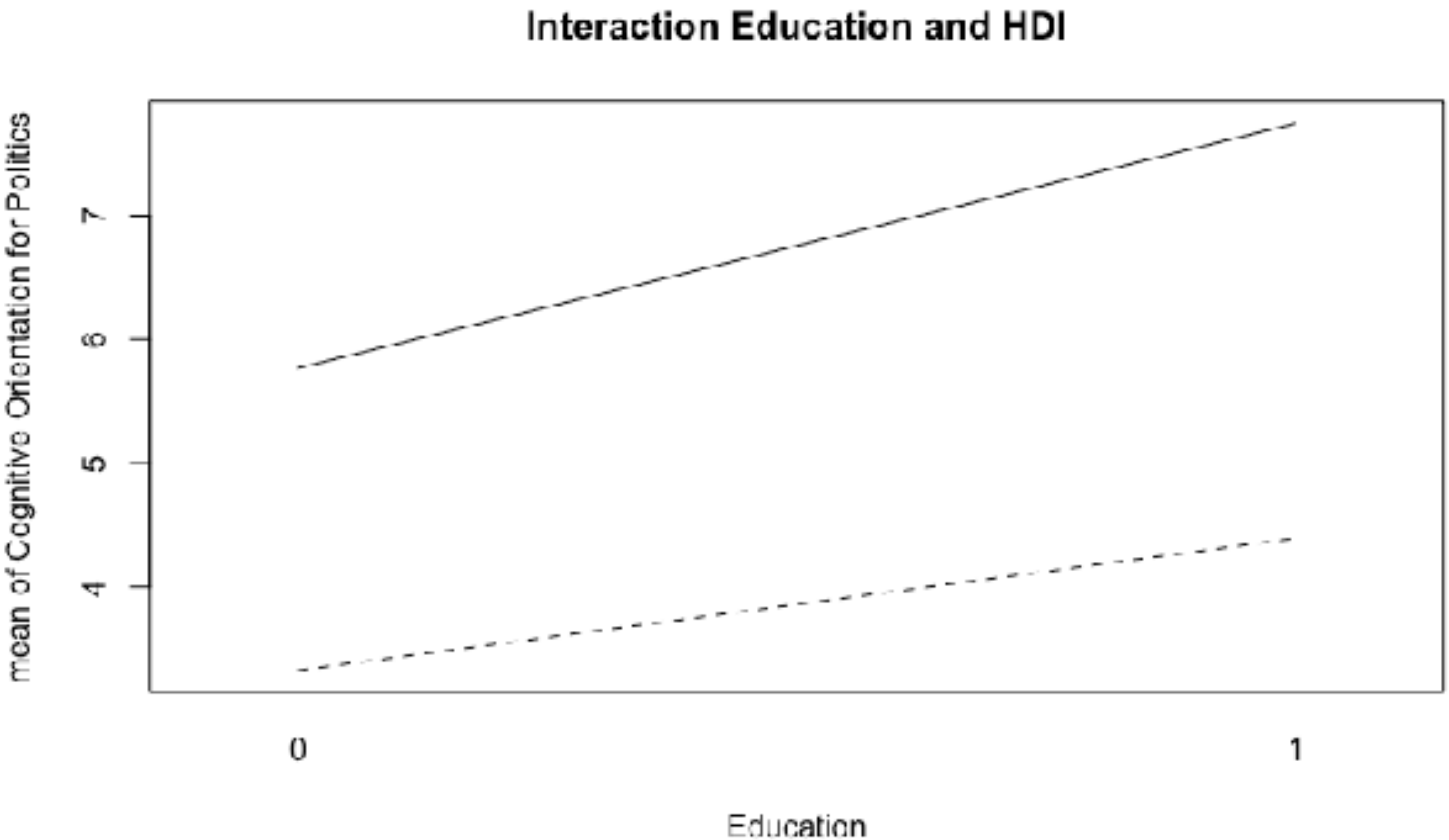
Visualização gráfica: Podemos construir um gráfico dessa interação para representar essa diferença contextual.

- Primeiro criamos uma variável em nosso banco de dados reduzido com os valores preditos

```
> TDAT.Att.HDI.Ed$PolAtt<-predict(ModelAtt.HDI.Ed,TDAT.Att.HDI.Ed,level=0)
```

- Na sequência usamos essa e outras variáveis do banco para criar o gráfico

```
with(TDAT.Att.HDI.Ed,interaction.plot(Ed,HDI100,PolAtt, legend=F,
xlab="Education", ylab="mean of Cognitive Orientation for Politics",
 main="Interaction Education and HDI"))
```



A linha tracejada representa o efeito da escolaridade no contexto de baixo IDH

Percebam como as inclinações são distintas